Universidad Nacional Autonóma de México
Posgrado en Ciencias e Ingeniería de la Computación
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Computación: Inteligencia Artificial

Sistemas de Clasificadores Adaptativos y su Aplicación a la Inteligencia Epidemiológica

T E S I S

QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
HUGO FLORES HUERTA

Director de Tesis:
Dr. Christopher R. Stephens S.
Instituto de Ciencias Nucleares

Ciudad Universitaria, Cd. Mx. Junio 2018

Resumen

UNIVERSIDAD NACIONAL AUTONÓMA DE MÉXICO

Sistemas de Clasificadores Adaptativos y su Aplicación a la Inteligencia Epidemiológica

por

Hugo Flores Huerta

El trabajo desarrollado a lo largo de este documento muestra la utilidad de usar algoritmos de minería de datos probabilisticos, en particular los basados en el teorema de Bayes como es el Clasificador Naive Bayes (NBC), para analizar y modelar enfermedades emergentes y presentar una alternativa a las técnicas estadísticas utilizadas por la Epidemiología para el análisis de las mismas. Hacemos análisis sobre datos de diabetes obtenidos de encuestas realizadas por instituciones publicas como son el IMSS y ENSANUT, otras de sectores privados pero de información publica como DxCG y el repositorio de la universidad de Irvine en california (UCI).

Mostramos los excelentes resultados y modelos que se pueden obtener de aplicar un clasificador tan simple como el NBC sobre estos datos, la utilidad de las herramientas de análisis como épsilon, la importancia de usar técnicas de suavizamiento como Laplace para lidiar con probabilidades cero e indeterminadas, en casos donde el conjunto de datos cuenta con una muestra muy pequeña o con ejemplos limitados. También exploramos diferentes técnicas para medición de desempeño de los modelos creados.

Proponemos diferentes algoritmos y sobre todo conceptos para la mejora tanto de los ranqueos como del desempeño en clasificación de los modelos creados por utilizar el NBC, uno de los cuales es el algoritmo de selección de características, el cual si bien conceptualmente ya existe en diferentes versiones en la literatura actual, no existía de tal forma que se adaptara a las necesidades que estábamos buscando en el desarrollo de este trabajo y para el tipo de clasificador propio que utilizamos a lo largo del trabajo, el cual es una version del NBC muy conocido. Mostramos los beneficios obtenidos de aplicar un algoritmo de selección de características como es: disminución del numero de atributos utilizados para crear el modelo, disminución del ruido extra agregado a los modelos por los atributos que no contribuyen en nada al modelo, lo que significa disminución de la variabilidad y varianza de los modelos obtenidos y por último en la mayoría de los caso hubo un ligero mejoramiento en el desempeño.

El clasificador NBC esta basado en una aproximación hecha al teorema de bayes (NBA) y es ampliamente utilizado por que ofrece un rendimiento sólido a través de un amplio espectro de dominios de problemas. Y dado que depende de un supuesto muy fuerte, el

cual es asumir la independencia de todos los atributos considerados como predictores para la creación del clasificador, esto es algo que usualmente desconcierta a las personas dedicadas a la modelación. Varias hipótesis han sido propuestas para explicar su éxito y también muchas mejoras al algoritmo donde se consideran las correlaciones o dependencias mas fuertes entre los atributos, las cuales son llamadas generalizaciones o Clasificador Naive Bayes Generalizado (NBG). En este documento proponemos un conjunto de medidas de error "local", asociadas con las funciones de probabilidad para subconjuntos de atributos y para cada clase, y mostramos explicitamente como estos errores locales se combinan para generar un error "global" asociado al conjunto completo de atributos. Al hacer esto formulamos un marco dentro del cual el fenómeno de la cancelación de errores, o el aumento, puede ser cuantificado y su impacto sobre el desempeño del clasificador estimado y predicho a priori. Estos diagnósticos nos permiten desarrollar una comprensión más profunda y cuantitativa de la razón por la cual la NBA es tan robusta y bajo cuales circunstancias se podría esperar que fallaría. Demostramos como estos diagnósticos pueden ser utilizados para seleccionar cuales atributos deben ser combinados y utilizados en una generalización simple de la NBA, dos tipos de generalizaciones se obtienen de este técnica una simétrica y la otra asimétrica en su aplicación a las funciones de probabilidad para la clase y no clase, a estos algoritmos los llamamos de forma simple como: Clasificador Naive Bayes Generalizado Simétrico (NBGs) y Asimétrico (NBGa), los cuales aplicamos a problemas del mundo real.

Otros dos conceptos o ideas innovadoras presentadas en este documento son las de minería de datos dinámica y adaptativa, en las cuales se busca capturar el efecto que produce el paso del tiempo sobre los fenómenos que se desea modelar, en el caso particular de las enfermedades emergentes y la epidemiología, como lo es la diabetes, es difícil pensar o creer que la enfermedad se mantendrá estática e invariable con el paso del tiempo, la enfermedades son fenómenos cambiantes y adaptables a las circunstancias variables de su medio ambiente e intervenciones creadas para combatirlas, de ahí la necesidad de contar con algoritmos de minería de datos capaces de capturar esta dinámica temporal presentada por este tipo de fenómenos, proponemos dos algoritmos uno dinámico y el otro adaptativo, los cuales son ideas diseñadas e implementadas en algoritmos probabilisticos con el fin de capturar estos efectos mencionados por el paso del tiempo y los aplicamos a datos reales relacionados con la diabetes.

Como se puede ver este documento es una colección de algoritmos probabilisticos, conceptos e ideas los cuales buscan modelar de la mejor forma los fenómenos tipo enfermedades emergentes y sus particularidades para presentar una alternativa a los análisis estadísticos hechos por la epidemiología. Y también podemos notar que estos algoritmos no necesariamente son solo útiles para la modelación de enfermedades si no para miles de conjuntos de problemas y datos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En la actualidad los datos son muy importantes, todos los datos recavados por empresas, encuestas, Internet, etc. pueden contener en si mismos un tesoro de conocimiento e información valiosa, el conocimiento generado por medio de estos es utilizado en los sistemas denominados actualmente como negocios inteligentes[1] pero no hay ninguna razón para no utilizar estas técnicas en la epidemiología, salud y enfermedades, este trabajo presenta la aplicación de las herramientas actuales, propuestas de modificaciones a estas herramientas diseñadas a lo largo de este trabajo y los beneficios que se pueden obtener con las mismas, para descubrir nuevo conocimiento a partir de datos y encuestas sobre un problema de salud muy fuerte en México como es la diabetes y proponer a la epidemiología nuevas herramientas útiles para los análisis que esta realiza cuando se presenta un problema de salud semejante al de México.

Y cabe mencionar que las nuevas técnicas presentadas en este documento no están restringidas a ser utilizadas únicamente en problemas de salud como la diabetes, si no en cualquier fenómeno susceptible de ser analizado, donde se puede descubrir nuevo conocimiento tal como puede ser, otros padecimientos, enfermedades, ventas, elecciones, seguridad, publicidad, educación, medio ambiente, etc. En fin cualquier fenómeno de la vida cotidiana el cual genere datos suficientes para poder ser analizado y modelado.

1.1. Marco General de la Epidemiología

La OMS (organización mundial de la salud) define a la epidemiología como: "el estudio de la distribución y los determinantes de estados o eventos (en particular de enfermedades) relacionados con la salud y la aplicación de esos estudios al control de enfermedades y otros problemas de salud"[2]. La epidemiología no se limita a estudiar las enfermedades infecciosas como muchos podrían suponer si no que es una palabra utilizada en el ámbito de salud para generalizar a cualquier enfermedad o problema de salud que afecte a un grupo de personas (población).

Esta ciencia no solo busca el origen biológico de la enfermedad sino que también se encarga de encontrar los factores sociales, ambientales, temporales, etc. Todos los ejes que tengan influencia en el grupo de personas las cuales padecen el problema de salud o enfermedad. Todo esto con el fin de encontrar los factores de riesgo (atributos)[3][4] de mayor relevancia para la descripción y el mejor entendimiento del padecimiento como son:

1. Causa del padecimiento, encontrar los factores genéticos y/o ambientales causantes del padecimiento.

2. Historia del padecimiento, desde la evolución (cambios y signos clínicos) hasta el resultado final ya sea recuperación o muerte del paciente.

3. Condiciones de salud, describir el estado de salud en el que se encuentra un grupo o población.

4. Factores de riesgo, descubrir los factores ambientales, genéticos y sociales por los cuales un grupo o población es vulnerable al padecimiento.

5. Control o prevención, definir si el problema de salud es prevenible o controlable.

6. Intervenciones, diseñar políticas de intervención para prevenir o controlar el problema y a su vez medir la eficacia.

La epidemiología es una ciencia que utiliza el método científico para obtener conocimiento [3]. Por todo lo descrito anteriormente es claro que la epidemiología utiliza ampliamente los métodos estadísticos y probabilísticos para poder cuantificar y calificar los factores de riesgo que describen un problema de salud y la eficacia al controlarlo.

Dada la complejidad de los padecimientos y enfermedades en los seres humanos la epidemiología tiene múltiples vías de estudio, a las cuales se les ha clasificado en tipos acorde con el estudio que realizan: [3][4].

- Epidemiología Observacional o Descriptiva, se basa en observar y recabar datos del grupo o población afectada para realizar un análisis estadístico de origen e impacto del problema.

- Epidemiología Experimental, esta podría ser llamada la primera intervención, se realizan estudios clínicos sobre distintos grupos expuesto y no expuestos al padecimiento para encontrar las formas de control y prevención inmediatas.

- Epidemiología Analítica, junta la información, datos y resultados de las dos anteriores para hacer análisis estadísticos y probabilísticos con el fin de encontrar y describir a detalle los factores de riesgo, causas y efectos del padecimiento, tendencias y el diseño de intervenciones de mayor efectividad, así como la medición de la eficacia de dichas intervenciones.

Es importante resaltar que la epidemiología analítica es tal vez el estudio de mayor importancia y relevancia, dado que mucho de las intervenciones y planes para control y prevención de los problemas de salud son definidos en esta etapa y es en esta donde las herramientas de la estadística y la probabilidad son ampliamente utilizadas pero también donde poco se ha explorado utilizando otras técnicas, como podría ser una de relativamente reciente creación y probada eficacia en la descripción, perfilamiento e identificación de patrones (en epidemiología factores de riesgo) de todo tipo de fenómenos (dado que una enfermedad o problema de salud puede ser considerado como tal) como es la minería de datos.

1.2. Los Sistemas Adaptativos Complejos

En la actualidad se reconocen dos tipos de sistemas (conjunto de elementos interconectados que trabajan juntos para que un todo funcione), los simples y los complejos, siendo estos últimos, los que con mayor frecuencia existen en la naturaleza o cualquier fenómeno de la vida, los sistemas simples son aquellos que cuenta con muy pocos elementos en su interacción, en cambio los sistemas complejos constan de múltiples elementos que regulan

su comportamiento[5].

El sistema adaptativo complejo es un tipo especial de sistema complejo[3], dado que además de contener múltiples elementos, los cuales interactúan entre sí, cada interacción entre elementos crea un nuevo factor capaz de alterar o influir en el comportamiento del sistema y además estas interconexiones y elementos dotan al sistema de la capacidad de aprender y cambiar, acorde a los factores externos en el que se desenvuelve el mismo, de ahí lo de adaptable.

En este tipo de sistemas lo más importante no son el número de variables o elementos que lo conforman, si no las relaciones (conexiones) existentes entre ellos, debido a que son estas conexiones quienes dotan al sistema de organización y comportamiento de acuerdo con un patrón, el cual puede ser identificable a través del tiempo[5].

John Holland [6] fue uno de los primeros en manejar el concepto de sistemas adaptativos complejos y lo definió mediante cuatro puntos principales:

1. Están conformados por un número grande de componentes o elementos.
2. Están organizados por niveles o estratos con elementos en cada uno de estos, que sirven de interconexión entre cada nivel, se podría decir también que cada uno de estos niveles son subsistemas y el sistema completo está conformado de estos subsistemas.
3. Estos sistemas nunca alcanzan un estado de equilibrio, es vital para ellos mantenerse cambiantes junto con el entorno para asegurar su propia existencia.
4. Los sistemas deben ser capaces de realizar predicciones sobre lo que pueda pasar a futuro, esto debido a que no solo es suficiente adaptarse al entorno, si no también predecir como cambiara, su entorno para poder hacer los ajustes necesarios antes de que este cambio llegue y así lograr lo que se denomina una ventaja competitiva , los sistemas que mejor se adaptan y son capaces de predecir los cambios en el entorno son los que tienden a sobrevivir con el paso del tiempo, los que no tienden a la extinción, este efecto lo logran gracias al aprendizaje que conlleva la adaptación continua al entorno, por lo cual son capaces de aprender e ir almacenando ese conocimiento, con lo cual pueden estimar los cambios futuros del entorno y crear estrategias para adaptarse a ellos.

1.3. La Revolución de los Datos

En un principio las computadoras y dispositivos de almacenamiento y procesamiento eran muy básicos y preciados, debido a lo caro que resultaba procesar y almacenar la información, por esto los datos que eran almacenados, guardados y analizados debían ser muy selectos y depurados, no se almacenaba cualquier información, si no solo aquella que desde un principio se sabía podría contener datos útiles y por lo tanto valía la pena almacenarlos para su posterior análisis. Posteriormente con el desarrollo tecnológico, se avanzo rapidamente hacia los bajos costos y la miniaturización de los componentes electrónicos (sobre todo microchips), en la década de 1970, se pudo dotar a las computadoras de mayor capacidad de almacenamiento y procesamiento de información, este crecimiento sostenido continua aun en la actualidad, por lo que hoy resulta relativamente barato almacenar volúmenes gigantescos de información y datos para su posterior procesamiento.

Ademas en la década de 1990 el internet alojo el WWW (World Wide Web), lo que desato el éxito y explosión mundial de este medio tan poderoso actualmente, el internet,

si a esto se añaden los dispositivos móviles y su incursión a la gigantesca red mundial, nos encontramos con miles de millones de bytes de información generados y almacenados diariamente en esta red, información conocida actualmente como el BIG DATA.

El BIG DATA [7] no es otra cosa que millones de terabytes de información que se generan, circulan y producen diariamente en internet, móviles y empresas (públicas y privadas), se cree que el 80 % de esta información es no estructurada (no está en una base de datos) y el otro 20 % es información estructurada (en BD), lo valioso de esta información radica en que no es preseleccionada, sesgada o filtrada se genera espontáneamente, no tiene un propósito inicial de ahí su valor, utilizando diferentes técnicas de análisis de datos se pueden encontrar patrones e información valiosa, como nunca se hubiera pensado que podría existir dentro de estos millones de terabytes de datos desorganizados y sin propósito hasta ahora, muchos almacenados en BD de compa nías solo para mantener una cartera de clientes.

Las técnicas utilizadas para la explotación (análisis de los datos e información con el propósito de generar conocimiento) del BIG DATA son: herramientas estadísticas, probabilistas y regresiones, pero sin duda la que mejor explota esta gran cantidad de información es la minería de datos (cap. 1.4) y la utilización de sus diferentes técnicas (cap. 2). El BIG DATA se ha definido con 3 dimensiones [7] y se les llama las tres V, las cuales se describen a continuación:

- Volumen: Usualmente se habla de terabytes y petabytes de información pero no siempre es así, lo primordial seria tener los datos pero esto no es posible para cada fenómeno o problema que se analiza, aun así las técnicas de minera de datos pueden lidiar con relativamente poca información y a pesar de esto generar conocimiento, señalando que ninguna técnica funcionaria si la información es demasiado escasa.
- Velocidad: La información sobre todo en la internet y debido al comportamiento del mercado cambia constantemente, en el análisis y toma de decisiones se requiere que los datos sean actuales para ser validos, la información demasiado antigua ya no es válida y el conocimiento que se podría generar de esta tampoco sería válido.
- Variedad: esta dimensión es muy importante, la variedad de los datos agrega riqueza a la información, entre mayor variedad se tenga en la información más rico es el conocimiento generado por las distintas técnicas.

Es claro que las enfermedades y padecimientos estudiados por la epidemiología concentran un gran cantidad de información, que cumple con todo lo descrito anteriormente, para ser considerado como un problema de BIG DATA (como son las tres V mencionadas anteriormente), por lo cual se pueden usar las técnicas de minería de datos (cap. 2), para generar conocimiento que aporte a la epidemiología herramientas y certeza al diseñar sus estrategias de control y prevención, así como a diseñar las intervenciones necesarias para cualquier problema de salud al que se enfrente, en este trabajo presentamos un análisis para un problema de salud fuerte en México, como lo es la diabetes utilizando datos e información almacenada y proporcionada por instituciones públicas de salud como son: el IMSS y ENSANUT, las cuales si bien no son un BIG DATA tal cual, al menos son un buen inicio para generar conocimiento e ilustrar que este tipo de análisis, no solo son factibles, si no que también entregan muy buenos resultados y ademas pueden ser aplicables sobre el BIG DATA.

1.4. La Minería de Datos

La minería de datos es una herramienta actualmente sub-utilizada en México, es principalmente usada en enfoques de negocios y ventas, en los modernamente llamados negocios inteligentes, los cuales buscan pronosticar el éxito de un producto y las caractersticas que deben tener los clientes apropiados o propensos a adquirir dicho producto (esto se conoce como perfilar), sin embargo con esta herramienta del cómputo también se puede modelar cualquier fenómeno, siempre y cuando se cuente con la información correcta y suficiente, existen diferentes algoritmos computacionales diseñados para hacer minería de datos (cap. 2) algunos funcionan mejor que otros dependiendo del problema a analizar y viceversa, no existe una sola herramienta o algoritmo que sea siempre la mejor en todos los problemas, de ahí la existencia de una variada gama de algoritmos, los cuales funcionan mejor en cada tipo de problema, el objetivo principal de la minería de datos es generar conocimiento nuevo, no existente, a partir de grandes volúmenes de información, la cual describe de manera directa o indirecta de alguna forma un fenómeno, el cual puede ser: ventas, campañas publicitarias, perfiles socio-econmicos, enfermedades, padecimientos de salud, tendencias, elecciones, preferencias, nuevos mercados, termodinmicos, qumicos, médicos etc. el campo de aplicación es casi inagotable, todo fenómeno del cual se deriven datos es susceptible de ser explotado y analizado utilizando la minería de datos.

La ventaja de la minería de datos sobre las técnicas estadísticas y probabilistas, utilizadas para analizar los datos, es que no requiere de un orden previo de los datos, en cierto sentido, no se requiere de una preselección de los datos ni de filtros, tampoco requiere de conocimientos previos sobre el fenómeno, para seleccionar la información adecuada, de hecho, la minería de datos prefiere que ningún filtro o preselección sea aplicado a los datos, debido a que los datos son su materia prima y entre más tenga es mejor, su objetivo es generar nuevo conocimiento, por lo que el conocimiento existente no debe ser aplicado para hacer preselecciones de información, este nuevo conocimiento puede estar oculto en la información menos pensada, en la que a simple vista del experto y en su opinión se pensaría que no esta relacionada con el comportamiento del fenómeno de estudio.

La minería busca en lo más profundo de la información, no importa si son BIG DATA, grandes BD o simples archivos, el poder de esta herramienta, está en el aprendizaje automatizado hecho por los algoritmos diseñados con este fin, ningún experto o analista puede sesgar la información o el aprendizaje del algoritmo ya que este busca, encuentra y aprende los patrones y perfiles de forma automática, con los cuales genera nuevo conocimiento en el campo y con el aprendizaje hecho por el algoritmo la minera de datos es capaz de manejar dos aspectos, el pronóstico y la clasificación.

Por lo anterior, la minería de datos tiene mucho éxito encontrando patrones, relaciones y correlaciones, en grandes volúmenes de datos, los cuales no podrían ser descubiertos o si quiera imaginados, sin la ayuda de las herramientas de esta técnica. En este trabajo en particular, aplicaremos lo algoritmos probabilisticos existentes, particularmente el clasificador Naive Bayes (NB) y versiones modificadas del mismo, las cuales son propuestas, diseñadas y probadas en este proyecto de investigación para la modelación del fenómeno de las enfermedades epidemiológicas y en particular la diabetes.

Una vez descritas estas técnicas en el capitulo 2, quedara claro que las innovaciones en el campo de la minería de datos buscan herramientas y algoritmos que disminuyan los errores y aumenten los aciertos a la hora de pronosticar y/o clasificar. Y también algoritmos adaptables a la presencia de nuevos datos y aún más deseable que capturen la dinámica temporal del fenómeno a modelar. Los dos algoritmos más utilizados y efectivos son: las

redes neuronales y el clasificador NB, este proyecto propone modificaciones, mejoras y herramientas para ser utilizadas en el clasificador NB, el cual es considerado como el algoritmo de minería de datos cuya naturaleza es más descriptiva y analítica, debido al uso y manejo de probabilidades.

1.4.1. Clasificación vs regresión

La minería de datos tiene como tarea principal analizar los datos, con el objetivo de extraer el conocimiento oculto en ellos, este conocimiento pueden ser: patrones, relaciones o descubrimiento de nuevo conocimiento a esto se le llama modelos. Los modelos solo pueden ser de dos tipos: predictivos o descriptivos [8]. Los modelos predictivos, como su nombre lo dice, tratan de hacer estimaciones de una variable denominada objetivo (o dependiente) a través de mirar en sus variables de entrada (o independientes), por otra parte los modelos descriptivos buscan características, patrones o perfiles en las variables independientes para identificar las características comunes (Perfil o Perfilamiento) entre los miembros que pertenecen al mismos objetivo (o resultado en la variable objetivo). Los modelos predictivos son Clasificación y Regresión y los descriptivos son agrupamientos, reglas de asociación y correlaciones. Aun cuando para la modelación en este proyecto utilizaremos ambas partes: predictiva y descriptiva, las innovaciones y aportaciones propuestas en este documento están enfocas a la predicción y en particular a la clasificación.

Los modelos de regresión, en una forma simplista de definirlos, son aproximaciones matemáticas, las cuales tratan de capturar el comportamiento del fenómeno que se esta modelando, el resultado de la modelación, es una función matemática de valores reales, el resultado de la modelacin crea una predición la cual es un número real, a diferencia de la clasificación cuyo resultado es un valor discreto.

1.4.2. El paisaje de la predictibilidad

Podemos definir a la predictibilidad como la cantidad de información contenida en los datos sobre un fenómeno, la cual ayuda o no a modelar dicho fenómeno, y por lo tanto, a predecir su comportamiento futuro, también se puede definir como la propensidad que tiene un fenómeno a ser predecible, a través de la creación de un modelo.

Pero por otra parte, un fenómeno puede ser totalmente predecible, pero si no se cuenta con la información suficiente y necesaria en los datos, ser imposible crear un modelo que pueda predecirlo. Por lo tanto podemos concluir que la predictibilidad de cualquier evento, fenómeno o cualquier cosa a modelar se encuentra dentro de los datos e información de la que se dispone. Los datos son entonces lo mas importante a la hora de modelar, dentro de ellos se encuentran los atributos, características y/o objetivos.

Para poder hacer predicción requerimos de datos suficientes, es imposible tener toda la información sobre un fenómeno, pero para los algorítmos de minería de datos eso no es necesario, pueden trabajar con una muestra de la información y estimar como se comporta el fenómeno, en lo general, para así poder predecir cualquier nuevo evento. Hay muchos problemas al tratar con datos, los cuales veremos mas adelante ref(cap 2), pero los mas importantes son la dimensionalidad y la detección de variables no útiles, las cuales solo aportan ruido a la modelación.

Dimensionalidad: Es algo que debe evitarse al modelar, esto para poder crear modelos generalizables y evitar procesamientos de datos de años o tal vez siglos, por ejemplo: en una encuesta de 50 preguntas con 10 posibles respuestas, el número de combinaciones

posibles seria de 10^{50}, si se aplicara a cada persona que habita en le planeta obtendríamos una muestra de 10^{40}, lo cual muestra que ni aun cuando se aplicara a todas las personas del planeta cubriríamos el total del espacio de búsqueda existente. Con esto es claro que jamas se contara con la inforamción completa al crear modelos.

1.5. Antecedentes de la Minería de Datos

El análisis de los datos siempre ha existido, no es que la minería de datos haya inventado este concepto o descubierto la utilidad de extraer conocimiento o experiencias de ellos, anteriormente el método tradicional utilizado para este propósito, era de forma "manual", esto es, se contrataban especialistas en la materia o área de la cual se trataban los datos, y estos, analizaban los datos utilizando técnicas estadísticas, revisando los mismos minuciosamente y generando informes, sus resultados eran la presentación de tendencias e hipótesis basadas en los análisis de estos datos. Resulta obvio ver, que este tipo de procedimiento resultaba en procesos de análisis muy lentos y caros, ademas de que existía un sesgo muy pronunciado en los resultados obtenidos, debido a que los resultados la mayoría de las veces representaban mas los conocimientos, opiniones y experiencias de los especialistas encargados de analizar los datos.

Posteriormente, los volúmenes de información generados en la actualidad, con las nuevas herramientas informáticas ha crecido de forma gigantesca, lo cual ha superado nuestra capacidad de análisis y manejo de estas cantidades de datos, las bases de datos y el estándar SQL surgieron como herramientas de ayuda para organizar y manejar estos grandes volúmenes de datos, pero estas herramientas son únicamente para obtener sumarización y descripción de los datos y no para generar conocimiento, por lo cual se tiene que recurrir nuevamente a los especialistas para su análisis, pero no todos los especialistas tenían esta capacidad de manejar esta herramientas, por el cual el proceso ahora conducía a recurrir a un intermediario informático para obtener información condensada de los grandes volúmenes de información, el cual entrega al especialista información pre-procesada agregando otro sesgo a un proceso de por si ya sesgado, por lo descrito anteriormente.

Debido a los problemas ya descritos que se tenían con el uso de las herramientas tradicionales, surge la necesidad de crear una nueva generación de herramientas y técnicas, las cuales puedan extraer conocimientos de grandes volúmenes de datos, sin ningún tipo de sesgo, por tener que recurrir a expertos en el área o comprimir la información disponible. De ahí el surgimientos del conjunto de herramientas y técnicas de extracción del conocimiento automático, llamados algoritmos de minería de datos o simplemente minería de datos, basados muchos de ellos en conceptos de inteligencia artificial y aprendizaje automático, los cuales no necesitan de intervención humana para la generación de reglas, patronés, clasificación, grupos y perfiles o lo que se llama simplemente conocimiento. Ahora la intervención de los expertos se emplea en la interpretación y aplicación de los resultados (conocimiento) obtenidos.

1.6. Epidemiología Tradicional vs Minería de Datos

La epidemiología y la minería de datos, como ya describimos en las secciones anteriores, tienen como objetivo común inferir patrones, tendencias, factores, etc., en resumen, describir de la mejor manera un fenómeno para poder entenderlo y predecirlo antes de que suceda. Esto es de vital importancia para la epidemiología, debido a que logrando la mejor descripción y predicción del fenómeno, así como los factores que lo producen, pueden ser

salvadas miles o millones de vidas, dependiendo del problema de salud analizado.

Las diferencias entre ambas técnicas son: los fenómenos que analiza la epidemiología, se particulariza al estudio de los fenómenos que afectan la salud de un grupo o población, como son enfermedades y padecimientos, al contrario la minería de datos puede analizar cualquier tipo de fenómeno, incluidos los de salud siempre y cuando exista información suficiente para hacerlo; las técnicas de análisis utilizadas por ambos son la principal diferencia, mientras la epidemiología utiliza las técnicas clásicas de la estadística y probabilidad, la minería de datos por el contrario utiliza las nuevas técnicas de análisis de datos, como son el naive bayes, redes neuronales, arboles de decisión, máquinas de soporte vectorial, k-means, etc., todas la cuales tiene en común que para la modelación y creación de modelos, así como, para encontrar los factores que influyen en el fenómeno, se utiliza aprendizaje automático y no supervisado, lo que produce mejores resultados al no estar sesgado por la experiencia o creencias de los expertos que utilizan las herramientas estadísticas para crear los modelos con técnicas clásicas.

Otra diferencia entre las técnicas clásicas (estadística y probabilidad) y las nuevas (minería de datos), es el volumen de datos que se puede manejar y lo más importante que se pueden analizar, la minería de datos brinda herramientas, las cuales ofrecen resumen de los factores mas simples de entender y analizar, a diferencia de una persona analizando millones de registros.

Capítulo 2

Clasificación y la Minería de Datos

La clasificación es una de las dos tareas llevadas acabo con las herramientas de la minería de datos (la otra es pronostico), la clasificación es de lo mas utilizado en la minería de datos, creando el modelo adecuado permite predecir la pertenencia o no a la clase de un elemento de acuerdo a sus caracterstícas (factores). Por esta razón son tan útiles estos modelos, ya que predicen quien, por ejemplo, será un comprador de cierto producto, o quien puede padecer diabetes u otra enfermedad o quien abandonara la escuela, etc., las posibilidades al clasificar son muchas.

2.1. Los Pasos en la Minería de Datos

La minería de datos, es una parte importante de algo mayor, que es la extracción del conocimiento a partir de información almacenada, y por lo tanto existen etapas o fases para poder lograr la adecuada extracción del conocimiento, de ahí viene la analogía de minería de datos, es buscar y separar el conocimiento (metales preciosos o minerales valiosos) del montn de información almacenada (minas, montaas, etc.), y después la difusión del modelo y conocimiento obtenido (convertir el metal o mineral en un producto). Todas las etapas descritas son iterativas, se puede regresar a cualquiera de ellas para cambiar o agregar información, buscando mejorar siempre el modelo y por ende que el conocimiento obtenido sea más preciso, entre las etapas siempre existirá interacción, para revisar los resultados parciales en cada una de ellas y vigilar la consistencia de lo generado hasta ese momento. La figura 2.1 ilustra estas etapas.

2.1.1. Recopilación y Almacenamiento de la Información

La información sobre un determinado fenómeno usualmente no esta disponible en un solo lugar, sobre todo, cuando no se planeo desde un principio almacenarla para este propósito e inclusive aunque fuera almacenada de esta forma no es adecuado limitar la extracción de conocimiento a una sola fuente de datos, por lo tanto en esta fase la tarea es encontrar toda la información disponible relacionada con el fenómeno a analizar, inclusive si es de distintas fuentes (pueden ser datos públicos o de otras instituciones, etc) siempre y cuando puedan ser integradas con el resto de la información para contar con materia prima (suficientes datos) para la extracción del conocimiento.

La información requerida puede estar en distintos formatos como pueden ser: archivos de texto, distintas tecnologías de base de datos (postgress, microsoft sql, dbase, etc), excel, access, etc., por lo cual es necesario no solo recopilarla, si no también, transformarla

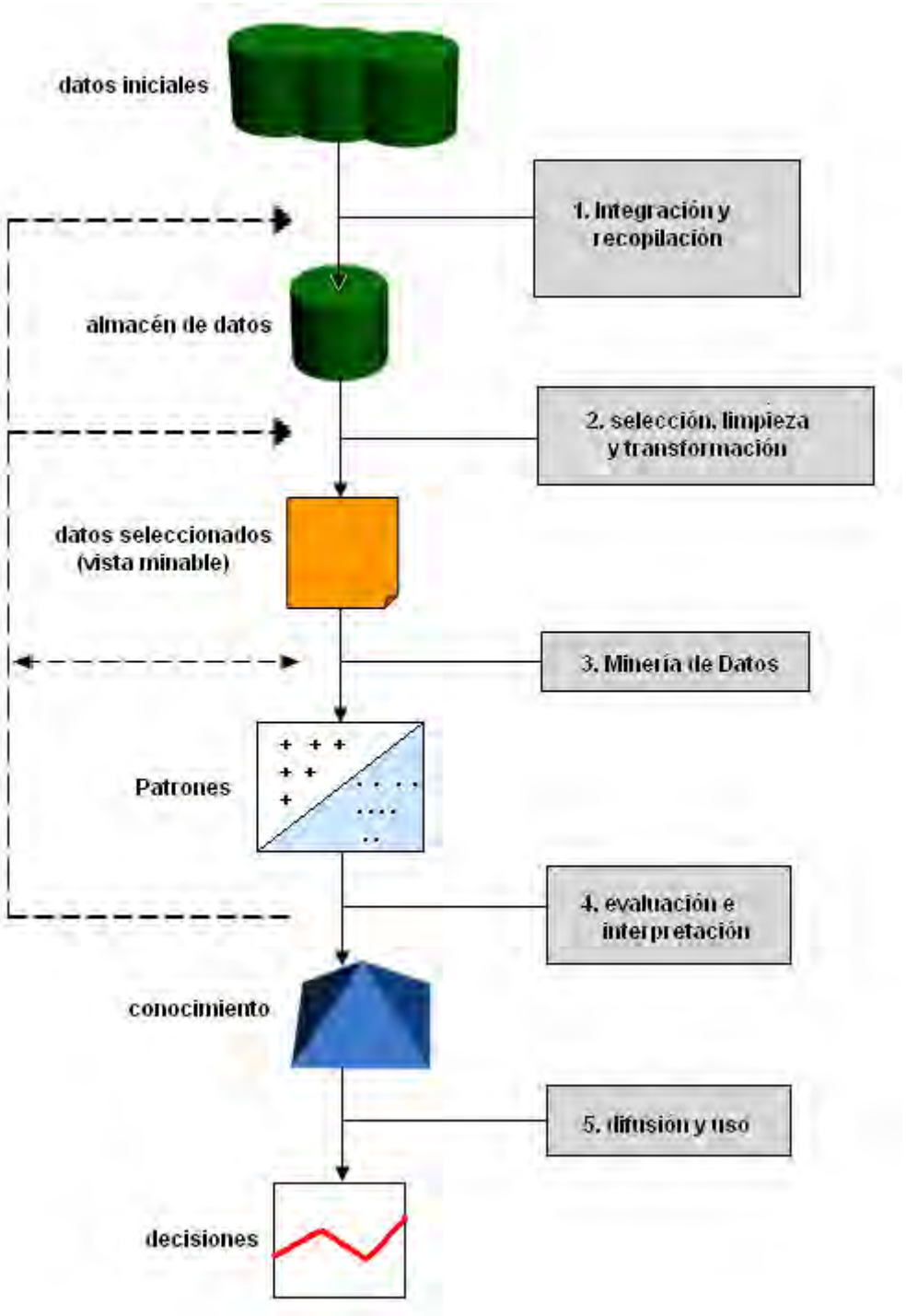


Figura 2.1: Fases de la Extracción del Conocimiento [8]

y hacer los procesos computacionales necesarios para tenerla en un solo formato bajo, un solo manejador de base de datos o archivos, para poder explotarla con las herramientas de la minería de datos.

Para la integración es necesario que la información pueda ser relacionada de alguna forma, para lo cual, la información recopilada debe tener datos en común, por ejemplo: si se quiere incluir la información sociodemográfica del INEGI al análisis de un fenómeno, será necesario contar con información de direcciones, códigos postales, colonias o delegaciones

para poder relacionar la información con los datos de los censos de INEGI, que precisamente son hechos usando esos niveles de detalle, y por lo tanto estarán disponibles como llaves para consultar la información sociodemográfica de cada colonia, código postal o delegación.

Toda información que no pueda se integrada y relacionada con el objetivo (clase) de estudio, no podrá ser utilizada para hacer minería de datos, por muy buena, interesante y basta que sea.

2.1.2. Limpieza y Transformación

Después de recopilar toda la información disponible y útil (la que fue posible de relacionar con el objetivo), es necesario limpiar los datos, esto es, algunas de las variables y datos recopilados son inútiles para los análisis, por ejemplo: las llaves (ids), variables de un solo valor o totalmente nulas, ya que no es posible con ellas lograr ningún tipo de discriminación, la limpieza solo debe consistir en eliminar aquellas variables evidentemente inútiles para la modelación, como los casos mencionados anteriormente, no se deben eliminar variables por consideraciones teóricas, experiencia o por evidencias documentadas en la literatura o pasadas, dado que el objetivo de la minería de datos es descubrir información nueva dentro de un conjunto de datos y al eliminar información, por tomar alguna de las consideraciones anteriores, puede sesgar el descubrimiento de la misma o limitarla. Por lo tanto, cualquier información por muy extraña o no relacionada con el objetivo que parezca, no debe ser eliminada a menos que sea evidentemente inutilizable.

La otra parte muy importante de la minería de datos es la creación de nuevas variables o transformación de las misma a partir de variables existentes, cuyo valor tal cual se encuentra no seria útil o utilizable por el algoritmo o en el mejor de los casos, la discriminación proporcionada por ellos no seria tan buena como la obtenida después de una transformación adecuada. Variables del tipo “fecha de nacimiento” por ejemplo, tal cual la fecha de nacimiento de alguien no es útil para la modelación, debido a que pueden existir miles de combinaciones posibles, es mas interesante si esta fecha de nacimiento, se utiliza para calcular la edad de la persona, esta nueva variable (edad) será más interesante como predictor que la fecha de nacimiento tal cual, a partir de esta se pueden obtener mejor patrones y discriminación, este seria un ejemplo de una transformación.

Otro tipo de transformación es discretizar la información existente de variables numéricas continuas, las cuales resulta, si no imposible utilizarlas en el estado natural en que se encuentran, no practico debido al número de infinitas posibilidades de valores que la variable tiene, por lo cual, en estos tipo de variables se aplica la discretización por rangos, por ejemplo: edad, temperatura, ingresos, mediciones, etc. variables numéricas con posiblemente un número infinito de valores posibles son transformadas a un número de rangos finitos mas manejables y con información mas valiosa y discriminante comparados con sus estado natural, por ejemplo la edad puede ser transformada de una variable númerica continua a rangos del tipo: 0 a 10 años 1, 11 a 20 años 2, ... 91 a 100 años 10, con lo cual al final tendríamos una variable edad discretizada, que en lugar de 100 valores posibles tendría solo 10 (1 a 10) posibles valores donde, por contar con una muestra mas grande (hay mayor número de personas y por ende registros en el rango 1 (0 a 10 años), que por separada cada edad), existe mayores posibilidades de encontrar patrones y una mejor descriminacíon.

Hay otra forma de crear rangos sin que eso implique la partición igualitaria como vimos en el ejemplo anterior, de 10 años de edad por cada rango, este seria un tipo particular de

crear rangos donde se busca que los rangos sean equitativos en su partición de intervalos, la otra forma y una de las mas utilizadas en este trabajo, por resultar ser muy util y dar mejores resultados en muchos de los casos de discretización, en cuanto a la discriminación que es posible lograr, consiste en hacer los rangos de tal forma que la equivalencia se encuentre en el número de registros contenido en cada rango, sin importar que los intervalos en estos rangos no sean equitativos, si no que el número de registros sea equitativo, esto da buenos resultados,dado que podemos partir la muestra de tal forma que existan el mismo número de personas en cada rango, este tipo de partción es mas justa con la muestra y evita sesgos, resultantes de una pobre muestra en un intervalo.

Por ejemplo en una partición equitativa por rangos como la mencionada primeramente en el rango 1 (0 a 10 años) podemos encontrar tal vez una muestra de 100 registros y en el segundo rango 1,000 y en el tercero 20,000 y en el último 10 esto hace que las muestras en los rangos no sean parejas y que podamos encontrarnos con sesgo en los resultados, de la otra forma podemos encontrarnos con rangos tan disparejos como: rango 1 (0 a 26 años), rango 2 (27 a 35 años), rango 3 ...., rango 10 (54 a 100 años), donde cada rango tendrá el mismo número de registros, es decir el mismo tamaño de muestra distribuida equitativamente entre cada uno de los rangos, lo que da la posibilidad de evitar el sesgo provocado por la transformación o discretización de la variable.

2.1.3. Selección de Características

La selección de características es una herramienta muy importante y utilizada en la modelación actual. El Clasificador Naive Bayes se beneficia enormemente de esta herramienta, tanto en aumento de desempeño, como en optimización en procesamiento. En general se ha encontrado que una buena pre-selección de las variables que describen el fenómeno a modelar, debe eliminar aquellas variables con un alto grado de correlación[9], lo cual ayuda al algoritmo de NB a tener un mejor desempeño y precisión a la hora de clasificar, porque se eliminan variables, que de otra forma, podrían ser redundantes y con esto dar el doble de peso a dichas variables desequilibrando la clasificación del modelo[9][10], también permite la reducción del número de variables que participan en el modelo, lo cual en modelos de predicción o clasificación de fenómenos con un número muy grande de variables a analizar, se traduce en ahorros significativos en tiempo de ejecución, como es en predicción de texto[11][12], clasificación de spam[13] y reconocimiento de escritura[14], entre otros.

Se han desarrollado muchas técnicas de selección de características combinándolas con el clasificador NB [9][10][11][12][13] y [14], las cuales son muy buenas, pero parecen particularizar el uso a la modelación de determinados fenómenos o aplicaciones especficas, la herramienta de interés para este proyecto es la que pueda seleccionar características a través de cumplir con múltiples restricciones u objetivos, dado que no se desea predisponer el aprendizaje del clasificador NB por discriminar variables haciendo consideraciones teóricas utilizando la experiencia y conocimientos actuales del fenómeno con lo que se pueda contar, si no que es preferible dejar a los datos"hablar"que sea el mismo modelo quien descubra la información útil para lograr la mejor predicción y desempeño. Existen múltiples trabajos elaborados hasta la actualidad, con variaciones entre ellos para crear el algoritmo de selección de características que mejor se adapta a las necesidades de cada problema, por lo cual se opto por hacer lo mismo, crear un algoritmo de selección de caractirística, el cual es adecuado para el tipo de fenómenos que deseamos analizar, este algoritmo tiene mucho en común con todos los existentes salvo por el algoritmo genético utilizado para optimizar la selección de estas características, utilizamos un algoritmo evo-

lutivo multi-objetivo (MOGA), para hacer la selección de características, al mismo tiempo que se optimizan múltiples parámetros (la mayoría solo optimizan a lo más 2 parámetros a la vez)[14][16][17] [18][19], en estos trabajos utilizan distintas herramientas de la minería de datos como son: redes neuronales[17][18], clúster[14] y modelación difusa[18] pero ninguno junta los dos factores, selección de características con MOGA y clasificador NB. Otros trabajos proponen múltiples objetivos a optimizar que puedan ser utilizados de forma general independientemente del clasificador[16][19]. En este trabajo proponemos particularmente optimizar el desempeño de un clasificador NB y encontrar las mejores variables que describen el fenómeno utilizando selección de características con MOGA.

Los trabajos publicados mencionados han empleado diversos MOGA para el algoritmo de selección de características, siendo los más populares el: NSGA[14][16][17] y MOGA[19], estos MOGA suelen tener problemas a la hora de optimizar más de dos características, por lo cual en este trabajo optamos por utilizar un MOGA de reciente creación el RankMOEA[20] que se describe como un mejor algoritmo comparado con los anteriores, puede manejar de buena forma más de dos objetivos a optimizar al mismo tiempo y además ofrece un espacio de soluciones a lo largo del frente de Pareto, lo que permite dar múltiples soluciones validas, dejando al modelador la decisión de tomar las soluciones que considere mejores o adecuadas para su objetivo.

2.1.4. Minado de Datos

Después de pasar los datos por las tres fases anteriores obtenemos los que se llama una “vista minable”, la cual no es otra cosa si no los datos pre-procesados, limpios y con las transformaciones necesarias para ahora si crear el modelo o hacer la extracción del conocimiento, el cual pueda contener esta información. Esta es la fase en donde se hace la minería de datos y su objetivo es generar conocimiento nuevo a partir de patrones, relaciones, etc.. toda nueva información que pueda ser extraída de estos datos.

Para extraer este nuevo conocimiento, lo primero es definir bien que deseamos hacer, después elegir la clase de algoritmo de minería de datos apropiada a nuestros datos y la tarea elegida, y por último elegir un objetivo (clase) para la extracción de conocimiento (patrones, clasificador, predicción, perfilamiento, etc.). Los procesos de la minería de datos se engloban en dos tareas principales que se pueden llevar acabo con esta herramienta: Predictivas y Descriptivas las cuales a su vez tienen sus propias técnicas para lograr el objetivo.

Los procesos predictivos, como su nombre lo indica, buscan predecir lo que puede pasar a futuro, a través de la experiencia vista en los datos con lo cuales se creo el modelo, es decir, busca patrones en los datos que le ayuden a suponer sobre un conjunto similar lo que ocurrirá, las dos tareas predictivas son:

- Clasificación: Este tipo de técnica es de las mas utilizadas y sobre la cual puede encontrarse mas información en la literatura, incluso hay concursos para ver cual algoritmo puede clasificar de mejor manera o con mas aciertos versus otros, como su nombre lo dice, el objetivo aquí es clasificar, en este tipo de problemas se define una clase u objetivo a buscar, por ejemplo en nuestro caso, las enfermedades, nuestra clase puede definirse como las personas que padecen diabetes y la no clase personas sin diabetes, personas con hipertensión y sin hipertensión, cáncer o sin cáncer, etc., para hacer clasificación cada registro en nuestro datos debe estar asociado a una clase, no siempre es bi-valuada, pero en la mayoría de los casos es lo mas utilizado,

una vez elegido el objetivo (clase), el resto de las variables (atributos) serán utilizados como predictores de esa clase, es decir sus valores son utilizados por el algoritmo de minería de datos para encontrar patrones, relaciones, etc., que le permitan a partir de ellos decir cual pertenece a la clase y cual no, en un conjunto de datos totalmente nuevo, del cual no se conoce la clase y se trata de predecir cual será esta. La utilidad de este tipo de algoritmo es evidente, por ejemplo: poder predecir cual persona es factible a padecer determinada enfermedad por solo tener su historial clínico o socio-demográfico o ambos (variables asociadas), igualmente en ventas o tarjetas de crédito o cualquier objetivo a clasificar.

- Regresión: Esta técnica es muy ampliamente utilizada en estadística, economía y ventas, el objetivo es muy simple, se asume de una dispersión de puntos que esta se debe a una relación existente entre una variable dependiente y una o muchas independientes, por lo cual se trata de capturar (aprender) el efecto de estas variables independientes sobre la dependiente, es decir capturar la influencia de como y cuando cambian estas variables independientes o como afectan o cambian a la dependiente, para así crear una función real que represente lo mas fiel posible a esta dependencia y los cambios cuantificados que sufre la variable dependiente a través de las independientes, la utilización es sobre todo en ventas, economía y estadística dado que al contar con una función real, representando a la información, se puede obtener un valor estimado para cualquier combinación de variables independientes de la variable dependiente, por ejemplo: cual será el volumen de ventas, dadas ciertas características del mercado, en cualquier punto del tiempo o los costos acorde al cambio de los indicadores económicos (variables).

Los procesos descriptivos, al igual por su nombre, nos dicen mucho de su objetivo, el cual es diferente al de los predictivos, los cuales intentan decirnos lo que pasara, en estas solo se desea saber lo que esta pasando o paso, se busca una descripción de la información disponible para hacer análisis e interpretaciones y tratar de entender el por que de una situación o tendencia. Las principales tareas son:

- Agrupamiento: Es de las técnicas mas utilizadas dentro de los procesos descriptivos y mas conocida aun por su nombre en ingles "clustering", el objetivo es extraer a partir de los datos grupos, es decir se trata de encontrar a los objetos muy similares y agruparlos en un conjunto, ademas los elementos de un grupo deben ser lo suficientemente distintos de los otros grupos creados, y pueden ser tantos como objetos distintos existan, el algoritmo trata de formar estos grupos, a diferencia de la clasificación donde previamente la clase (grupo) es conocido y definido, aquí se busca dar esa pertenencia a los objetos, el objetivo es encontrar esos grupos y los objetos que los componen. Como ejemplo de una aplicación de este tipo de técnica están los sitios de compras por internet, cuando un usuario compra un producto el algoritmo coloca a este usuario en un grupo de preferencias y así puede recomendar al usuario que otros artículos adquirir.

- Correlaciones: Es otra de las técnicas aplicadas en las tareas descriptivas, es una forma de cuantificar, que tan parecidas son dos variables y sus valores, usualmente estas variables deben ser numéricas, aunque existen técnicas, las cuales permitirían convertir o simular como numéricas, variables categóricas para poder utilizar la correlación en ellas, existen muchas formas y formulas para obtener la correlación entre dos variables, a este valor usualmente se le conoce como el coeficiente de correlación, ademas es usualmente normalizado entre 1 y -1, entre mas cerca a 0 este el valor, quiere decir que no existe correlación entre estas variables, valores positivos cercanos a 1, hablaran de una fuerte similitud entre estas variables y un valor negativo cerca

no a -1, nos habla de anti-correlación, esto es, cuando hay correlación positiva si una variable crece la otro crece al mismo tiempo y por el contrario en anti-correlación cuando una crece la otra decrece en el mismo instante, se puede ver claramente la utilidad de esta técnica, por ejemplo: si se quiere evitar fugas de agua constantes en una instalación utilizando mangueras y se necesita saber cual es el grosor adecuado de esta manguera para prevenir al máximo las fugas de agua, entonces podría buscar información relacionada con eventos de fugas y los grosores de mangueras utilizados en estas, para encontrar la correlación negativa o positiva entre ambas variables y, así poder definir cual sería el grosor mas apropiado.

- Reglas de Asociación: Otra de las tareas descriptivas, cuya funcionalidad es muy similar a la anterior técnica (correlaciones), pero para el caso donde las variables son categóricas, lo cual resulta muy útil, dado que en el mundo real no todas variables son numéricas o pueden transformase en una, lo que nos evitaría encontrar relaciones entre este tipo de variables. Estas reglas de asociación son enunciados, los cuales pueden ser tan complejos o sencillos como sean útiles, un ejemplo de una regla de asociación sencilla y de la mas utilizadas es la siguiente: "Si X es A entonces Y es B" donde X y Y son las variables categóricas y A y B sus valores respectivamente, como podemos ver la creación de estas reglas es sencilla, pero para poder medir que tan precisa es la regla, se utilizan varios tipos de mediciones las mas comunes son precisión y soporte, la precisión nos habla de cuantas veces se repite la regla, es decir, cuantas veces ocurre que si X = A sucede Y = B y el soporte habla de cuantas veces aparecen estos casos, en comparación con todas las combinaciones posibles, es decir, si X puede tomar mas de un valor al igual de Y, cuantas veces ocurre la combinación mencionada con respecto a otras combinaciones posibles, esto nos dice que tan frecuentemente ocurre la regla, por ejemplo volviendo al caso de las compras en linea, podríamos estar interesados en crear un regla que nos indique si alguien compra un producto A, que tan posible es que compra también el producto B y supongamos que la información en nuestra base de datos indica que el 80 % de las personas que compran el producto A también compran el producto B y del total de personas en nuestra base de datos esto lo hace el 30 % de los clientes, por lo tanto la regla: "Si compra A entonces compra B" tendría una precisión del 80 % y un soporte del 30 %.

En esta etapa es posible que se pruebe con distintos algoritmos para al creación del modelo y en algún punto se tenga que regresar para cambiar el formato de los datos y modificar los mismos, para poder aplicar ciertos algoritmos, no existe un algoritmo correcto o único para todo tipo de problemas, siempre se debe probar con varios para obtener al final la decisión de cual es el mas adecuado para el tipo de problema y resultados que esperamos obtener, la forma de entrenar los algoritmos predictivos, es usualmente, dividiendo la información total en dos archivo uno para entrenamiento y otro para prueba y la partición usualmente utilizada es 70 % y 30 % respectivamente, pero puede variar según las necesidades del problema y el algoritmo, en los descriptivos usualmente no es necesarios dividir el conjunto, ya que buscamos evidencias y entre mayor información se tenga, más robustas serán las evidencias obtenidas. El resultado final de esta etapa, es un modelo (ecuación, algoritmo, reglas de asociación, pesos, scores, etc) que describe o contiene los patrones encontrados y es capaz de encontrarlos o predecirlos en información nueva (en el caso de modelos predictivos).

2.1.5. Evaluación e Interpretación

Existen muchas formas de medir o evaluar a un modelo creado por los algoritmos de minería de datos, los modelos predictivos usualmente son medidos por desempeño, esto es, mide el número de aciertos (predicciones correctas de la clase) hechas por el modelo creado encontradas al aplicar el mismo en un archivo de prueba, ejemplos de esto es la matriz de confusión y el área bajo la curva ROC, técnicas típicas para medir el desempeño de un modelo creado con el objetivo de predecir o clasificar, aunque esta es la forma mas simplista de medir un modelo, ya que la minería de datos y el descubrimiento de conocimiento van mas álla de tan solo predecir o clasificar bien, tanto para modelos predictivos como descriptivos, algo de mayor valor es encontrar nuevo conocimiento útil y novedoso, el cual además pueda ser explicado y comprensible para quien desee utilizarlo, otra parte importante de la minería de datos actualmente, es dar o describir una serie de atributos propios de las instancias pertenecientes a la clase u objetivo de estudio, lo cual es llamado "perfil" o "perfilamiento" este consiste en describir una serie de características únicas pertenecientes al objeto de estudio las cuales hacen al mismo objeto mas proclive o probable de pertenecer a la clase (objetivo de estudio), por ejemplo para el caso de la diabetes podemos crear el siguiente perfil: edad mayor de 40 años, con sobrepeso u obesidad, alimentación alta en grasas y azucares, poco ejercicio y padre y/o madre diabetico(s), lo anterior es un perfil donde las características o variables propias de la instancia de estudio (personas) como son: edad, indice de masa corporal, tipo alimentación, actividad física y herencia son las variables (características) que sirven para crear en este caso un perfil de la personas propensas a padecer la enfermedad.

Como se menciono anteriormente, para validar el resultado del modelo, en el caso de modelos predictivos, es necesario contar con un conjunto de entrenamiento y otro de prueba, la forma mas sencilla es dividir la información total con la que se cuenta en porcentajes, puede ser 50 a 50 para entrenamiento y prueba u otras variaciones (60 a 40, 70 a 30, 80 a 20, etc..) siempre teniendo igual o mas cantidad en el archivo de entrenamiento, cuando no es posible dividir de esta forma, por que la cantidad de información disponible sea muy poca, entonces se pueden utilizar técnicas como validación cruzada o re-muestreo (bootstrapping).

En validación cruzada los datos se dividen de forma equitativa en dos archivos iguales y el proceso es el siguiente: primero se entrena el modelo con el archivo 1 y se prueba en el 2 y viceversa se entrena el modelo con el archivo 2 y se prueba en el 1, finalmente se construye un modelo con todos los datos y se prueba de igual forma en todos, con esto se obtendrán promedios del error encontrado en cada archivo de prueba para estimar de igual forma el error general del modelo, existen variaciones de esta tecnica las cuales consiste en hacer mas de una partición haciendo hasta n particiones de los datos y el mismo número de procesos entrenamiento/prueba para obtener, al final, el promedio de error de estas n pruebas. El re-muestreo tiene el mismo proceso que el anterior, al aplicar entrenamiento y prueba sobre varios conjuntos y obteniendo al final el promedio del error de estos archivos de prueba, la diferencia consiste en como se seleccionan estos archivos, a diferencia de la técnica anterior, en esta no se hacen particiones del total de los datos, si no que, las instancias pertenecientes a cada archivo (entrenamiento y prueba) se seleccionan al azar, donde puede haber repetición en la selección hasta conformar el archivo de entrenamiento dejando para el archivo de prueba el complemento de las instancias que no están en el archivo de entrenamiento.

La interpretación de los resultados obtenidos, es tal vez, la parte mas importante en esta etapa, los números no significan mucho si no se tiene un buen entendimiento del

área de aplicación del modelo y el problema que se esta analizando, es aquí donde un buen modelo requiere del apropiado conocimiento y un buen entendimiento del dominio, por ejemplo usualmente en problemas médicos y análisis de enfermedades y en general en problemas reales, la clase o complemento suelen ser relativamente pequeñas en en comparación con la otra, veamos el ejemplo de la diabetes para los datos de la encuesta de ENSANUT 2006 alrededor del 8 % de las personas entrevistadas tienen diabetes y el 92 % no, por lo cual el mejor modelo que se podría crear seria un clasificador el cual nos pronosticara que ninguna persona tendrá diabetes y desde el punto de vista de la medidas numéricas utilizadas para definir el desempeño del modelo este tendría un desempeño del 92 % de predecir correctamente, lo cual suena maravilloso, pero si hacemos una buena interpretación del problema de salud que significa la diabetes, nos daremos cuenta que el alto costo por dejar de predecir ese 8 % de la población, la cual tendrá alto riesgo de padecer diabetes, es muy alto, debido a que múltiples reportes del sector salud en México han informado que dedicando 15 % del presupuesto total al tratamiento de pacientes con diabetes y esto sigue en aumento año con año, por lo cual, en este tipo de modelación lo importante no es crear un algoritmo con la mayor precisión posible, si no aquel que nos de un mejor ranqueo y perfil de las personas con mayor riesgo a padecer esta enfermedad y así poder intervenir a tiempo en aquellas de mayor riesgo.

2.1.6. Difusión y Uso

Esta es la ultima parte del proceso de minería de datos y tal vez una de las mas importantes, desde el punto de vista comercial, económico y productivo, de nada sirve crear un buen modelo y tener un buen entendimiento del mismo, con una perfecta identificación de los perfiles y predictores (variables de alto peso a la hora de predecir la clase), si no es posible sacarle provecho aplicando el modelo a los problemas reales, para obtener beneficios ya sean económicos, en salud, mercantiles o para crear intervenciones que permitan mejorar el panorama de la situación presente.

Existen muchas formas de aprovechar el conocimiento generado por los modelos de minería de datos, una de ellas es a través de los perfiles y esta tal vez sea la mas interesante, ya que no solo le permite a los expertos en el área diseñar una politica de detección temprana, por ejemplo en los casos de salud y enfermedades, si no también crear políticas de prevención mediante el diseño y creación de intervenciones, a partir de los perfiles de riesgo encontrados para las personas propensas a padecer determinada enfermedad, por ejemplo, en el caso de la diabetes si encontramos que las mujeres son mas propensas a padecer diabetes, el experto en salud podría lanzar camapañas dirigidas a ellas para prevenirlas sobre el riesgo alto en el que se encuentran, igualmente podría pedir a los médicos hacer énfasis en las consultas a las mujeres de los altos riesgos que tienen de padecer la enfermedad.

2.2. Algortimos Tradicionales

Una de las ventajas evidentes de la minería de datos es el contar con múltiples técnicas (algoritmos) para lograr su propósito, existen algoritmos basados en teorías estadísticas, probabilistas y de inteligencia artificial entre las mas utilizadas y exitosas, en esta sección describiremos las mas comúnmente utilizadas en el campo de la minería de datos, las cuales además son las que ofrecen mejores resultados.

2.2.1. Regresión

Este algoritmo es utilizado para tareas predictivas y es una de las herramientas estadísticas mas fuertes utilizada por la minería de datos, su tarea es aprender una función real, lo cual es la principal diferencia con una tarea de clasificación, el valor a predecir por el algoritmo es un valor numérico. Al entrenar este tipo de algoritmo su objetivo es minimizar el error (el mas comúnmente utilizado es el error cuadrático medio) entre el valor predicho por el algoritmo y su valor real. Por lo general la formula para crear una **regresión lineal** es del tipo 2.1

$$y = a + a_0 * x_1 + ... + a_n * x_n \tag{2.1}$$

donde x_i son llamados predictores, dado que son los que varían la variable dependiente (y) para aproximarla a la nube de puntos y al final obtener el menor error total para todas las aproximaciones a los puntos que conforman la nube (representación de una función real), es natural que no todos los problemas a modelar a través de una regresión lineal se comporten o tengan una representación en su nube de puntos del tipo lineal, cuando nos enfrentamos a uno de estos casos los atributos x_i de la ecuación 2.1 pueden ser sustituidos por funciones (cuadrados, potencias, logaritmos, senos, cosenos, inversa, etc..) con lo que la ecuación 2.1 quedaría de la siguiente forma

$$y = a + a_0 * f_1(x_1) + ... + a_n * f_n(x_n) \tag{2.2}$$

la cual es conocida como **regresión no lineal**. la cuales pueden minimizar el error en patrones de nube de puntos no lineales.

Para crear la función de regresión ya sea lineal o no lineal se requiere de conocer los coeficientes a_j, por lo tanto el algoritmo en su fase de entrenamiento tratara de encontrar los valores de los coeficientes a_j que minimizan el error entre la y_i calculada y la real, en cada uno de los puntos i, el entrenamiento es como cualquier algoritmo de minería de datos, se elige un porcentaje de los datos de forma aleatoria para entrenamiento y se reserva el resto para prueba (lo usual es 70 % y 30 % respectivamente), una vez calculados los coeficiente en la fase de entrenamiento, se procede a probar la función obtenida en los datos de prueba y se calcula su error, entre menor sea el error mejor es la aproximación encontrada.

2.2.2. Reglas de Asociación

Este tipo de algoritmo es utilizado para tareas descriptivas de la minería de datos y su objetivo es identificar las relaciones existentes entre un conjunto de atributos y sus valores, relaciones que no son obvias a simple vista, los atributos pueden ser categóricos esto proporciona un plus a esta técnica. El objetivo de las reglas de asociación es difundir los patrones inherentes a un conjunto de datos, esto nos permite conocer mejor el problema que se describe en los datos lo cual nos ayuda a la toma de decisiones. Estas reglas son creadas de la forma "SI .. algo se cumple .. ENTOCES .. otra cosa se cumple tambien".

El aprendizaje en los algoritmos típicos de este tipo esta basado en dos parámetros: cobertura y confianza, es decir el algoritmo en su fase de aprendizaje busca todas las reglas definidas tipo "**SI** X_i AND Y_j AND ..... **ENTONCES** Z_k" donde X_i, Y_j, Z_k son oraciones o sentencias del tipo: comprar producto x, binarias (comprar, tomar, existe, falla) o temperatura alta o presión menor de; las cuales cumplan con magnitudes mínimas de cobertura y confianza especificadas previamente antes del aprendizaje del algoritmo.

2.2.3. Técnicas Probabilísticas

Este tipo de algoritmos son utilizados para tareas de clasificación, pero esta técnica va un poco mas allá, esto es, se pueden utilizar cuando no solo se desea hacer una clasificación simple de pertenencia binaria (esta en la clase o no), si no cuando se desea conocer también cual es la probabilidad de pertenecer a determinada clase, esto es muy útil ya que no solo se predice quien pertenece a la clase, si no también la probabilidad de pertenencias y al tener las probabilidades, se puede decir, para cada uno de los objetos el grado de riesgo de pertenecer o no a la clase, en muchos problemas se desea no solo conocer si estará o no en la clase, si no el grado de probabilidad de pertenecer a la misma, sobre todo en casos de estudio como el presentado en esta tesis, las enfermedades, cuando se crean modelos para este tipo de problemas, el interés no se centra en si padecerá o no la enfermedad determinada persona, si no en si puedo prevenir que la padezca, esto se logra creando perfiles de riesgo para todas las personas, donde al asignar una probabilidad tienes la oportunidad de saber que personas por ahora sanas pueden en un futuro padecer la enfermedad, por tener una probabilidad alta de acuerdo con el modelo, esto es de gran valor y ayuda a la prevención y creación de intervenciones.

De las técnicas probabilisticas, el método mas utilizado es el **Naive Bayes**, el cual esta basado en el teorema de bayes, como su nombre lo dice (ingenuo) basa su funcionamiento en asumir la independencia entre si de los atributos, dada un clase en especifico, la ecuación que representa a este algoritmo es al siguiente:

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)} \tag{2.3}$$

donde x es un vector de atributos asociados a las clase, c es la clase u objetivo a predecir, P(c) es la probabilidad de que ocurra la clase o probabilidad de la clase, $P(c|x)$ es la probabilidad de que ocurra c dado el vector de atributos x, lo cual es básicamente el pronostico de probabilidad de pertenencia a la clase, $P(x|c)$ es la probabilidad de que ocurra el valor del atributo del vector x dado que pertenece a la clase, esto también puede ser llamado como evidencia y se calcula para cada uno de los atributos del vector x en fase de entrenamiento y por ultimo $P(x)$ es la probabilidad de que ocurra cada uno de los atributos del vector x, este es una constante y por lo tanto puede ser descartada de la ecuación cuando esta se normaliza, ya que es independiente de la clase c. Esta técnica es la utilizada en este trabajo por lo ya explicado anteriormente y ampliaremos en su explicación y uso en el capitulo 3.3.

Otra técnica no muy utilizada pero si útil es el **Naive Bayes Generalizado o Aumentado**, este es un esfuerzo por omitir la parte "ingenua" del algoritmo por considerar que no todas las variables son independientes entre si, este trata de encontrar un punto medio, este punto medio es considerar las dependencias mas fuertes que existan entre los atributos para incluirlas en el calculo o creación del modelo, esto permite al algoritmo tener un mejor desempeño en conjuntos de datos donde existan correlaciones fuertes entre algunos de sus atributos, los cuales serian omitidos por el Naive Bayes simple, existen muchos desarrollos, diseños y propuestas de este tipo de algoritmos en la literatura, sin poder elegir alguno que este por encima de todos, ya que su funcionamiento depende mas de la forma de encontrar las correlaciones fuertes, por lo que algunos funcionan muy bien para ciertos tipos de problemas y otros mejor para el resto de esos problemas, un análisis mas extenso y una propuesta del mismo se presentan mas adelante en el capítulo 5.

Otro tipo de clasificador probabilistico son las **redes bayesianas**, este algoritmo consiste en representar el conocimiento contenido en los datos en forma de un grafo dirigido, donde los nodos son los atributos y las inter-conexiones entre estos son las relaciones de

dependencia o independencia entre los atributos, estas relaciones de dependencia ademas son cuantificadas, por lo que su valor expresa que tan "fuerte" es la relación entres estas siendo 0 representación de una independencia total entre las variables, el valor de la fuerza de la relación entre los atributos viene dado por distribuciones de probabilidad del tipo: $P(X_i|P_a(X_i))$ donde X_i es el vector de atributos y $P_a(X_i)$ representa a los padres de los atributos X_i representados en el grafo. Por lo tanto un red bayesiana es una tupla $RB = (G, \theta)$, donde G es el grafo y θ es el conjunto de distribuciones, como se muestra en la figura 2.2, la cual es un ejemplo de una red bayesiana creada para predecir quien comprara un auto.

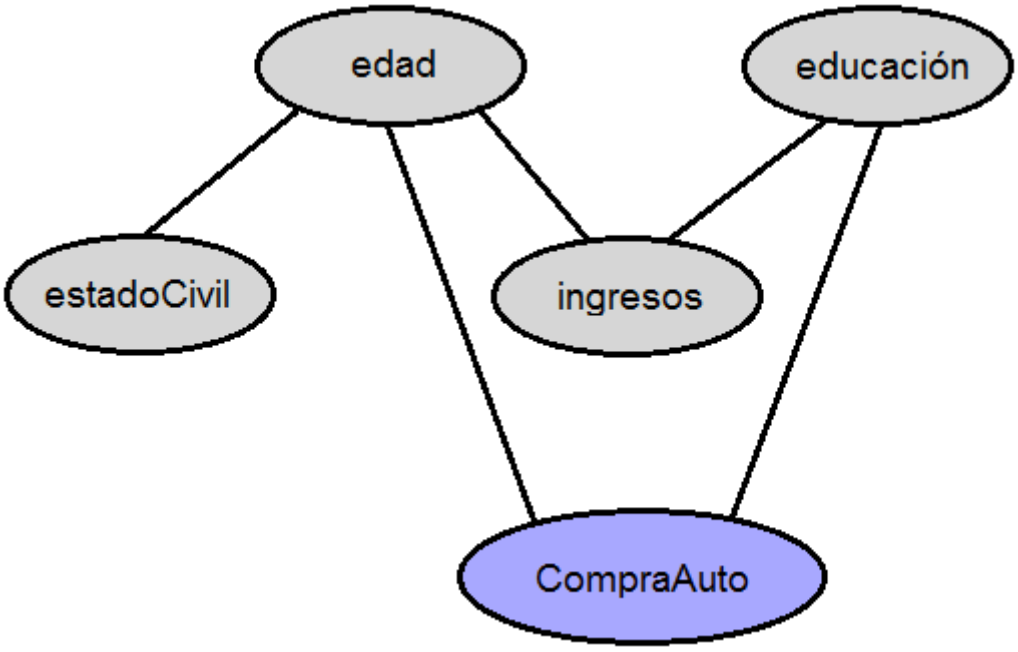


Figura 2.2: Ejemplo de un grafo de red bayesiana

2.2.4. Árboles de Decisión

Este tipo de algoritmos son muy versátiles ya que pueden ser utilizados en tareas de clasificación, agrupamiento y regresión, ademas se pueden utilizar en problemas donde los atributos están mezclados en categóricos y numéricos, pueden ser representados con un gráfico tipo árbol donde se parte de un nodo raiz el cual puede ser la pregunta a responder o problema a resolver y se continua de forma jerárquica a traves de cada uno de los nodos hasta llegar el final donde estará la respuesta o solución al problem, cada nodo significa tomar una decisión por lo cual en cada nodo puedes cambiar de dirección dependiendo de cada decisión, los nodos están inter-conectados con arcos los cuales cada uno representa un posible valor del atributo que representa cada nodo, cada uno de los arcos esta conectado a otro nodo o a una hoja en cuyo caso con esta ultima se llega a la solución del problema, en este caso podemos decir que el algoritmo de el resultado de su clasificación o pronostico de clasificación.

Los arboles de decisión pueden ser considerados también como otra metodología tipo de aprendizaje de reglas ya que cada una de las ramas puede ser interpretada como una regla donde se toman decisiones, un ejemplo de árbol de decisión se muestra en la figura

2.3 donde el arbol fue entrenado en su etapa de aprendizaje para predecir si se debe jugar o no el encuentro de un partido de algún deporte (puede ser baseball, soccer, etc..) o no, de acuerdo con la información disponible el árbol de decisión obtuvo el modelo presentado en la figura, al final el algoritmo es capaz de predecir si se debe o no jugar el partido.

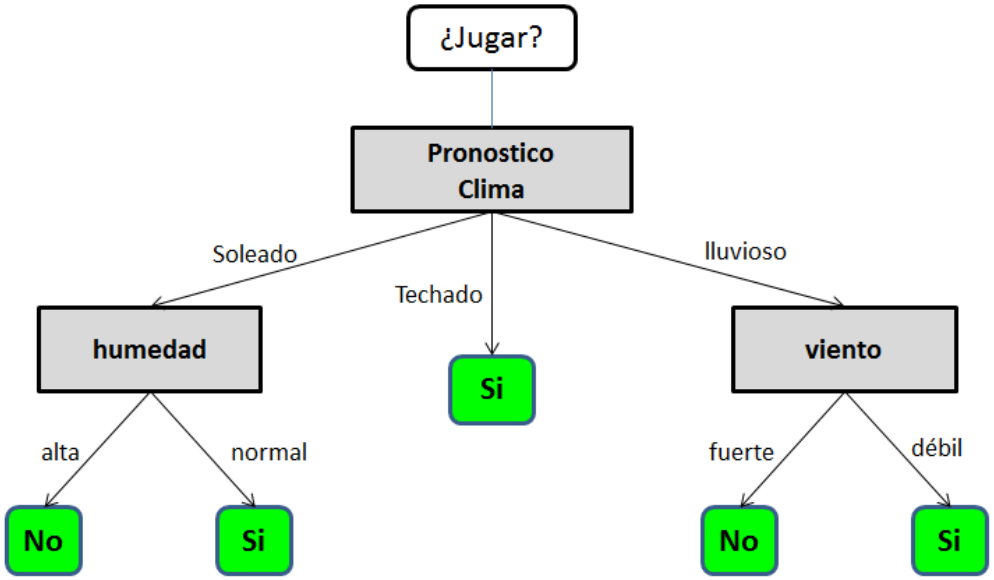


Figura 2.3: Ejemplo de árbol de decisión entrenado para determinar si se debe jugar o no partido de algún deporte [8]

Los algoritmos mas utilizados y famosos para la creación de arboles de decisión son el CART [21], ID3 [22] y C4.5 [23], la parte importante para la creación de un modelo basado en arboles de decisión, es definir el número de particiones a considerar (nodos) y como seleccionar estas particiones, las 3 técnicas mas utilizadas y mencionadas anteriormente, proponen diferentes métodos para obtener estas características y en si cualquier otro algoritmo de arboles de decisión basara su innovación en esta parte, no es propósito de este documento ahondar en el tema, por lo que si es de su interés puede referirse a la bibliográfia.

2.2.5. Redes Neuronales

Este tipo de algoritmos son utilizados para clasificación, y al igual que los arboles de decisión, su representación y toma de decisiones es en estructura de grafo dirigido, pero la toma de decisiones en cada nodo es mucho mas complejo y muchas veces oculto, debido a que en la fase de entrenamiento son creados de forma automática los nodos y las inter-conexiones (arcos) entre ellos, dando así la sensación de una caja cerrada, donde solo conocemos la entrada y la salida, pero no el proceso que se lleva al interior de la misma.

Este tipo de algoritmo fue creado con el estudio de la inteligencia artificial, y es una emulación al proceso que se lleva acabo en el cerebro humano entre las neuronas y sus conexiones, así como la toma de decisiones. Su estructura esta conformada como se muestra en la figura 2.4, se cuenta con una capa de entrada, donde a lo máximo habrá un nodo por cada atributo del problema a resolver, puede tener una o n capas ocultas (aunque se ha visto en muchos ejemplares que con hasta 3 es suficiente para obtener buenos modelos, poner mas capaz ocultas no beneficia mucho el desempeño y si afecta el tiempo de procesamiento), y una capa de salida, la cual puede tener un solo nodo, el cual indicara la

pertenencia o no a la clase. Una de sus limitantes es que solo trabaja con datos numéricos, por lo cual cuando se desea utilizar datos categóricos, estos deben ser transformados primeramente a una version numérica representativa de cada uno de sus valores categóricos.

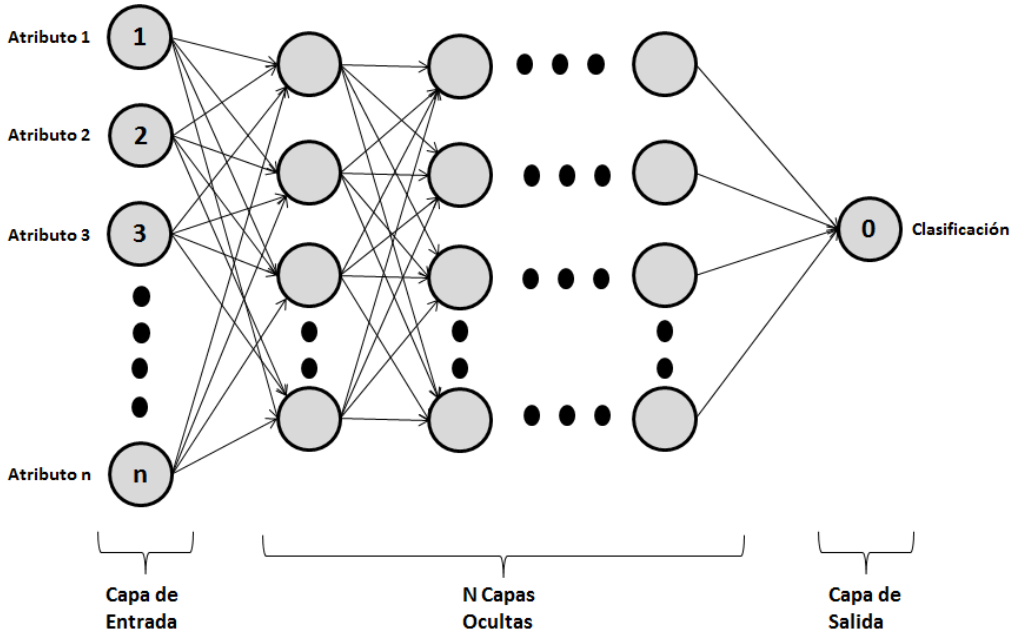


Figura 2.4: Ejemplo de una red neuronal general

Este tipo de algoritmos son de aprendizaje automático, es decir se presentan los datos al mismo y este calcula los pesos entro los arcos que inter-conectan todos los nodos, como ya se menciono, estos valores están ocultos, no se sabrá su valor, ni el número de conexiones entre los nodos dentro de las capas ocultas, el objetivo del algoritmo es calcular todos estos y los nodos dentro de las capas ocultas, los cuales minimicen el error de clasificación a la salida, con respecto de los datos de entrenamiento. Una de las desventajas de este tipo de algoritmos es su nula habilidad descriptiva, dado que son cajas negras donde no se puede interpretar el tipo de conexiones internas, y por lo tanto no se puede inferir un patrón, o crear reglas o perfiles a partir de estos modelos, por otra parte son bastante robustas y de desempeños generalmente mejores a cualquier otra técnica de la minería de datos. Los algoritmos mas comúnmente utilizados son las redes perceptrón simple, el cual cuenta con una capa de entrada, de uno o varios nodos, y una capa de salida, con uno o mas nodos, en este algoritmo no existe capa oculta; y el mas utilizado el perceptrón multicapa, el cual puede contar con mas de una capa oculta, además de la capa de entrada y la de salida, otro concepto importante es la forma en que los pesos entre las capas son calculados, la técnica mas utilizada para esto es la que se llama retropropgación (backpropagation). Una vez mas no ahondaremos en este técnica, ya que no la utilizamos en este trabajo, en la literatura existen muchas referencias y ejemplos para estos tipos de algoritmos.

2.2.6. Agrupamiento

El método de este tipo de algoritmos esta basado en instancias o ejemplos a diferencia de los algoritmos vistos anteriormente, en su fase de entrenamiento, este tipo de algoritmo no termina su aprendizaje, si no que reserva cada instancia (ejemplar) distinto en su memoria, así que cuando va procesando cada una de las instancias, estas son comparadas con las existentes para ver si es posible encontrar algunas parecidas y ponerlas juntas, lo cual

logra mediante una métrica de distancia calculada entre cada una de las diferentes clases de instancias que tiene almacenadas en su memoria, si la métrica es lo suficientemente pequeña o cercana (lo cual es definido como parámetro para la ejecución del algoritmo) a una determinada instancia, el nuevo ejemplar pasara a formar parte de ese grupo (o clase), en caso contrario, esta formara su propio nuevo grupo y así con cada instancia contenida en los datos, cabe mencionar que también a diferencia de los algoritmos de clasificación vistos anteriormente en este documento, en este caso se desconoce la clase u objetivo de clasificación, tanto en fase de entrenamiento como de prueba, es mediante el uso del algoritmo que se debe determinar la clase a la que pertenece cada instancia y el número de estas que existe.

En la figura 2.5 se muestra la evolución de un algoritmo de agrupamiento (k-means), en este caso el algoritmo parte colocando centroides al azar (el usuario puede indicar al algoritmo cuantas clases considera que existen dentro de los datos, para tener un punto de partida), en el caso de la figura 2.5 se colocan 3 centroides, lo cual nos indica que se buscaran agrupar las instancias en 3 clases, conforme el algoritmo avanza y mas instancias son presentadas al mismo, se ve la evolución de los grupos y el desplazamiento de los centroides a un lugar mas adecuado para los mismos, hasta que al final podemos ver como los tres grupos se encuentran bien definidos, así cuando una nueva instancia requiera de ser clasificada el algoritmo sabrá a que grupo se asemeja mas y la colocara en dicha clase.
Los algoritmos más utilizados de agrupamiento (clustering), en la minería de datos son

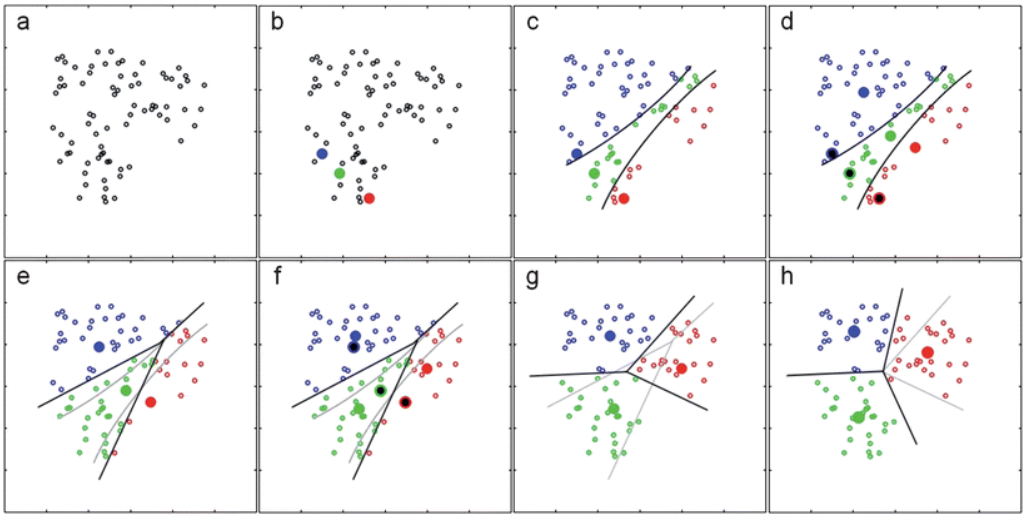


Figura 2.5: Ejemplo del proceso de un algoritmo de agrupación [24]

los siguientes: “Mapas auto-organizados de Kohonen”, la idea principal de este algoritmo esta basada en las redes neuronales maneja el concepto de capas, K-medias es de los mas utilizados, se basa en la teoría de vecindad, es el algoritmo típico de agrupamiento, su funcionamiento es estándar con los primeros datos, se colocan los centroides y posteriormente decide cuales son las instancias mas cercanas a cada uno de los centroides (clases) definidas para asignarlo a la que pertenece, K-vecinos es una variante también de estas técnicas de agrupación por vecindad, esta es la técnica mas sencilla y a la vez las mas errónea por la simplicidad con que decide la pertenencia de las instancias a cada uno de lo grupos y Redes de Cuantización Vectorial (LVQ), es un algoritmo basado en redes neuronal al igual que Kohonen . Como ya lo mencionamos anteriormente no ahondamos tampoco en estos algoritmos por no ser parte de las técnicas utilizadas en el mismo.

2.3. Algortimo de Selección de Características

La selección de características es una herramienta muy importante, utilizada en la modelación actual. El Clasificador Naive Bayes se beneficia enormemente de esta herramienta, tanto en aumento de desempeño como en optimización en procesamiento. En general se ha encontrado que una buena pre-selección de las variables que describen el fenómeno a modelar, debe eliminar aquellas variables con un alto grado de correlación [9], lo cual ayuda al algoritmo de NB a tener un mejor desempeño y precisión a la hora de clasificar, porque se eliminan variables, que de otra forma podrían ser redundantes y con esto dar el doble de peso a dichas variables desequilibrando la clasificación del modelo[9][10], también permite la reducción del número de variables que participan en el modelo, lo cual en modelos de predicción o clasificación de fenómenos con un número muy grande de variables a analizar, se traduce en ahorros significativos en tiempo de ejecución, como es en predicción de texto[11][12], clasificación de spam[13] y reconocimiento de escritura[14], entre otros.

Se han desarrollado muchas técnicas de selección de características combinándolas con el clasificador NB[9][10][11][12][13], las cuales son muy buenas, pero parecen particularizar el uso a la modelación de determinados fenómenos o aplicaciones especficas, la herramienta de interés para este proyecto, es la que pueda seleccionar características a través de cumplir con múltiples restricciones u objetivos, dado que no se desea predisponer el aprendizaje del clasificador NB, por discriminar variables haciendo consideraciones teóricas, utilizando la experiencia y conocimientos actuales del fenómeno con lo que se pueda contar, si no que es preferible dejar a los datos "hablar", y sea el mismo modelo quien descubra la información útil para lograr la mejor predicción y desempeño. Existen múltiples trabajos elaborados hasta la actualidad, con variaciones entre ellos para crear el algoritmo de selección de características que mejor se adapta a las necesidades de cada problema, por lo cual se opto por hacer lo mismo, crear un algoritmo de selección de caractirísticas, el cual es adecuado para el tipo de fenómenos que deseamos analizar este algoritmo tiene mucho en común con todos los existentes, salvo por el algoritmo genético utilizado para optimizar la selección de estas características, la mejor forma para nuestro propósito es utilizando un algoritmo evolutivo multi-objetivo (MOGA), para hacer la selección de caractersticas, al mismo tiempo que se optimizan múltiples parámetros (la mayora solo optimizan a lo más 2 parámetros a la vez)[14][16][17] [18][19], en estos trabajos utilizan distintas herramientas de la minera de datos como son: redes neuronales[17][18], clúster[14] y modelación difusa[18] pero ninguno junta los dos factores, seleccin de caractersticas con MOGA y clasificador NB.

Otros trabajos proponen múltiples objetivos a optimizar que puedan ser utilizados de forma general independientemente del clasificador[16][19]. En este trabajo proponemos particularmente optimizar el desempeño de un clasificador NB y encontrar las mejores variables que describen el fenómeno, utilizando selección de características con MOGA. Los trabajos publicados mencionados han empleado diversos MOGA para el algoritmo de selección de características, siendo los ms populares el: NSGA[14][16][17] y MOGA[19], estos MOGA suelen tener problemas a la hora de optimizar más de dos características, por lo cual en este trabajo optamos por utilizar un MOGA de reciente creación, el RankMOEA[20] que se describe como un mejor algoritmo comparado con los anteriores, puede manejar de buena forma más de dos objetivos a optimizar al mismo tiempo y además ofrece un espacio de soluciones a lo largo del frente de Pareto, lo que permite dar múltiples soluciones validas, dejando al modelador la decisión de tomar las soluciones que considere mejores o adecuadas para su objetivo.

Capítulo 3

Un Marco Explicito Simple

En este capítulo analizaremos los distintos métodos existentes para seleccionar las variables que participaran en la creación de un modelo determinado, así como estadísticas para determinar la importancia e influencia relativa, que tiene el valor de una variable sobre una determinada clase, para este último solo utilizaremos algo llamado épsilon ε [25] debido a que es una métrica la cual conocemos y utilizamos muy frecuentemente, por lo cual se nos facilita el entendimiento profundo de los resultados arrojados por esta, y también métodos automatizados de selección de variables, los cuales están ligados a distintas métricas de desempeño para la elección de las variables útiles o calificadas para participar como elementos importantes del modelo y no simples variables de relleno o ruido.

3.1. Épsilon

Es una métrica probabilística utilizada en modelos del mismo tipo, aunque no limitada a utilizarse en otro tipo de algoritmos, ya que es independiente del tipo de algoritmo usado para la creción del modelo, épsilon ayuda a cuantificar la importancia, injerencia, relevancia e impacto que tiene una variable y sus valores sobre el objetivo, el cual se desea pronosticar (clase), de forma cuantitativa. Esto tiene una ventaja dado que al ser un número es facil decidir cuales combinaciones de variable/valor están asociadas a la clase y al ser numérico incluso nos cuantifica el tamaño de esa relación, entre mas grande es el valor, mayor es la relacíon con la clase que tiene esta variable/valor.

La siguiente ecuación muestra como se obtiene el valor de épsilon para cada una de estas combinaciones de variable/valor

$$epsilon_{X_i} = \frac{N_{X_i}(P_{CX_i} - P_C)}{\sqrt{N_{X_i} * P_C(1 - P_C)}} \tag{3.1}$$

donde:
N_{X_i} es la población total con variable x y valor i
N_{CX_i} es la población total que está en la clase y tiene la variable x con valor i
N_C es la población total en la clase
N es la población total
$P_{CX_i} = \frac{N_{CX_i}}{N_{X_i}}$
$P_C = \frac{N_C}{N}$

Una de las ventajas de esta métrica es que funciona, tanto en variables categóricas, como numéricas, aunque en el caso de variables numéricas, lo más conveniente es crear

grupos (categorías) para obtener mejor información de la variables y sus valores, de otra forma la información estaría muy disgregada a través de los muchos valores posibles de la variable numérica, también puede trabajar con variables discretas y continuas, lo cual nos permite aplicar esta métrica en cualquier tipo de variable sin restricción.

La parte del numerador puede ser conocida también como "Señal" y el denominador es llamado o interpretado como "Ruido", por lo tanto esta métrica en resumen mide variable por variable y valor por valor, que tan relacionada esta la combinación con la clase o no clase (señal) y que tanto sobresale esta relación comparada con las variaciones derivadas de la muestra (ruido). Típicamente si épsilon es mayor o igual a 2 o menor a -2 se dice que la variable esta suficientemente relacionada con la clase y puede ser tomada en cuenta para participar en la creación del modelo, cualquier valor entre 2 y -2 puede ser descartado como importante en cuanto a su relación entre la variable/valor y la clase en cuestión.

3.2. Metodos de Selección de Características

En la literatura esta ampliamente documentado y estudiado este tema, la conclusión general e importante es que aplicar una técnica de selección de características apropiada ayuda al desempeño de cualquier algoritmo o técnica de minería de datos por reducir el número de variables, las cuales pueden considerarse como no útiles o ruido para el modelo, el problema o el dilema, es elegir la técnica adecuada para la selección de características por que una técnica no apropiada puede empeorar el desempeño del modelo, comparado con usar todas las variables, por eso las selección de la técnica de minería de datos a aplicar en el problema en particular es muy importante, no todas las técnicas sirven para todos los problemas y todos los algoritmos de miner'a de datos.

3.2.1. Análisis de Componentes Principales

Esta es una técnica estadística denominada análisis de componentes principales (PCA por sus siglas en ingles) y es utilizada para reducir la dimensionalidad un conjunto de datos [3]. "El ACP construye una transformacin lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamao del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza ms grande es el segundo eje, y as sucesivamente. Para construir esta transformacin lineal debe construirse primero la matriz de covarianza o matriz de coeficientes de correlacin. Debido a la simetra de esta matriz existe una base completa de vectores propios de la misma. La transformacin que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformacin lineal necesaria para reducir la dimensionalidad de datos. Adems las coordenadas en la nueva base dan la composicin en factores subyacentes de los datos iniciales" [3].

El propósito de esta tranformación es que las nuevas variables $(f_1, f_2, ...)$ se generan de manera que son independientes entre sí y además, se debe lograr que las primeras variables f deben ser las que sean más relevantes, las que contengas más información. En otras palabras f_1 debe ser el atributo ms relevante del nuevo conjunto, con esto se tiene que f_1 es mas relevante que f_2 y f_2 es mas relevante que f_3 y así sucesivamente. Lo anterior permite seleccionar los k primeros atributos garantizando que se seleccionaran los atributos mas relevantes y representativos del conjunto de datos original [15]. Por la tanto, esta técnica más que reducir el número de variables las transforma en un nuevo conjunto del mismo tamao, pero donde podemos asegurar que las primeras componentes principales contienen

la gran mayoría de la información contenida en los datos, es ahí donde se encuentra la reducción en la dimensionalidad del conjunto de datos ya que podemos elegir las primeras componentes principales y con esto tener un representatividad del conjunto de entre 90 y 95 % de la información.

Esta aun cuando es una técnica muy utilizada en la estadística para reducir la dimensionalidad de los datos, para el propósito de este trabajo tiene varias limitantes por las cuales no podemos utilizarla: 1) dado que la técnica involucra el cálculo de la varianza, mínimos cuadrados y proyecciones, esta limitada al uso de variables numéricas, y como veremos a lo largo de este proyecto los fenómenos de estudio abordados en este trabajo incluyen muchas variables categóricas como raza, sexo, tipos de ejercicio, etc. , las cuales si bien es cierto se pueden transformar de categóricas a numéricas esta practica no es muy recomendable de hacer dado que de no elegir la correcta relación entre en número a asignar el valor categórico se estaría destruyendo la integridad de los datos y así destruir las posibilidades de modelar de forma adecuada el fenómeno de estudio; 2) al ser una técnica que involucra la creación de nuevas variables (componentes principales), basado en la muestra o conjunto de datos presentado al algoritmo, se corre el riesgo de que al entrenar el modelo de minería de datos con este conjunto, y tratar de aplicarlo a nuevos conjuntos del mismo fenómeno, existan varios conjuntos atípicos (con diferentes distribuciones en su muestreo) para los cuales las componentes principales no seas las mismas que las utilizadas para entrenar el modelo y esto provocaría que el modelo no fuera valido para nuevos conjuntos de información, aun tratándose de información relacionada con el mismo fenómeno, lo cual rompe con el espíritu de la minería de datos, el cual es crear un modelo a partir de una muestra de información que describe un fenómeno dado y con este modelo predecir sobre cualquier nuevo conjunto de información (muestra) relacionado con el fenómeno y predecir su comportamiento; 3) es una técnica que no se asocia con el desempeño del modelo, es decir las variables no son elegidas para mejorar el desempeño del modelo creado con estas, si no su elección solo depende de reducir la dimensionalidad del problema, lo cual tiene el mismo problema descrito para la técnica abordada en el sub-capítulo siguiente.

3.2.2. Selección de Características por Atributo

Épsilon es una de estas técnicas para selección de características basada en la teoría de probabilidades y el Naive Bayes, lo que la hace una técnica más adecuada para nuestro trabajo, al ser probabilistica al igual que las técnicas desarrolladas en este trabajo y el propio Niave Bayes, otras técnicas muy utilizadas en la literatura son: Ganancia de Información (Information Gain), Razón de Ganancia (Gain Ratio), Incertidumbre Simétrica (Symmetrical Uncertainty), Relief-F, ONE-R y Chi-Cuadrada, todos los cuales son analizados por Novakovic [9] y comparados para desempeño en NB, en el articulo se puede ver que para ciertos problemas (conjuntos de datos), el desempeño mejora y para otros empeora comparados con utilizando el conjunto completo de variables, esto confirma lo que mencionamos anteriormente, ninguna técnica de selección de características es la mejor para todos los problemas, por lo cual la elección de la técnica a utilizar depende mas del tipo de problemas a analizar, el entendimiento que tengamos sobre ella, en cuanto a su funcionamiento, la interpretación adecuada de los resultados obtenidos, y lo familiar que resulte para quien la usa, en nuestro caso estas son las razones por las cuales decidimos utilizar épsilon a lo largo de este trabajo sobre algún otro tipo de técnica.

Tanto épsilon como el resto de las técnicas mencionadas anteriormente para selección de características, son denominadas de “evaluación de variable”, esto es por que analizan variable por variable y valor por valor de forma independiente, al desempeño del modelo, interesándose solamente por la variable analizada en cuestión y su relación con la clase,

todas estas técnicas miden de una forma u otra la "fuerza" de esa relación y la convierten en una cantidad numérica interpretable y por ende discriminante, es decir se puede colocar un umbral a partir del cual las variables califican para ser utilizadas o no en el modelo, de esa forma, discriminando entre lo que se puede considerar como variables, las cuales solo aportan ruido al modelo y las que aportan señal (discriminación) al mismo.

3.2.3. Selección de Características por Desempeño

Otras técnicas de selección de características, son aquellas las cuales van directamente dirigidas al desempeño, esto es se busca analizar el mayor número posible de combinaciones de variables con la cuales se obtenga un mejor desempeño del modelo, resultando la de mejor desempeño la combinación ganadora y por ende el resultado de la selección de características, la forma de hacerlo es creando un vector, donde cada elemento del vector es un valor binario, el cual representa a cada una de las variables (atributos) con los que se cuenta en los datos para la creación del modelo, el valor binario, el cual puede encontrarse en cada elemento del vector (en realidad cada elemento del vector representa a cada una de las variables disponibles), es 1 y 0 o Si y No o Falso y Verdadero o etc., es decir una bandera que permita al código, el cual ejecuta el algoritmo de minería de datos, saber si utilizará o no dicha variable y así para el conjunto completo de todas las variables, al final se obtiene un subcojunto de variables elegidas de alguna forma para entrenar el modelo y obtener un desempeño, después se crean un nuevo vector de valores binarios diferentes al primero y se calcula su desempeño y así sucesivamente se puede hacer N veces o hasta cumplir todas las combinaciones posibles, al final se utiliza la combinación de variables con el vector de mejor desempeño.

Cuando tenemos pocas variables podemos ir a todas las posibles combinaciones de variables y calcular sus desempeños, y así elegir con certeza la mejor combinación, pero cuando vamos a números mas grandes de variables, la cantidad de posibles combinaciones en este caso es de N^2, lo cual es muy costoso en términos computacionales si tratamos de buscar en todas y cada una de las posibles combinaciones, esto es debido a que por cada vector (combinación) es necesario aplicar el algoritmo de minería de datos y obtener su desempeño, por lo cual un algoritmo de selección de características con búsqueda exhaustiva para seleccionar el conjunto de variables que produzcan el mejor desempeño no es inteligente.

Esto es por lo cual, las técnicas con mayor éxito en la utilización de esta teoría, son aquellas las cuales utilizan los algoritmos genéticos y sus bondades para realizar búsquedas "inteligentes", evitando así hacerlo de forma exhaustiva, para encontrar la selección de caracteristicas (vector) de mejor desempeño, además podemos, dependiendo del problema en el cual estamos trabajando, requerir de más de un parámetro a optimizar, parámetros que funcionarían o nos darían diversas formas de medir, ya sea el desempeño del modelo o métricas asociadas con las variables seleccionadas, para lo cual podemos utilizar los Algoritmos Genéticos Multi-Objetivo (MOGA por sus siglas en ingles), los cuales tienen las mismas bondades que los algoritmos genéticos estándar, además de la capacidad de optimizar 2 o mas parámetros simultáneamente un su búsqueda, pueden encontrarse ejemplos de las bondades, aplicación y funcionamiento de este tipo de algoritmos para selección de características utilizando MOGAs en los artículos de Morita, Emmanouilidis, Oliveira y Venkatadri [14][16][18][17][19].

Por lo expuesto anteriormente, es que, en este trabajo decidimos crear nuestro propio algoritmo de selección de características, el cual se adapte a las métricas (parámetros)

los cuales deseamos optimizar acorde con el tipo de problemas (datos) con los cuales trabajamos en este documento, este se presenta en el capítulo 4, donde se podrá notar que mas allá de inventar una nueva técnica, algoritmo o teoría, utilizamos las experiencias y teorías existentes en la literatura para crear un algoritmo que se acopla a nuestras necesidades y por lo tanto, las métricas a optimizar y el MOGA utilizado, son diferentes a los mencionados en estos articulos y están dirigidos completamente a ser utilizados en con el clasificador Naive Bayes.

La diferencias entre los algoritmos de selección de características por atributos y por desempeño son claras, el segundo puede ser utilizado de forma automatizada y permitirnos no preocuparnos por las variables elegidas, ya que seguramente son la combinación con la cual obtenemos el mejor desempeño en esa muestra, con el primero la selección de características requiere mas de la intervención humana, para decidir acerca de las variables y su selección, aunque permite tener un mejor entendimiento de todos y cada uno de los atributos, los cuales, participaran en el entrenamiento del modelo, ya que se puede desmenuzar todas y cada una de las variables y "cuestionar" al algoritmo el por que si o no utilizar la variable y al final decidir si usarla o no independientemente del resultado de la técnica, aunque si se desea estas técnicas también podrían ser automatizadas, poniendo un umbral de aceptación o rechazo para cada variable; la segunda técnica tiene un costo computacional más alto, debido a que requiere de entrenar un modelo y medir el desempeño del mismo múltiples veces, el primero tiene la desventaja de que al hacer la selección de características, podemos obtener un conjunto de variables, que nos lleve a un peor desempeño comparado con el obtenido por utilizar todas las variables, esto no ocurre con el segundo, dado que el resultado de la selección de variables, al menos sería de igual desempeño comparado con el de todas las variables.

Como podemos ver, no hay mejor ni peor técnica para seleccionar características, solo la apropiada para el tipo de problema, cada una tiene sus ventajas y desventajas, ademas no existe ningún impedimento para usar ambos tipos para la depuración de las variables, podemos usar primero la técnica de selección de características por atributo y posteriormente al sub-conjunto resultante aplicar la técnica por desempeño o viceversa.

3.3. Naive Bayes Estándard

El Naive Bayes Estándar (NBS), es una algoritmo de clasificación ampliamente utilizado por la minería de datos, lo presentamos brevemente en el capítulo 2.2.3, ahora vamos a ahondar en esta metodología, por ser la utilizada durante el desarrollo de todo el proyecto y el presente documento. Como ya mencionamos, esta es una técnica probabilistica basada en el teorema de Bayes el cual nos dice lo siguiente: sea $X_1, X_2, X_3, ...X_n$ un conjunto de sucesos (atributos) mutuamente excluyentes, y sea C un suceso (objetivo, clase) del cual se conocen las probabilidades condicionales $P(C|X_i)$, entonces la probabilidad $P(X_i|C)$ es definida por la ecuación:

$$P(X_i|C) = \frac{P(C|Xi) * P(X_i)}{P(C)} \tag{3.2}$$

donde $P(X_i)$ son las probabilidades a priori, $P(C|X_i)$ es la probabilidad de C en la hipótesis X_i, $P(X_i|C)$ son las probabilidades a posteriori.

De la ecuación 3.2 podemos despejar $P(C|X_i)$ con lo que se obtiene lo siguiente:

$$P(C|X_1, X_2, ..., X_n) = \frac{P(C) * P(X_1, X_2, ..., X_n|C)}{P(X_1, X_2, ...X_n)} \tag{3.3}$$

la expresión muchas veces es interpretada en términos no matemáticos como:

$$Posterior = \frac{Anterior * Probabilidad}{Evidencia} \tag{3.4}$$

De la ecuación 3.3, el numerador es la parte importante para la solución de la misma, el denominador es constante ya que no depende de C, por lo tanto, no es importante para el cálculo, los valores $X_1, X_2, ..., X_n$ son datos, por lo cual siempre se conoce su valor y por lo tanto la probabilidad es conocida y constante, en cambio el numerador es una probabilidad compuesta, la cual depende de C y puede ser expresada por:

$$P(C) * P(X_1, X_2, ..., X_n|C) = P(C, X_1, X_2, ..., X_n) \tag{3.5}$$

y puede ser derivada aplicando la definición de probabilidad condicional repetidamente como sigue:

$$\begin{aligned} P(C, X_1, X_2, ..., X_n) &= P(C) * P(X_1, X_2, ..., X_n|C) \\ &= P(C) * P(X_1|C) * P(X_2, X_3, ..., X_n|C, X_1) \\ &= P(C) * P(X_1|C) * P(X_2|C, X_1) * P(X_3, X_4, ..., X_n|C, X_1, X_2) \\ &= P(C) * P(X_1|C) * P(X_2|C, X_1) * P(X_3|C, X_1, X_2) * \\ &\quad P(X_4, X_5, ..., X_n|C, X_1, X_2, X_3) \\ &= P(C) * P(X_1|C) * P(X_2|C, X_1) * P(X_3|C, X_1, X_2) * ..... * \\ &\quad P(X_n|C, X_1, X_2, X_3, ..., X_{n-1}) \end{aligned} \tag{3.6}$$

como se puede apreciar en la ecuación 3.6, las probabilidades condicionales compuestas se vuelven muy complejas para su cálculo, y cuando el vector X tiene muchos atributos se vuelve casi imposible de calcular y procesar esta probabilidad, además como lo mencionamos en el cap. 1.4.2 “la maldición de la demensionalidad”, aplicar el teorema nos daría instancias únicas, es decir habría 1 o 0 elementos en cada posible combinación, lo cual no sirve de ninguna forma para generalizar y crear un modelo de clasificación. Aquí es donde entra la asunción de independencia entre los atributos de X, lo que al final permite al NBS ser un clasificador muy exitoso y posible de aplicar y generalizar para miles de problemas y datos, se asume que cada X_i es independiente de cualquier otro X_j para $j \neq i$, lo cual significa que $P(X_i|C, X_j) = P(X_i|C)$, con lo cual podemos simplificar la ecuacíon 3.6 como sigue:

$$\begin{aligned} P(C, X_1, X_2, ..., X_n) &= P(C) * P(X_1|C) * P(X_2|C) * P(X_3|C) * ... * P(X_n|C) \\ &= P(C) * \prod_{i=1}^{n} P(X_i|C) \end{aligned} \tag{3.7}$$

con lo cual la ecuación 3.3 puede quedar expresada como:

$$P(C|X_1, X_2, ..., X_n) = \frac{1}{P(X_i)} * P(C) * \prod_{i=1}^{n} P(X_i|C) \tag{3.8}$$

donde C es el atributo a clasificar y $P(X_i)$ es un valor constante, cuando todos los valores de X_i son conocidos.

Ahora la pregunta es: ¿como se vuelve un clasificador el NBS?, si asumimos que en el problema de clasificación solo existen dos grupos posibles, y por lo tanto, solo dos clases,

entonces la clasificación se reduce a decir si la instancia se encuentra en la clase o se esta en la no clase. Lo cual nos permite asumir que un elemento se encuentra en la clase si cumple con la siguiente condición:

$$P(C|X_i) > P(\sim C|X_i) \tag{3.9}$$

Ahora dado que no estamos interesados en una simple decisión de si o no la instancia pertenece a la clase, podemos obtener la razón entre la probabilidades, para obtener a la vez un ranqueo de los elementos mas proclives a pertenecer a la clase, esto es:

$$\frac{P(C|X_i)}{P(\sim C|X_i)} \tag{3.10}$$

Basados en el teorema de Bayes visto anteriormente, podemos sustituir las probabilidades por la ecucación 3.8. con lo cual obtenemos lo siguiente:

$$\begin{aligned}
\frac{P(C|X_i)}{P(\sim C|X_i)} &= \frac{\frac{1}{P(X_i)} * P(C) * \prod_{i=1}^{n} P(X_i|C)}{\frac{1}{P(X_i)} * P(\sim C) * \prod_{i=1}^{n} P(X_i| \sim C)} \\
&= \frac{P(C) * \prod_{i=1}^{n} P(X_i|C)}{P(\sim C) * \prod_{i=1}^{n} P(X_i| \sim C)} \\
&= \frac{P(C)}{P(\sim C)} * \prod_{i=1}^{n} \frac{P(X_i|C)}{P(X_i| \sim C)}
\end{aligned} \tag{3.11}$$

aplicando logaritmos en ambos lados de la ecuación 3.11, para facilitar la aplicación y cálculos, aprovechando las propiedades de los mismos, obtenemos los siguiente:

$$\begin{aligned}
\ln\left(\frac{P(C|X_i)}{P(\sim C|X_i)}\right) &= \ln\left(\frac{P(C)}{P(\sim C)} * \prod_{i=1}^{n} \frac{P(X_i|C)}{P(X_i| \sim C)}\right) \\
&= \ln\left(\frac{P(C)}{P(\sim C)}\right) + \sum_{i=1}^{n} \ln\left(\frac{P(X_i|C)}{P(X_i| \sim C)}\right)
\end{aligned} \tag{3.12}$$

la expresión $\ln(\frac{P(C)}{P(\sim C}$), es una constante que para propósitos del ranqueo no es importante, debido a que su inclusión, solo desplaza en un valor constante a todos los elementos ranqueados (pero no impacta en la posición de las instancias ranqueadas), con esto obtenemos la métrica que llamaremos "Score", la cual será una medida de la propensidad de la instancia a pertenecer a la clase. Ahora podemos definir el cálculo de score como sigue:

$$Score = \sum_{i=1}^{n} \ln\left(\frac{P(X_i|C)}{P(X_i| \sim C)}\right) \tag{3.13}$$

donde $P(X_i|C) = \frac{N_{XC}}{N_C}$, $P(X_i| \sim C) = \frac{N_{X\sim C}}{N_{\sim C}}$, N_{XC} es el número de instancias con el atributo x las cuales pertenecen a la clase, N_C número total de instancias las cuales pertenecen a la clase, $N_{X\sim C}$ es el número de instancias con el atributo x las cuales no pertenecen a la clase y $N_{\sim C}$ número total de instancias las cuales no pertenecen a la clase.

Como podemos ver de la ecuación 3.13, el cálculo del score se vuelve bastante intuitivo y sencillo, teniendo como elementos los aportes parciales de cada uno de los atributos que participan en el modelo, para a el final contribuir todos a un score total, el cual da a la instancia un ranqueo, y con esto, un nivel de propensidad de la misma, con respecto a su pertenencia en la clase. El ranqueo conseguido sirve, tanto para clasificar, como para crear perfiles de riesgo utilizando las instancias en la parte más alta del ranqueo, estas son las de mayor riesgo o prepensidad de pertenencia a la clase, en cuestión de claificación, es suficiente con definir un umbral de score, a partir del cual se decide cualquier elemento con score mayor a ese umbral en la clase y los de menor score al umbral estarán en la no clase.

Los intentos de mejora de este algoritmo, son los denominados Naive Bayes Generalizado o Aumentado (NBG), en los cuales la oportunidad de mejora esta basada en la "debilidad" del NBS, que es la consideración de la independencia entre todas las variables entre si, los algoritmos aumentados buscan encontrar las correlaciones más importantes entre todas las variables, para ser tomadas dentro del algoritmo y así obtener mejor desempeño, por esto los algoritmos están basados en tener la mejor técnica para encontrar las correlaciones más importantes, explicamos a mayor detalle este caso en el capítulo 5.

3.4. Métricas de Desempeño

Existen muchas técnicas de la minería de datos para medir el desempeño, la más utilizada cuando se desea comparar diferentes tipos de clasificadores es la llamada curva ROC, debido a que es una técnica general y aplicable a cualquier algoritmo de clasificación, las otras dos que mostramos en este trabajo, son menos generalizables y dirigidas al tema de ranqueo, que como ya explicamos anteriormente, es lo de mayor interés para nuestro objetivo de estudio, las que utilizamos son: "Score Decile" y "TOP Porcentaje", ambas son técnicas que aprovechan el ranqueo creado por el score, para darnos una métrica de desempeño para cada nivel de partición y avance a través del ranqueo resultante, con esto podemos obtener más información de propensidad de pertenencia a la clase de grupos de instancias a diferentes niveles de score.

3.4.1. Área Bajo la Curva ROC

Este es un método ideal para comparar diferentes tipos de algoritmos clasificadores, ya que sus métricas están basadas solo en el acierto y error, hecho por el clasificador al pronosticar, sin importarle como llego a ese resultado.

El área bajo la curva ROC, es una representación gráfica de la sensibilidad frente a (1-especifidad). Proporciona herramientas para seleccionar los modelos posiblemente óptimos, y descartar modelos subóptimos, independientemente del coste de la distribución de las dos clases sobre las que se decide. Lo que esta técnica ofrece en resumen, es una forma de medir el desempeño de un clasificador en todas las direcciones posibles, tanto en pertenencia, como no, a la clase u objetivo. [3]

La siguiente figura 3.1 muestra la llamada matriz de confusión, la cual consiste en el conteo de estos aciertos y errores del clasificador, para calcular el resto de las métricas, estos valores son calculados fijando un umbral en el resultado del pronostico y simplemente contando las coincidencias con la clase a partir de ese umbral.

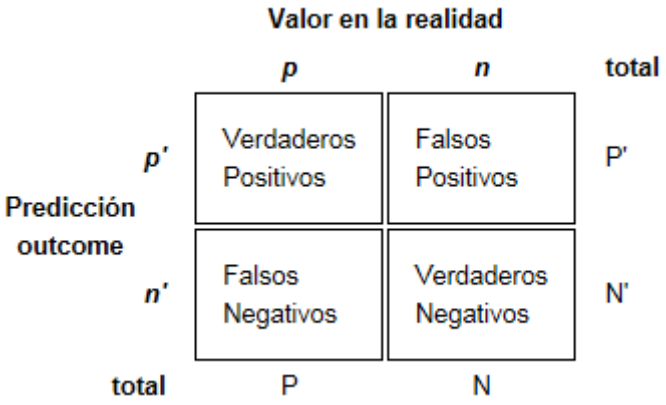


Figura 3.1: Concepto comparativo curva ROC [3]

Dónde:

Sensibilidad = VP/ P = VP /(VP + FN)
Especificidad = VN/N = VN/(FP + VN)
Precisión = (VP + VN) / (VP + FP + VN + FN)
Error = 1 - Precisión

y

VP = Verdaderos Positivo, VN = Verdaderos Negativos, FP = Falsos Positivos FN = Falsos Negativos

La curva ROC mide la robustez del modelo predictivo a diferentes umbrales de especificidad, como se muestra en la siguiente figura 3.2, esta es calculada utilizando la misma técnica anterior, solo que en este caso para obtener distintos puntos de sensibilidad y especificidad, se varia el umbral hacia arriba y abajo, se pueden poner tantos puntos como se desee, tomando en consideración que entre mas puntos se dibujen, mejor es la aproximación del área bajo la curva aunque también se incrementa el tiempo de cálculo de la misma.

3.4.2. Desempeño por Deciles de Score

El desempeño "Score Decile", mide el porcentaje de aciertos que se tienen, en 10 intervalos creados, por dividir el ranqueo generado por el modelo en 10 partes iguales (deciles), donde los deciles 10 y 9 son para el 20 % de los ranqueos ms altos (10 % para cada decile) y el 1 y 2 son para el 20 % de los raqueos más bajos respectivamente, lo cual significa que entre mayor es el número de aciertos en los deciles altos, mejor es el modelo al predecir y/o clasificar a los elementos de mayor probabilidad a pertenecer a la clase. Lo cual como hemos explicado permite generar perfiles de riesgo, ordenando las instancias por el score más alto, con esta métrica podemos decir que tan probable es que dichas instancias pertenezcan a la clase, obteniendo no solo una clasificación, si no también un perfil de riesgo. Como se muestra en la siguiente figura 3.3, cuando tenemos un buen modelo, esperamos que el porcentaje de aciertos vaya disminuyendo conforme descendemos en los deciles, con lo cual tendríamos la mayoría de los aciertos en los deciles más altos de score, esta métrica es aplicada en el archivo de prueba, al igual que todas las formas de medir desempeño, para dar validez a su resultado.

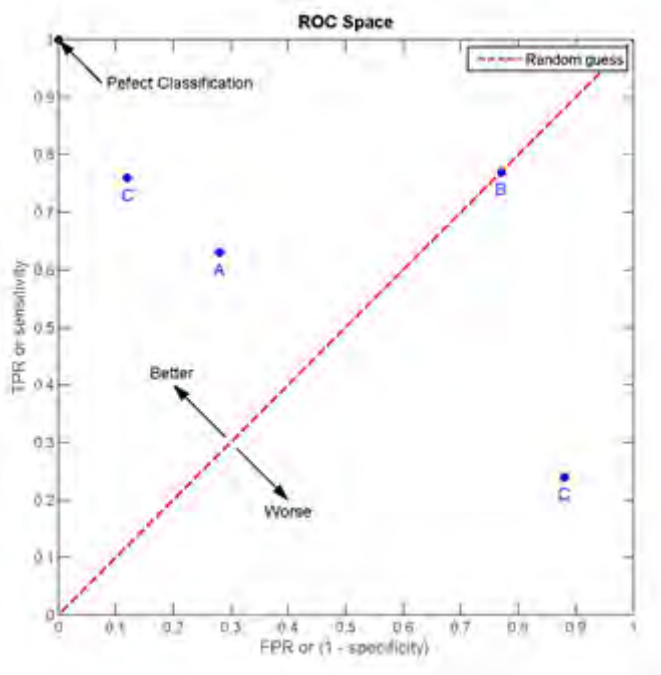


Figura 3.2: Espacio de Curva ROC [3]

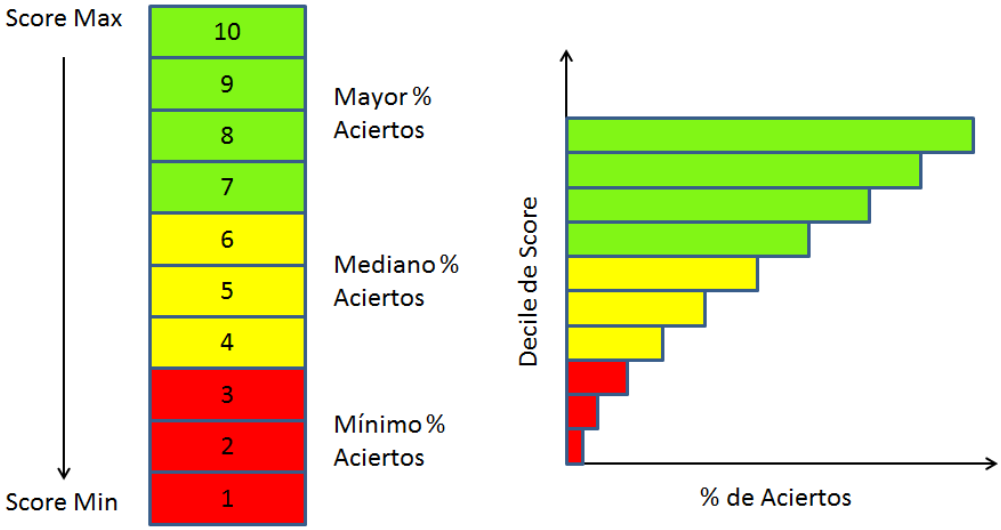


Figura 3.3: Desempeño por Deciles de Score

Entre mayor sea el porcentaje de aciertos en los deciles en verde y mucho menor los sea en deciles en rojo, el modelo sera mejor en su capacidad de ranqueo y por ende en su capacidad de crear perfiles de riesgo.

3.4.3. Desempeño por TOP Probabilidad

El desempeño "TOP Probabilidad", obtiene la probabilidad de predecir de forma adecuada la clase, a distintos niveles de porcentaje del ranqueo hecho por el modelo, partiendo desde el top, así que el 1 % es el 1 por ciento más alto del ranqueo en score y así sucesiva-

mente, esta probabilidad puede ser comparada con la probabilidad de elegir un elemento de la clase al azar (probabilidad de la clase), con lo cual se puede cuantificar a distintos niveles de score, que tanto es mejor utilizar el modelo para predecir o clasificar, comparado con seleccionar al azar. Figura 3.4.

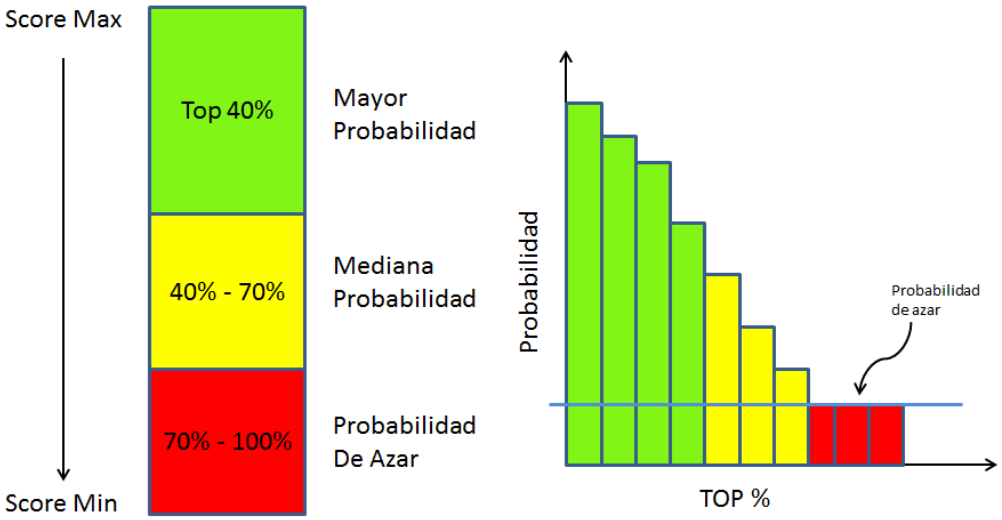


Figura 3.4: Desempeño por Top Probabilidad

Al igual que la métrica anterior, entre mayor sea la probabilidad en los top porcentajes en verde y mucho menor e igual a la probabilidad de azar, lo sea en porcentajes en rojo, el modelo sera mejor en su capacidad de ranqueo y por ende en su capacidad de asignar una probabilidad a una instancia en su perfil de riesgo.

3.5. Probabilidad Cero y Corrección de Laplace

Uno de los problemas a los que se enfrentan los clasificadores y en general los problemas de modelación es a la falta de muestra suficiente, cuando esto sucede es común enfrentarse a probabilidades cero o indeterminadas, cuando no se cuenta con evidencia de que un valor para cierta variable ha ocurrido o se ha presentado ya sea para la clase o la no clase, lo cual nos deja con el dilema para decidir si debemos considerar a este valor como cero, lo cual desde el punto de vista probabilistico no es del todo correcto, el hecho de que en la muestra de datos con la cual se cuenta no exista un valor para la clase o su completo, no significa que si tuviéramos una muestra mayor este caso estaría presente.

Existe una solución diseñada para cuando esto ocurre, llamada corrección de Laplace [3] (Laplace smooothing o suavizado de Laplace por su nombre en ingles), el cual precisamente lidia con este problema suavizando el cálculo, al añadir una compensación al calculo, es este caso al de probabilidades, evitamos que existan probabilidades cero o indefinidas cuando su presencia no es encontrada en la muestra. La correción de Laplace es definida como se muestra en la ecucación 3.14:

$$\theta = \frac{X_i + \alpha}{N + \alpha * d} \tag{3.14}$$

donde $\theta = (\theta_1, ..., \theta_d)$, $X = (X_1, ..., X_d)$ y $i(1, 2....d)$.

Si aplicamos este concepto de suavizado a la ecuación 3.13, para cada una de las probabilidades respectivamente, podríamos modificar el calculo de probabilidad actual por 3.15 y 3.16:

$$P(X_i|C) = \frac{N_{xc} + \alpha}{N_c + \beta * d} \tag{3.15}$$

y

$$P(X_i| \sim C) = \frac{N_{x\sim c} + \alpha}{N_{\sim c} + \beta * d} \tag{3.16}$$

Donde α y β son valores por definir y d es el número de clases que hay en el problema a clasificar, en nuestro caso, usualmente este valor es 2, por consideran la clase y su complemento (no clase) solamente. Muchas veces en artículos y trabajos relacionados con este tipo de problemáticas, se utiliza sumar 1 en el denominador, a esto se le conoce por su nombre en ingles "ADD-ONE Smoothing", donde se considera a $\alpha = \beta = 1$, lo cual es una posible solución al problema, pero como veremos en los resultados de los análisis, esto no siempre funciona de manera adecuada. Por lo que trataremos de encontrar la correcta relación entre α y β para mantener el correcto balance de las probabilidades, para esto consideraremos el caso de una sola variable, tomando del vector X_i alguno de sus elementos X_1, con lo cual obtenemos la ecuación 3.17:

$$P(X_1|C) = \frac{N_{X_1C} + \alpha}{N_C + \beta * d} \tag{3.17}$$

y tomando en cuenta que la sumatoria sobre todos los valores de X1 es igual a 1 3.18 :

$$\sum_{X_1} P(X_1|C) = 1 \tag{3.18}$$

aplicando la sumatoria como en 3.18 a 3.17 tenemos los siguiente:

$$\sum_{X_1}^{Z} P(X_1|C) = \frac{N_C + \alpha * Z}{N_C + \beta * d} = 1 \tag{3.19}$$

y despejando α de la ecuación 3.19 :

$$\begin{aligned} N_C + \alpha * Z &= N_C + \beta * d \\ \alpha &= \frac{\beta * d}{Z} \end{aligned} \tag{3.20}$$

donde Z es la cardinalidad de X_1, esto es el número de valores distintos posibles en este atributo, d es el número de clases y β sera un valor fijo definido por el usuario (lo mas comúnmente utilizado es 1).

3.6. Resultados Clasificador Naive Bayes Estándar

En este apartado presentamos los resultados obtenidos al aplicar el clasificador NBS sobre datos relacionados con enfermedades emergentes, como es la diabetes y los costos de la misma, para lo cual se obtuvo información de fuentes como ENCOPREVENIMSS 2006, DxCG compañia dedicada al análisis de riesgo utilizando modelos predictivos [26] y datos obtenidos del repositorio de la universidad de Irvine California [27] comúnmente utilizados para la prueba y comparación de distintos clasificadores y algoritmos, por su disponibilidad y utilización en muchos de los artículos académicos relacionados con esta área.

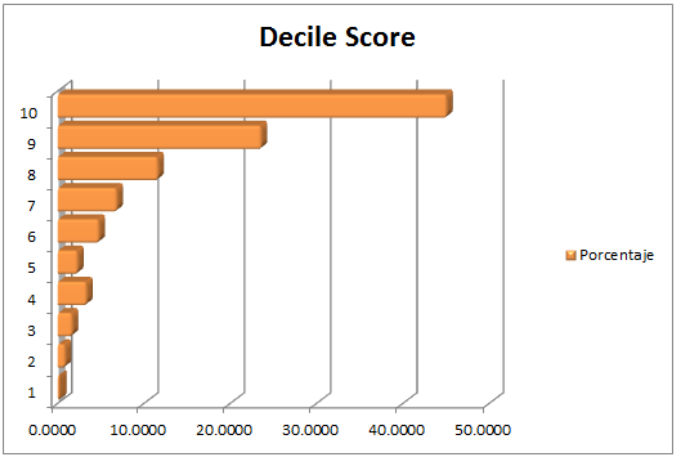
3.6.1. ENCOPREVENIMSS 2006

Las Encuestas Nacionales de Coberturas (ENCOPREVENIMSS), son encuestas aplicadas a una muestra de la población de los derecho-habientes del Instituto Mexicano del Seguro Social (IMSS), para este trabajo se pudo contar con acceso a los datos correspondientes a la encuesta realizada en el año de 2006, gracias a la cooperación con personal del IMSS, de lo cuales se pudieron crear 2 modelos, uno para hombres y otro para mujeres, utilizando el clasificador NBS tomando como objetivo (clase) a aquellas personas que padecían diabetes, lo resultados obtenidos fueron muy interesantes, así como los modelos resultantes y su desempeño, pero sobre todo por el lado de las variables mas predictivas (factores de riesgo), las cuales no contraponen a las ya conocidas, si no que, además se descubrieron nuevos factores de riesgo, los cuales son altamente probables de conducir a padecer diabetes.

Los resultados en desempeño de los modelos respectivos se muestran a continuación en la figura 3.5, tanto para el modelo de hombre (a), como para el de mujer (b), como se puede observar, en ambos modelos tenemos un destacado desempeño en los resultados de los raqueos creados por la aplicación del clasificador NBS, para el modelo de hombre, la figura 3.5a muestra que en el decile mas alto de score (decile 10), se tiene el 45.04 % de los aciertos totales (es decir elementos de la clase, en este caso, personas con diabetes), a diferencia del decile de menor score (decile 1), donde solo tenemos el 0.21 % de los aciertos, esto es una diferencia enorme, con la cual podemos asegurar que al aplicar el modelo a una población dada, de la cual no sabemos quien padece o no diabetes, el modelo ranqueara por score y en el decile 10 encontrara al 45 % de personas con diabetes, o altas probabilidades de padecerla, lo cual es interesante ya que se pueden crear intervenciones para estos grupos de alto riesgo, para prevenir, que al final, terminen padeciendo la enfermedad, al igual si tomamos los primeros 3 deciles, es decir 30 %, los aciertos crecen al 80.1 % figura 3.6a muestra que en el top 10 % tenemos una probabilidad de 29.76 % de encontrar a personas con o próximas a padecer diabetes, lo cual podría parecer poca efectividad del modelo, pero si lo comparamos con la probabilidad de azar (probabilidad general de la población a padecer la enfermedad o porcentaje de la población de la muestra de datos con diabetes), la cual es el del 6.58 %, entonces el modelo es 4.5 veces mejor que hacerlo al azar, es decir hay 450 % mas probabilidades de escoger correctamente a una persona que padecerá la diabetes o ya la tiene, y si tomamos el top 1 %, entonces las oportunidades aumentan hasta 6.7 veces mejor que al azar o 670 % mejor probabilidad.

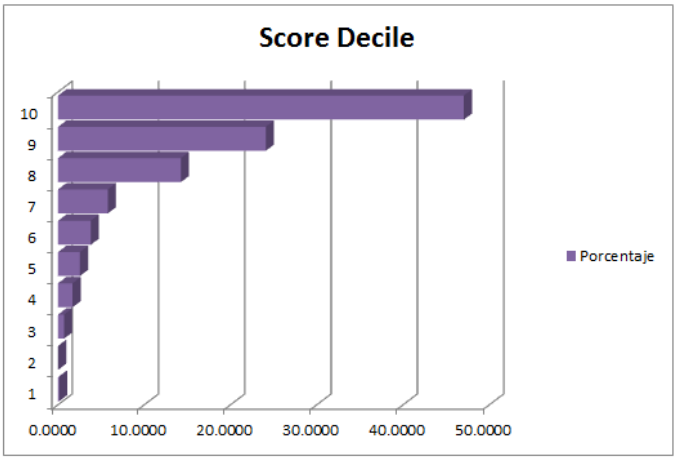
Lo anterior importa mucho a la hora de diseñar intervenciones, por que en enfermedades y sobre todo en diabetes, los recursos en infraestructura, atención y económicos son limitados, por lo que invertir bien ese dinero y recursos es crucial, y la forma de hacerlo correctamente es intervenir en las personas que realmente tienen riesgo de padecer diabetes, usando este ejemplo, si alguna institución publica o privada de salud decidiera invertir cierta cantidad de dinero en prevenir que un grupo de personas de determinada población padeciera diabetes, y eligiera a un grupo de 20 personas al azar, en realidad estaría ayudando solo a 1 persona, la cual esta en verdadero riesgo (20 * 0.0658 que es la probabilidad al azar de padecer diabetes es 1.2), entonces la inversión de recursos realmente serviría solo para ayudar a una sola persona, en cambio si aplicara el modelo y escogiera a las 20 personas basada en el ranqueo del top 1 %, estaría ayudando a 9 personas que realmente lo necesitan (20 * 0.4459 probabilidad de elegir un diabético en el top 1 % de score, es igual a 8.9) entonces la inversión, en este caso, de los recursos, ayudaría a prevenir 9 personas con diabetes, esto comparado con el resultado de elegir al azar es 9 veces mejor en costo/beneficio. De ahí la importancia de la creación y utilización de este tipo modelos, para ayudar en la lucha contra las enfermedades emergentes y la mejor inversión de los

Decile	Aciertos	Porcentaje
10	218	45.0400
9	114	23.5500
8	56	11.5700
7	33	6.8200
6	23	4.7500
5	11	2.2700
4	16	3.3100
3	8	1.6500
2	4	0.8300
1	1	0.2100



(a) Deciles de Score Modelo Hombre

Decile	Aciertos	Porcentaje
10	334	47.0400
9	171	24.0800
8	101	14.2300
7	41	5.7700
6	27	3.8000
5	18	2.5400
4	12	1.6900
3	5	0.7000
2	0	0.0000
1	1	0.1400



(b) Deciles de Score Modelo Mujer

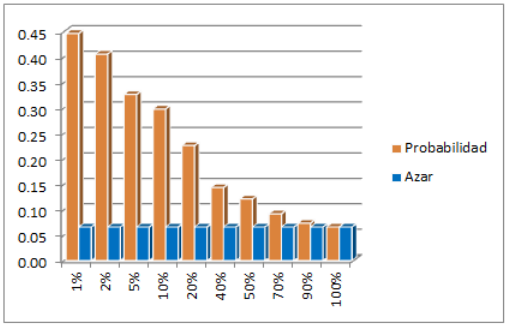
Figura 3.5: Desempeño por Deciles de Score Modelos Diabetes ENCOPREVENIMSS

recursos destinados para esta causa.

En el caso del modelo de mujer, los resultados son similares, figura 3.5b 47.04 % de los aciertos en el decile 10 y 0.14 % en decile 1, 85.35 % de los aciertos en los primeros 3 deciles y la figura 3.6b muestra en el top 10 % una probabilidad 37.15 % comparado con la de azar de 7.91 %, lo cual es 4.7 veces mejor usar el modelo y en el top 1 % es 46.67 % probabilidad, lo cual es 5.9 veces mejor, los resultados son similares al modelo de los hombres, salvo que vemos un ligero aumento en la incidencia de diabetes para mujeres, lo cual es un hallazgo interesante, de acuerdo a las estadísticas en esta encuesta y al modelo, las mujeres son mas propensas a padecer diabetes que los hombres.

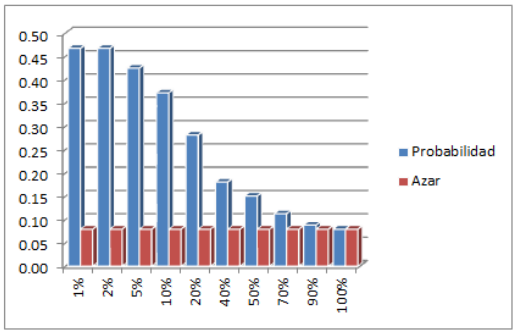
La otra parte de los resultados que interesan, son la creación de perfiles, lo cual se logra con la identificación de los atributos característicos de las personas pertenecientes a un mismo grupo, la clase (en este caso de estudio son los diabeticos), para identificar estos perfiles utilizamos la métrica llamada “épsilon” y las probabilidades correspondientes de cada par variable/valor del cual se tienen datos. Vamos a presentar aquellos cuyos resultados son menos intuitivos, como el mostrado a continuación, a la pregunta expresa de la encuesta ¿Sabe leer o escribir un recado? la persona tenia 4 opciones de respuesta valida: “Si”, “No”, “No sabe” y “No responde”, al hacer el análisis de epsilon y probabilidades

Decile	Aciertos	Probabilidad	Probabilidad Azar
1%	33	0.4459	0.0658
2%	60	0.4054	0.0658
5%	120	0.3261	0.0658
10%	219	0.2976	0.0658
20%	332	0.2257	0.0658
40%	421	0.1431	0.0658
50%	444	0.1208	0.0658
70%	471	0.0915	0.0658
90%	483	0.0730	0.0658
100%	484	0.0658	0.0658



(a) Top Porciento Modelo Hombre

Decile	Aciertos	Probabilidad	Probabilidad Azar
1%	42	0.4667	0.0791
2%	84	0.4667	0.0791
5%	191	0.4244	0.0791
10%	334	0.3715	0.0791
20%	505	0.2810	0.0791
40%	647	0.1801	0.0791
50%	674	0.1501	0.0791
70%	704	0.1120	0.0791
90%	710	0.0878	0.0791
100%	710	0.0791	0.0791



(b) Top Porciento Modelo Mujer

Figura 3.6: Desempeño por "Top Porciento"Modelos Diabetes ENCOPREVENIMSS

para esta pregunta y sus posibles respuestas, se obtuvo lo mostrado en la figura 3.7, donde podemos ver que la incidencia de diabetes en las personas analfabetas (10.24 %), es aproximadamente 60 % mas en comparación con las alfabetizadas, este atributo podría ser considerado como un proxy para nivel socio-económico, ademas podemos ver que su valor de épsilon en ambos casos es mayor a 2 o -2, lo cual nos permite validar el significado estadístico de este hallazgo y su correlación con la clase, para el caso de "Si" es de -3.6268, lo cual indica una anticorrelación con la clase (diabetes), y por lo tanto, lo podemos considerar como un factor protector, para el caso de "No", el valor de épsilon es 10.3469 esto es altamente correlacionado a la clase, es decir, es un factor de riesgo. El resto de los posibles valores de respuesta aun cuando presentan altos valores de probabilidad, su valor de épsilon nos indica que no tenemos muestra suficiente para considerarlos con un significado estadístico relevante, por eso podemos descartarlos como proxies (factores protectores o de riesgo).

Variable	Valor	Epsilon	P(C/X)
¿Sabe leer y escribir un recado?	Sí	-3.6268	0.0611
¿Sabe leer y escribir un recado?	No	10.3469	0.1024
¿Sabe leer y escribir un recado?	No sabe	0.8896	0.1071
¿Sabe leer y escribir un recado?	No responde	1.7182	0.2000

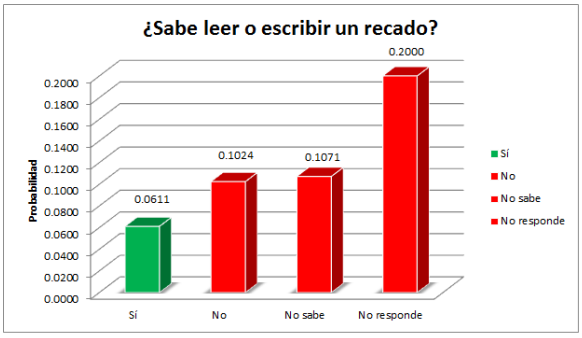


Figura 3.7: Atributo (pregunta) de encuesta ¿Sabe leer o escribir un recado?

Otro resultado interesante se obtuvo al analizar la pregunta ¿Sabe que es el sexo protegido?, cuyas posibles respuestas contemplaban las siguientes opciones: “Otro”, “No sabe/responde”, “Proteger los derechos sexuales de la gente”, “No tener relaciones sexuales”, “Evitar el embarazo”, “Tener relaciones sexuales únicamente con pareja”, “Protegerte de la enfermedades sexuales” y “Utilizar condón”, a pesar de las muchas opciones podemos ver que hay solo una respuesta correcta, por lo cual, agregamos una columna para medir el grado de ignorancia sobre el tema acorde a la respuesta elegida, como podemos apreciar en la figura 3.8, las 2 repuestas de mayor grado de ignorancia tienen el doble de riego que la respuesta correcta, ademas de esta ultima ser un factor protector, de ambas preguntas podemos concluir que la ignorancia es un factor de riesgo a la hora de padecer diabetes y sobre todo la ignorancia en temas de salud, lo cual suena lógico visto así, ya que alguien no interesado o documentado sobre temas generales de salud, es alguien también, muy propenso a padecer enfermedades por esa misma falta de conocimiento, incluida la diabetes.

Respuesta	Grado de ignorancia	Epsilon	P(C/X)
Otro	8	2.4499	0.1327
No sabe/responde	7	4.8657	0.1272
Proteger los derechos sexuales de la gente	6	-0.3692	0.0667
No tener relaciones sexuales	5	1.2927	0.0774
Evitar embarazo	4	0.0537	0.0699
Tener relaciones sexuales unicamente con pareja	3	0.5774	0.0711
Protegerte de las enfermedades sexuales	2	-1.0710	0.0665
Utilizar condon	1	-3.1397	0.0629

Figura 3.8: Atributo (pregunta) de encuesta ¿Sabe que es sexo protegido?

En relación con la actividad física y el ejercicio, se obtuvieron también resultados interesantes a través de la preguntas ¿Practica algún deporte? figura 3.9 y ¿Cuantos días a la semana hace ejercicio? figura 3.10, de estas preguntas podemos observar, que la incidencia de diabetes es 30 % mayor para aquellas personas que no hacen ejercicio vs aquellos que si lo hacen, y además para quienes lo hacen 5 o mas días a la semana, la incidencia de diabetes es menor a la de la población general, esto puede ser considerado como un factor protector.

Respuesta	Epsilon	P(C/X)
Si	-4.4084	0.0586
No	3.7665	0.0777

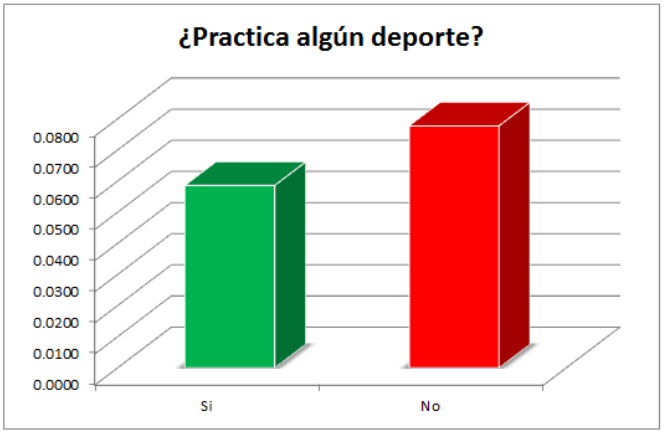


Figura 3.9: Atributo (pregunta) de encuesta ¿Practica algn deporte?

Días	Epsilon	P(C/X)
7	-6.4664	0.0342
6	-5.1536	0.0422
5	-2.3031	0.0570
4	0.2279	0.0718
3	0.7318	0.0753
2	0.1661	0.0718
1	5.3413	0.1110

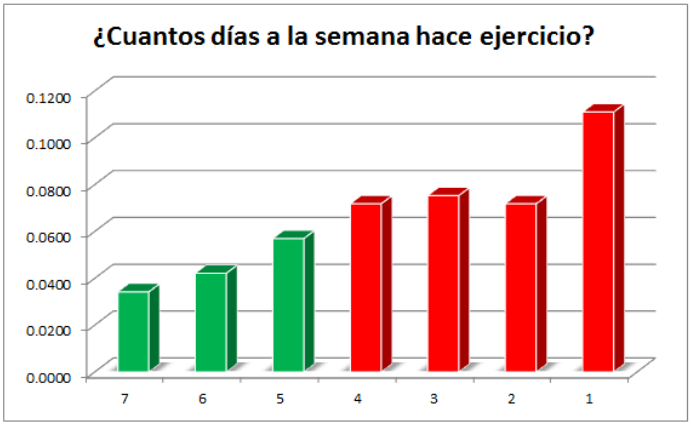


Figura 3.10: Atributo (pregunta) de encuesta ¿Cuantos días a la semana hace ejercicio?

Sumado a lo anterior, tenemos dos preguntas también relacionadas con la actividad física y el ejercicio, las preguntas son ¿Cuantos minutos hace de ejercicio? y ¿Que tipo de deporte practica?, en la figura 3.11 podemos observar, que el ejercicio solo se convierte en un factor protector, si el ejercicio se realiza durante 50 minutos o mas y en la figura 3.12 podemos observar que en especial el tipo de deporte, también es un factor protector pero puede haber muchas posibles causas detrás de los resultados, por ejemplo, en esta misma figura 3.12 se puede ver que las personas con actividad física "Caminar", tienen una incidencia alta de diabetes 14 %, esto no significa que caminar sea malo (no sano) o sea un factor de riesgo para padecer diabetes, la interpretación adecuada tal vez sea: cuando las personas son diagnosticadas con diabetes por indicaciones medicas, se les pide hacer ejercicio, pero es posible que sus posibilidades de ejercitarse, dado su deterioro físico, solo les permite caminar y solo por algunos minutos.

Minutos	Epsilon	P(C/X)
5	0.7608	0.1429
10	3.6481	0.1728
20	2.6581	0.0989
30	5.2324	0.1005
40	0.5835	0.0746
50	-1.9752	0.0469
60	-4.0293	0.0489
70	-8.4363	0.0356

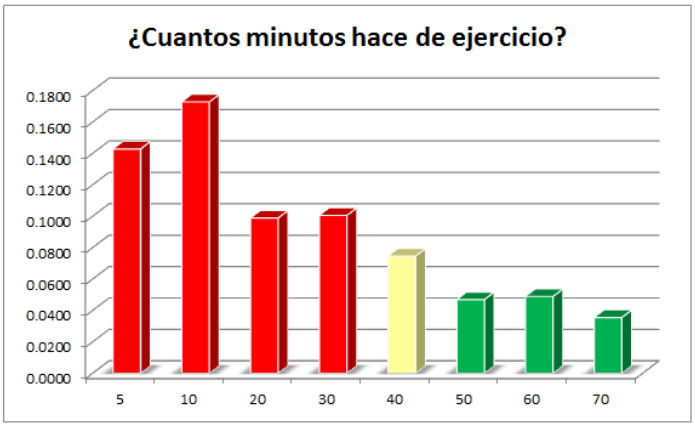


Figura 3.11: Atributo (pregunta) de encuesta ¿Practica algn deporte?

De estas variables sobre actividad física, podríamos concluir que la hipótesis sobre el beneficio de hacer ejercicio, como factor preventivo de la diabetes, es correcta pero únicamente si la frecuencia (número de días) y duración (minutos por día), son adecuados. Desafortunadamente, para confirmar varias de las hipótesis arriba mencionadas, se requieren datos longitudinales y detallados, esencialmente las historias clínicas de las personas. En la ausencia de datos de esa naturaleza, las hipótesis se quedan como hipótesis. Las cuales tendrían que comprobarse en algún tipo de experimento o con datos, los cuales

Tipo Ejercicio	Epsilon	P(C/X)
baloncesto	-3.8513	0.0259
bicy fija	1.7133	0.1064
artes marciales	0.0339	0.0706
beisbol	-2.3892	0.0407
caminar	14.4840	0.1470
correr	-3.4413	0.0431
natacion	-2.2335	0.0248
bicycleta	-1.0813	0.0588
otro	-0.5993	0.0614
baile	-1.4040	0.0196
pesas	-3.9344	0.0254
futbol	-11.3227	0.0284

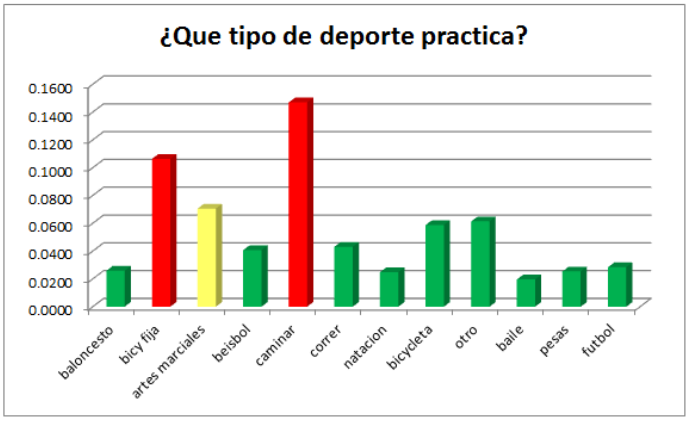


Figura 3.12: Atributo (pregunta) de encuesta ¿Cuantos días a la semana hace ejercicio?

contengan toda la información necesaria. Esto es una gran desventaja, debido a que seguramente los factores de riesgo de estilo de vida dependen del tiempo. Aun así como un primer acercamiento a los perfiles y factores de riesgo de los diabéticos, y la identificación y de personas propensas a padecer la enfermedad, el modelo NBS hace una gran trabajo como lo mostramos, en este capítulo.

3.6.2. DxCG 97-99

Estos datos fueron obtenidos de una compañía dedicada al análisis de riesgo y modelos predictivos llamada Verisk Helath [26] y en especifico de datos utilizados en lo que ellos llaman DCG (Diagnostic Cost Group), estos datos contienen la información de los gastos generados por personas diabeticas, las cuales tienen un seguro de gastos médicos mayores, los gastos son de 3 tipos: de hospitalización, consultas medicas y exámenes de laboratorio/medicina, además de los diagnósticos de las enfermedades por las que recibieron atención y algunos otros atributos correspondientes a los pacientes. Se cuentan con 2 periodos de información 97-98 y 98-99, por lo cual, para los modelos decidimos utilizar el periodo 97-98 y 98-99 para predecir los costos en 99, el modelo esta enfocado en encontrar a las personas de mayor costo y sus factores de riesgo, para esto se definió la clase como aquellas personas en el top 5 % de costos totales (hospitalización + consultas + medicina).

El desempeño del modelo se muestra a continuación en la figura 3.13 y en la figura **??**, como se puede observar el desempeõ aciertos (personas en el top 5 % de gastos totales), en el decile más alto de score (decile 10) y al 76.57 %, en los primeros tres deciles, a diferencia del decile más bajo (decile 1), donde solo hay 1.8 % de los aciertos, además la probabilidad en el mismo top 10 % es de 26.15 %, lo cual es 6 veces mas que la probabilidad de azar (encontrar escogiendo al azar a alguien que estará en el top 5 % de costos), y si vamos al top 1 % únicamente, el modelo es 10 veces mejor, lo cual se convierte en 1000 % mejor que el azar. Con esto podemos ver la utilidad de la creación de modelos muy básicos, utilizando la técnica de NBS, es por esto que el proyecto presentado propone algunas mejoras y teorías que podrían ayudar a mejorar, aunque sea un poco, los desempeños logrados al utilizar este clasificador.

En cuanto a los atributos destacados, presentamos en la tabla 3.1, los top 10 más importantes identificados por la métrica épsilon, y al analizarlos uno por uno podemos entender que tienen mucha lógica, el primero de ellos es Clase_98, este atributo significa “La persona estuvo en el top 5 % de costos el año pasado”, de esto podemos concluir que es altamente probable que cuando una persona tuvo gastos médicos altos el año anterior los siga tenien-

Decile	Aciertos	Porcentaje
10	119	53.6000
9	33	14.8600
8	18	8.1100
7	16	7.2100
6	6	2.7000
5	11	4.9500
4	5	2.2500
3	4	1.8000
2	6	2.7000
1	4	1.8000

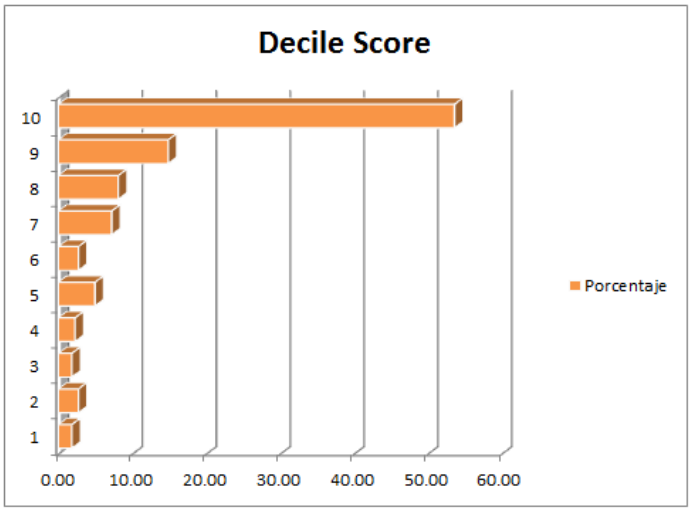


Figura 3.13: Desempeño por Deciles de Score Modelo Costos DxCG

Decile	Aciertos	Probabilidad	Probabilidad Azar
1%	24	0.5217	0.0488
2%	47	0.5165	0.0488
5%	86	0.3772	0.0488
10%	119	0.2615	0.0488
20%	152	0.1672	0.0488
40%	186	0.1023	0.0488
50%	192	0.0845	0.0488
70%	208	0.0654	0.0488
90%	218	0.0533	0.0488
100%	222	0.0488	0.0488

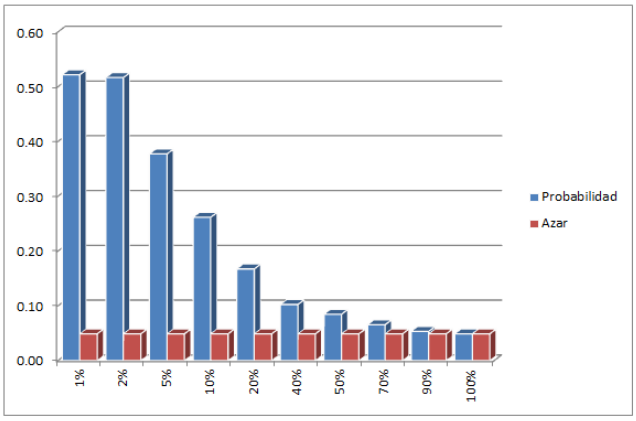


Figura 3.14: Desempeño por "Top Porciento"Modelo Costos DxCG

do en el siguiente año, lo cual es lógico, si entendemos que los altos costos van asociados con el deterioro de la salud de la persona y cuando son enfermedades como la diabetes, se espera en la mayoría de los casos, que este deterioro continue, por lo cual probablemente, cada año que pase, la persona incremente sus costos, el atributo HCC131_98 habla de un diagnostico medico para falla renal, esto nos indica que cuando una persona con diabetes presenta falla renal, el deterioro en su salud llego a un punto sin retorno, en el cual requerirá de mayor atención medica y hospitalaria, medicina y por ende sus costos médicos se irán elevando cada vez mas, los atributos CG_ACOV98 (gastos totales en 98), CG_COV98O (gastos en consultas medicas en 98), CG_ACOV97 (gastos totales en 97) y CG_COV98I (gastos en hospitalización en 98), estos 5 atributos confirman al primero, una vez que vemos un incremento en el gasto de los pacientes es alguna de las áreas o en todas, como por ejemplo CG_COV98O donde el paciente empieza a tener mas y mas consultas medicas, por evidentes molestias que le causa la enfermedad o CG_COV98I donde la persona requiere de hospitalizaciones medicas de urgencia muchas veces al año nos indica igualmente el deterioro progresivo de la enfermedad y el consiguiente aumento en costos; los atributos HCC130_98 (en diálisis), HCC131_97 (falla renal hace dos años) y HCC015_98 (diabetes con síntomas renales) nos dejan ver que la peor parte de la diabetes es llegar a la falla renal, además de que es la parte mas costosa de los tratamientos, y por último CG_TLOS_98 (días totales de hospitalización) con un valor medio, lo cual significa que no

son de los que mas tiempo pasaron hospitalizados, pero si tiempos cortos muchas veces, nos indica tal vez personas con urgencias medicas continuas, que requieren de algunos días para estabilizarlos y regresar a casa, pero que pueden tener 2 o mas eventos de este tipo al año, por lo cual sus costos médicos son altos. Como podemos apreciar, utilizando estas métricas y el clasificador NBS, se pueden obtener muchos resultados valiosos, pero sobre todo lógicos, a los cuales con un poco de estudio y pensamiento pueden ser entendidos y clasificados como factores de riesgo, ademas se puede observar que las técnicas de minería de datos, pueden ser una herramienta muy poderosa, que ayuden al estudio de las enfermedades y a la epidemiología.

Atributo	Valor	Epsilon	$P(C\|X)$	Descripción
TOP5_97	Si	32.25	0.3047	Estaba en top 5 % de costos en 97
HCC131_0	Si	34.71	0.4543	Falla renal 98
CG_ACOV98	intervalo mas alto	32.99	0.2183	Costos total en 98
CG_COV98O	intervalo mas alto	28.59	0.1975	Costos totales de consultas en 98
CG_ACOV97	intervalo mas alto	26.60	0.1956	Costos totales en 97
CG_COV98I	intervalo medio	25.87	0.2702	Costos totales de hospitalización 98
HCC130_0	Si	25.04	0.6800	En diálisis 98
HCC131_1	Si	24.91	0.3913	Falla renal 97
CG_TLOS_0	intervalo medio	21.08	0.4783	Total de días en hospitalización
HCC015_0	Si	21.04	0.2688	Diabetes con síntomas renales

Tabla 3.1: TOP 10 atributos en valor de épsilon.

3.6.3. Datos de Irvine Universidad de California

Estos datos se encuentran disponible en el sitio web de las universidad UCI (University of California Irvine), es un repositorio con muchos tipos de datos disponibles para propósitos de minería de datos, los datos son donaciones y por lo tanto están disponibles a cualquier persona que desee utilizarlos, en muchos de los artículos sobre minería de datos y aprendizaje de maquina son utilizados estos repositorios, por el hecho de que están disponible a cualquiera y por lo tanto se pueden utilizar para comparar algoritmos, también para hacer múltiples pruebas y ver la consistencia de los mismos, en este trabajo las utilizaremos para dos cosas, la primera en este apartado para probar la eficiencia y conveniencia de utilizar la corrección de la Laplace y los beneficios que aporta al NBS y mas adelante en cap. 5 para comparar los desempeños entre NBS y el NBG propuesto en el mismo capítulo.

Para probar la eficiencia y beneficios de utilizar la correción de Laplace, junto con algoritmos de minería de datos, en particular con NBS, se tomaron 10 conjuntos de datos del repositorio y se ejecuto 20 veces sobre cada conjunto el NBS, para obtener las mediciones de: “Sensibilidad”, “Especificidad”, “Error”, “Precisión” y “Área bajo la curva ROC”, con el fin de poder observar los beneficios en la clasificación, por usar o no usar la correción de Laplace, en la tabla 3.2, se muestran los promedios y desviación estándar de estas 20 corridas para cada conjunto de datos en tres tipos: “Sin Correción de Laplace”, “Con Correción de Laplce ADD-ONE” y “Con Correción de Laplace Generalizada”, la cual fue derivada en en la sección 3.2 en la ecuación 3.20, para todos los experimentos se fijo el valor de $\beta = 1$.

Conjuto de Datos	Correción de Laplace	Métrica	Sensitividad	Especificidad	Precisión	Error	AUC	% Disminución Error
Adultos 48,882 registros	No	Promedio	0.7938	0.8246	0.8172	0.1828	0.9003	X
		Desv. Est.	0.0061	0.0033	0.0031	0.0031	0.0028	X
	Si - ADD-ONE	Promedio	0.7958	0.8247	0.8178	0.1822	0.9007	0.30 %
		Desv. Est.	0.0051	0.0038	0.0034	0.0034	0.0026	X
	Si - Generalizado	Promedio	0.7966	0.8257	0.8187	0.1813	0.9011	0.81 %
		Desv. Est.	0.0053	0.0038	0.0034	0.0034	0.0023	X
Wine 178 registros	No	Promedio	1.0000	0.8200	0.8660	0.1340	0.9637	X
		Desv. Est.	0.0000	0.0494	0.0401	0.0401	0.0094	X
	Si - ADD-ONE	Promedio	1.0000	0.9415	0.9575	0.0425	0.9691	68.31 %
		Desv. Est.	0.0000	0.0317	0.0251	0.0251	0.0116	X
	Si - Generalizado	Promedio	1.0000	0.9895	0.9925	0.0075	0.9744	94.37 %
		Desv. Est.	0.0000	0.0132	0.0095	0.0095	0.0015	X
Segmentation 2,310 registros	No	Promedio	0.8335	0.8581	0.8545	0.1455	0.9227	X
		Desv. Est.	0.0468	0.0120	0.0120	0.0120	0.0157	X
	Si - ADD-ONE	Promedio	0.8364	0.8742	0.8689	0.1311	0.9440	9.87 %
		Desv. Est.	0.0247	0.0118	0.0098	0.0098	0.0071	X
	Si - Generalizado	Promedio	0.8202	0.8821	0.8737	0.1263	0.9483	13.19 %
		Desv. Est.	0.0313	0.0127	0.0119	0.0119	0.0066	X
Vehicle 946 registros	No	Promedio	0.7964	0.8494	0.8358	0.1642	0.9122	X
		Desv. Est.	0.0447	0.0209	0.0194	0.0194	0.0164	X
	Si - ADD-ONE	Promedio	0.8032	0.8907	0.8670	0.1330	0.9326	19.01 %
		Desv. Est.	0.0420	0.0198	0.0183	0.0183	0.0096	X
	Si - Generalizado	Promedio	0.8437	0.9201	0.9000	0.1000	0.9463	39.11 %
		Desv. Est.	0.0353	0.0185	0.0210	0.0210	0.0105	X
Pendigits 10,992 registros	No	Promedio	0.9011	0.9555	0.9499	0.0501	0.9742	X
		Desv. Est.	0.0141	0.0032	0.0033	0.0033	0.0041	X
	Si - ADD-ONE	Promedio	0.9013	0.9568	0.9511	0.0489	0.9766	2.48 %
		Desv. Est.	0.0110	0.0023	0.0027	0.0027	0.0037	X
	Si - Generalizado	Promedio	0.8937	0.9634	0.9562	0.0438	0.9784	12.59 %
		Desv. Est.	0.0132	0.0035	0.0029	0.0029	0.0031	X
Hypothyroid 3,163 registros	No	Promedio	0.9658	0.9511	0.9518	0.0482	0.9865	X
		Desv. Est.	0.0199	0.0074	0.0074	0.0074	0.0089	X
	Si - ADD-ONE	Promedio	0.9539	0.9680	0.9674	0.0326	0.9889	32.28 %
		Desv. Est.	0.0242	0.0039	0.0042	0.0042	0.0052	X
	Si - Generalizado	Promedio	0.9546	0.9714	0.9706	0.0294	0.9911	39.06 %
		Desv. Est.	0.0212	0.0053	0.0053	0.0053	0.0022	X
	No	Promedio	0.8579	0.8140	0.8383	0.1617	0.8932	X
		Desv. Est.	0.0403	0.0719	0.0362	0.0362	0.0254	X
		Promedio	0.8424	0.8702	0.8574	0.1426	0.9052	11.83 %

Continua en la siguiente página

Conjuto de Datos	Correción de Laplace	Métrica	Sensitividad	Especificidad	Precisión	Error	AUC	% Disminución Error
Statlog-Hearth 270 registros	Si - ADD-ONE	Desv. Est.	0.0488	0.0514	0.0317	0.0317	0.0229	X
	Si - Generalizado	Promedio	0.8479	0.8702	0.8611	0.1389	0.9012	14.12 %
		Desv. Est.	0.0465	0.0379	0.0297	0.0297	0.0233	X
Annealing 898 registros	No	Promedio	0.9624	0.7036	0.7320	0.2680	0.9522	X
		Desv. Est.	0.0262	0.0291	0.0275	0.0275	0.0156	X
	Si - ADD-ONE	Promedio	0.9775	0.9766	0.9766	0.0234	0.9937	91.26 %
		Desv. Est.	0.0281	0.0081	0.0075	0.0075	0.0016	X
	Si - Generalizado	Promedio	0.9836	0.9879	0.9874	0.0126	0.9946	95.28 %
		Desv. Est.	0.0170	0.0065	0.0058	0.0058	0.0009	X
Satellite 6,435 registros	No	Promedio	0.8706	0.8453	0.8477	0.1523	0.9270	X
		Desv. Est.	0.0247	0.0065	0.0069	0.0069	0.0084	X
	Si - ADD-ONE	Promedio	0.8732	0.8396	0.8429	0.1571	0.9282	-3.20 %
		Desv. Est.	0.0140	0.0073	0.0071	0.0071	0.0055	X
	Si - Generalizado	Promedio	0.8561	0.8479	0.8487	0.1513	0.9242	0.61 %
		Desv. Est.	0.0227	0.0064	0.0058	0.0058	0.0077	X
Ionosphere 351 registros	No	Promedio	0.9484	0.8081	0.8986	0.1014	0.9229	X
		Desv. Est.	0.0193	0.0563	0.0261	0.0261	0.0186	X
	Si - ADD-ONE	Promedio	0.9378	0.8454	0.9052	0.0948	0.9356	6.57 %
		Desv. Est.	0.0223	0.0530	0.0228	0.0228	0.0194	X
	Si - Generalizado	Promedio	0.9465	0.8424	0.9067	0.0933	0.9429	7.98 %
		Desv. Est.	0.0270	0.0544	0.0258	0.0258	0.0151	X

Tabla 3.2: Resultados sobre pruebas usando correción de Laplace

De los resultados obtenidos al aplicar estos experimentos, los cuales se pueden observar en la tabla 3.2, vemos que los conjuntos de datos con mayor número de registros de muestra, son los menos beneficiados (Adults, Pendigits y Satellite) y los conjuntos con menor número de registros de muestra son los mayormente beneficiados (Wine, Vehicle, Statlog-Hearth, Annealing y Ionosphere), el resto obtienen beneficios aun cuando su muestra es considerable, pero no seria necesariamente cierto decir que cuando una muestra tiene muy poca cantidad de registros, la correción de Laplace funcionara y el contrario cuando se tiene una muestra suficiente, basta observar el caso de “Ionosphere”, el cual comparado con “Segmentation”, “Vehicle”, “Hypothyroid” y “Annealing” obtienen una disminución de error menor, aun cuando estas ultimas cuentan con mayor número de registros en su muestra, haciendo una análisis mas profundo en el contenido de esas muestras, pudimos observar que si bien el tamaño de la muestra es un factor determinante para el éxito o fracaso de la corrección de Laplace, esto no es debido a la cantidad de registros, si no mas bien a la cantidad de información, con esto nos referimos a que el contenido de la muestra sea representativo, es decir, que contenga la mayor cantidad de casos en cuanto a combinaciones de atributo/valor, sin importar si es solo un ejemplar, es en estos caso cuando corrección de Laplace funciona mejor, cuando la muestra no es lo suficientemente representativa de todos los casos posibles que se pueden presentar en el fenómeno de estudio, lo cual es lógico, si lo vemos de esta forma, ya que la corrección de Laplace se encarga de llenar estos huecos en la muestra de información, esto es por lo cual se ve el

beneficio dramático de un conjunto a otro poco beneficio, cuando la muestra es suficiente y heterogénea (Adults) versus un gran beneficio cuando es lo contrario (Wine y Annealing); en conclusión podemos decir que siempre es mejor utilizar la corrección de Laplace, dado que no afecta el desempeño y a lo mas puede ser que el beneficio sea muy poco, pero nunca sera negativo.

Capítulo 4

Selección de Características

Como se comento en el capítulo 2, la herramienta de selección de características es muy útil para elegir aquellos atributos útiles al modelo, con esto nos referimos a aquellos que aportan descriminación al mismo, contribuyen a su desempeño y no son redundantes con otros atributos contenidos en el mismo conjunto de datos, además para que sea una selección valida y justa, en el campo de la minería, esta debe ser de forma automática y sin intervención humana de ningún tipo.

En este trabajo presentamos nuestra propia versión del algoritmo de selección de características el cual fue diseñado para elegir solo aquellas atributos útiles para un clasificador a partir de un conjunto grande de variables candidatas de forma automática, sin perder predictibilidad, ni desempeño del mismo, esta optimizado para trabajar con el clasificador Naive Bayes Estandar y su optimización se basa en las métricas utilizadas por dicho clasificador.

4.1. Metodología

La idea es optimizar el número de atributos a utilizar por parte del modelo al mínimo, con lo cual se puede eliminar la redundancia de los atributos que estén correlacionados o aquellos cuya aportación es nula o solo ruido, sin afectar el desempeño del modelo, pero dado que esta selección debe hacerse de tal forma que ningún criterio teórico o de conocimiento previo humano sesgue esta selección, entonces esto debe ser hecho de forma automática por el mismo algoritmo, una forma sería buscar una a una y en todas las combinaciones posibles de atributo/valor aquellas cuya similitud o correlación sea muy alta (y para esto se podría utilizar cualquier indicador estadístico de correlación), para lograr esto se requeririá de utilizar un método exhaustivo, el cual crecería en complejidad conforme el fenómeno a analizar contara con más y más atributos candidatos, además faltaría por eliminar aquellos atributos que aunque no son correlacionados con ninguno otro, tampoco aportan gran descriminación o conocimiento para el modelo.

Existe una mejor alternativa al método exhaustivo, el cual es probar de forma automatizada con diferentes combinaciones de atributos y utilizando el mínimo posible de estas, sin perder o sacrificar el desempeño del modelo, se escogen las combinaciones de atributos con mayor éxito en el desempeño y con menor cantidad de ellas, con lo cual se logra que sea el mismo algoritmo quien decida, cuales son las variables útiles, aquellas que aportan descriminación y conocimiento sin redundancia al modelo y por ende un mejor desempeño. Esto podría llevarnos a suponer que se requiere nuevamente de una búsqueda exhaustiva, pero afortunadamente existe una herramienta del cómputo, la cual puede ser utilizada para hacer búsqueda inteligente, con el mismo propósito y con excelentes resultados, sin te-

ner que ir por el camino de la exhaustividad, esta herramienta son los algoritmos genéticos.

Con el uso de algoritmos genéticos, lo que se hace es optimizar la búsqueda, se pasa de una búsqueda exhaustiva a una inteligente, donde el objetivo es generación a generación (iteración a iteración), ir encontrando las combinaciones de variables que maximizan el desempeño del modelo. Los algoritmos genéticos estándar, solo pueden optimizar un parámetro a la vez, pero afortunadamente, la evolución de estos algoritmos nos ha llevado a los MOGA (Multi-Objetive Genetic Algorithm) o algoritmos genéticos multi-objetivo, con los cuales se puede optimizar más de una parámetro a la vez, lo cual nos interesa, contenga el algoritmo de selección de características que desarrollaremos en este trabajo.

Al utilizar un algoritmo genético se requiere crear individuos, en este caso estarán formados por todas los atributos y únicamente se activara o desactivara el uso de dichos atributos figura 4.1, los mejores individuos serán aquellos cuya combinación de atributos activos y no activos optimicen los objetivos pedidos al MOGA. 1 significa que la variable debe incluirse y 0 que el atributo no se incluirá en la creación del modelo. Cabe mencionar que esta forma de crear los individuos para la aplicación del MOGA, es la más comúnmente utilizada en los trabajos ya mencionados y en general en la literatura.

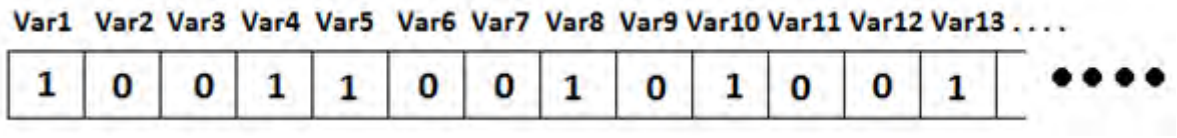


Figura 4.1: Individuos y sus Genes a Optimizar por el Algoritmo Gentico Multi-Objetivo

Se eligieron 4 objetivos a optimizar, los cuales juntos cumplen con el propósito de crear un buen modelo, estos parámetros a optimizar se enlistan a continuación:

- Desempeño por Score Decile.
- Desempeño por Top Percent.
- Área Bajo la Curva ROC [73]
- Número de Características (Atributos)

En resumen, la selección de características es la optimización del número de características que participan en el modelo, para lograr el mejor desempeño con el mínimo número de atributos utilizados para describir el fenómeno.

4.2. Algoritmo Genético Multi-Objetivo RankMOEA

La optimización es un proceso, el cual consiste en obtener el máximo o mínimo de un objetivo con el mínimo de recursos o esfuerzo, desde este punto de vista, los algoritmos genéticos son herramientas computacionales, para encontrar los mínimos y máximos globales, o muy cercanos a los objetivos planteados, sin tener que buscar en todo el espacio de búsqueda posible (lo cual se convertiría en una tarea imposible cuando la dimensionalidad del problema aumente), la búsqueda en los algoritmos genéticos es del tipo inteligente, esto es hace muestreos variados sobre el espacio de búsqueda, desechando aquellas zonas del espacio, donde no se encuentra indicios de candidatos óptimos (máximos o mínimos)

y priorizando aquellas donde se van obteniendo resultados buenos o mejores, esto se realiza de forma iterativa (a esto se le llama generaciones en los algoritmos genéticos), por lo cual en cada iteración los mejores individuos avanzan y los peores se quedan, existen muchas técnicas tanto para la selección de los individuos que pasan a la siguiente iteración como para la creación de nuevos individuos (mutación, cruza), como ya mencionamos los algoritmos genéticos simples fueron creados para optimizar un solo objetivo a la vez, de ahí la necesidad de crear algoritmos genéticos capaces de optimizar simultáneamente 2 o mas objetivos a la vez, esto dio paso a los denominados MOGA (multi-objetive genetic algorithm) o algoritmos genéticos multi-objetivo, los cuales dependiendo del número de objetivos a optimizar, complicaran su búsqueda y tendrán que utilizar mejores técnicas.

Existen gran variedad de algoritmos genéticos multi-objetivo en la literatura, cuya variación depende sobre todo, en la técnica utilizada para hacer selección, cruza y mutación, no existen algoritmos genéticos malos o buenos, o técnicas adecuadas, si no mas bien para determinados problemas unos funcionaran mejor que otros y viceversa, es por eso que la selección del algoritmo genético a utilizar depende más del problema a resolver y del tipo de resultado que se desea obtener, por tal motivo antes de hacer la selección del algoritmo a utilizar se debe leer cuidadosamente los resultados obtenidos con cada algoritmo y las conclusiones donde explican en cuales problemas funciona mejor o peor el algoritmo y que beneficios otorga sobre otros, para hacer una selección de la técnica adecuada acorde a las necesidades del problemas a resolver.

En el caso de este trabajo, el número de objetivos a optimizar es de 4 simultáneamente, los cuales, si bien es cierto no son totalmente independientes uno del otro, ya que tres de ellos tienen que ver con desempeño, si es necesario optimizar los 4 objetivos de forma simultanea, para lograrlo se requiere de utilizar un algoritmo genético multi-objetivo los cuales pueden optimizar más de dos objetivos a la ves simultáneamente. Existen varios tipos de algoritmos genéticos multi-objetivo, en este proyecto se decidió utilizar el Rank-MOEA [20], debido a ciertas caractersticas que lo favorecen: 1) trabaja bien con más de 2 objetivos simultáneos a optimizar (varios de los algoritmos existentes tienen problemas cuando se deben optimizar más de 2 objetivos de forma simultanea) y 2) ofrece como solución toda una gama de individuos igualmente óptimos y no comparables entre sí, lo cual se traduce en un conjunto de clasificadores, de los cuales, el usuario puede elegir el que más se adapte a sus necesidades.

4.3. Funcionamiento del Algoritmo

La forma en que el algoritmo diseñado en este proyecto trabaja es la siguiente:

1. Creación de archivos de entrenamiento y prueba: En este sentido podemos crear N archivos diferentes de entrenamiento y prueba, para hacer N diferentes selecciones aleatorias del mismo conjunto de datos, la proporción es 70 % para entrenamiento y 30 % prueba, esto se hace para evitar que un determinado conjunto de atributos tenga un buen funcionamiento solo por suerte, es decir, la partición 70/30 % resulto favorable para un buen desempeño usando determinado conjunto de atributos, pero al probarlo en otra partición 70/30 % no tiene el mismo desempeño, al hacer esto estamos evitando que esto pase, por medir el desempeño promedio y la desviación estándar, tras probar el conjunto de atributos a través de N distintas particiones 70/30 %.

2. Creación de arreglos binarios: Se leen todos los atributos que participaran en la creación del modelo y el objetivo de clasificación (clase), para después mapearlos a la

estructura mostrada en la figura 4.1, esto es, asignar a cada variable un sobrenombre (Var_i), donde i es el número de atributo, con esto se crea un arreglo de variables todas con valor 1, que significa usar la variable para la creación del clasificador (y 0 es no usarla), este arreglo es la primer entrada al algoritmo genético multi-objetivo junto con las N particiones 70/30 % de archivos de entrenamiento y prueba.

3. Aplicación del algoritmo genético: Este es en si, un modulo cerrado, dado que su funcionamiento ya es definido [20], las distintas técnicas y variaciones aplicadas pueden ser consultadas en el articulo mencionado en las referencias de ser de interés, los pasos de forma general que sigue el algoritmo genético son los siguientes:

   - Crear la población inicial: esto lo hace usando el arreglo de variables que se le proporciono como entrada, a partir del cual, genera M arreglos iguales, donde se varían aleatoriamente los contenidos del arreglo, es decir los 1 y 0, en otras palabras lo que hace en este paso el algoritmo es crear una población de tamaño M con diferentes conjuntos de atributos formados aleatoriamente, donde aquellos con valor 1, serán utilizados para entrenar el clasificador y los de valor 0 son desechados.
   - Entrenar los clasificadores: Al tener una población de M conjuntos de atributos diferentes, esto es igual a tener M clasificadores diferentes, lo cuales deben ser entrenados usando el clasificador Naive Bayes Estándar, lo cual es facilitado por la misma naturaleza de este clasificador (asume la independencia entre cada uno de los atributos), dado que los scores individuales de cada una de las variables pueden ser calculados de forma individual y as solo sumar aquellas donde exista valor 1, por cada conjunto (individuo).
   - Probar Clasificadores: Cada uno de los M clasificadores es aplicado a los N archivos de test creados, para al final obtener de estos archivos el desempeño en sus tres modalidades mencionadas (Score Decile, Top percent y Área bajo la curva ROC (AUC)), además del promedio y desviación estándar de estos desempeños.
   - Evaluación de objetivos: El algoritmo toma cada uno de los objetivos a optimizar y los evalúa para encontrar a los mejores individuos (grupos de atributos o clasificadores), los cuales pasan a formar parte de la población de la siguiente generación, las generaciones también son configurables, puede haber P de estas.
   - Creación de población siguiente generación: Los algoritmos genéticos funcionan a través de generaciones, a lo largo de cada una de las generaciones, la siguiente siempre contiene individuos mejores, o al menos igual de buenos que los de la generación anterior (esto medido por medio de los objetivos), hasta converger en la ultima generación al mejor o mejores de los individuos que la búsqueda inteligente pudo encontrar, individuo(s) óptimo(s), en este caso el clasificador óptimo (el conjunto de atributos mínimo con desempeño máximo) que existe dentro de los datos a modelar. La nueva población en cada generación esta formada por los mejores individuos de la generación anterior y los “hijos” de estos, los cuales son creados con la técnica denominada “Cruza”, en la cual los individuos (arreglos de 1 y 0 pertenecientes a cada atributo), son combinados para formar nuevos individuos (clasificadores o arreglos de atributos con combinaciones distintas de 1 y 0), y por último, se utiliza la técnica de “mutación” para cada uno de los individuos de la población, esto es, se altera uno de los valores (1 o 0 cambian a 0 o 1), esto con la finalidad de evitar estancamiento del algoritmo en mínimos o máximos locales, con esto da por finalizada la creación de nueva población.

- P iteraciones: Por último, los procesos descritos anteriormente deben repetirse P veces (generaciones), para al último entregar a los mejor individuos (conjuntos de atributos o clasificadores), que cumplen con la optimización de los objetivos propuestos al algoritmo.

4. Resultados: Dado que el algoritmo genético RANKMOEA entrega un conjunto de posibles soluciones, es decir clasificadores igualmente buenos y no comparables, se muestran todas estas junto con el clasificador, utilizando todos los atributos para que visualmente el usuario pueda decir cual de los clasificadores ofrecidos satisface mejor sus necesidades (como pueden ser mejor desempeño en el top 10 % o en los primeros 30 % o mejor AUC o menor desviación estándar o el mínimo número de atributos o combinaciones de los anteriores).

4.3.1. Pseudocódigo

A continuación se presenta el pseudocódigo del funcionamiento del algoritmo de selección de características:

Función Selección de Características():

{

Parámetros de configuración (entrada)
N = # de particiones del archivo
Arr_Var = Arreglo con los atributos totales
Mat_Datos = Totalidad del conjunto de datos para crear modelo

Parámetros de configuración algoritmo genético
M = Tamaño de la población
P = # de generaciones
Prob_mut = Probabilidad de mutación
Prob_cruza = Probabilidad de cruza

For i= 1 hasta N, (crear las particiones archivos test y train)
train_i = Aleatorio(70 % Mat_Datos)
test_i = Complemento(Mat_Datos - train_i)

Algoritmo Genético[Arr_Var, Array(con N train files), Array(con N test files), M, P, Prob_mut, Prob_cruza]
{
For j = 1 hasta M, (crear población inicial)
Individuo_i = Aleatorio(Arr_Var, 1s y 0s)

Población = Arreglo(M individuos)

For k = 1 hasta P, (itera por las p generaciones)
Scores = Entrenar [Población, Array(con N train files)],
(entrena cada individuo N veces una por train file)

Scoring = Aplica_Scores [Scores, Array(con N test files)],
(aplica los scores de cada individuo N veces una por test file)

Score_Decile = Evalúa [Scoring],
(calcula score decile para cada test file e individuo)

Top_Percent = Evalúa [Scoring],
(calcula top percent para cada test file e individuo)

AUC = Evalúa [Scoring],
(calcula Área bajo Curva ROC para cada test file e individuo)

Sel_mejores = [individuos, Score_Decile, Top_Percent, AUC],
(Selecciona a los mejores individuos para crear nueva población)

Población = [Sel_mejores, Prob_mut, Prob_cruza],
(Crea nueva población)

}

Resultado = Sel_mejores,
(los mejores individuos o clasificadores después de las P generaciones)

Imprime Resultado;

}

4.4. Resultados Selección de Características

Al igual que en el capítulo anterior, presentaremos los resultados de aplicar el algoritmos de selección de características a los conjuntos de datos ya descritos en el capitulo previo, con la intención de comparar los resultados obtenidos al utilizar Niave Bayes Estandar sin ningún tipo de filtro, en las atributos usados en el modelo, obviamente descartando los atributos de por si obvios, como fueron los atributos llave o con exceso de nulos.

4.4.1. ENCOPREVENIMSS 2006

Aplicando el algoritmo de selección de características a los datos de las encuestas hechas por el IMSS en 2006, los cuales fueron introducidos en el capítulo anterior, con los siguientes parámetros de configuración:

- Generaciones = 500
- Probabilidad Mutación = 0.01
- Probabilidad Cruza = 0.9
- Número Archivos Train/Test = 5

Se obtuvieron los resultados mostrados en la tabla 4.1 y 4.2

Como podemos observar en la tabla 4.1, al aplicar el algoritmo de selección de características (ASC), a la vista minable del modelo creado en 3.6.1, para el diagnostico de diabetes en hombres, obtenemos una disminución del 95.7 % en el número de atributos utilizados para la creación del modelo, sin perder predictibilidad ni desempeño, al contrario se obtiene una mejora en las medidas del mismo, como son: 28.9 % mejor para la métrica Top 5 %, 15.5 %

	Modelo Hombre				
	Con SC	Sin SC	ACP1	ACP2	Épsilon
Desempeño Promedio Top 5 %	44.43	34.46	26.42	29.56	29.62
Desempeño Promedio Top 3 Deciles	88.95	76.98	68.28	71.32	70.57
Área Bajo Curva ROC Promedio	0.8835	0.8295	0.7302	0.7512	0.7631
# Atributos	7	163	7	54	54
Varianza Top 5 %	3.3447	9.0193	8.8721	9.1525	9.3211
Varianza Top 3 Deciles	0.4948	8.9091	8.3510	8.6123	8.7723
Varianza Cuva ROC	0.00001	0.00016	0.00042	0.00035	0.00032
% mejora en desempeño Top 5 %	28.9 %		–	–	–
% mejora en desempeño Top 3 Deciles	15.5 %		–	–	–
% mejora en curva ROC	6.5 %		–	–	–
% disminucin de # atributos	95.7 %		95.7 %	66.87 %	66.87 %

Tabla 4.1: Resultado de aplicar el algoritmo de selección de características a datos ENCOPREVENIMSS 2006 (Modelo Hombre).

mejor para la métrica top 3 deciles y 6.5 % mejor para curva ROC, por lo tanto podemos concluir, que en este caso en especifico, los beneficios de aplicar el algoritmo de selección de características van mas allá de la disminución del número de variables requeridas para la creación del modelo, por que además se consigue una mejora en el desempeño del modelo creado. Por otra parte utilizando técnicas de selección de características por variable como son: Análisis de Componentes Principales (ACP) y épsilon, obtenemos resultados similares en desempeño al modelo sin filtros e incluso ligeramente peores en desempeño, ACP1 significa utilizar las primeras n componentes principales, donde n es el nmero de variables mínimo encontrado por el algoritmo de selección de características, en este caso 7, ACP2 son las n componentes principales, donde n es el número de variables elegidas a través de la técnica de épsilon. Esta elección de componentes principales se hace con el propósito de poder comparar los diferentes algoritmos con ACP.

En resumen, los 7 atributos seleccionados por el ASC se presentan en la tabla 4.3, donde podemos ver que el resultado del algoritmo y las variables seleccionadas por el mismo son bastante lógicas, en primer lugar sobre el ejercicio, esta pregunta en especifico descarta una multitud de ejercicios posibles, entre ellos los de mayor demanda fsica, al poner otro, el paciente puede estar escondiendo el hecho de que no realiza ninguno, o lo que el considera como ejercicio no es algo muy demandante fisícamente, por lo cual no aportaría mucho a la buena salud, la segunda es sobre el conocimiento en contagio de enfermedades sexuales, lo que se puede extrapolar a un conocimiento mas bien general sobre la salud y su cuidado, lo cual es muy importante, la tercera habla sobre diagnostico previo de presión arterial alta, sabemos que la presión arterial esta muy relacionada a la obesidad y esta a la diabetes, la cuarta es sobre si lo han enviado a exámenes de laboratorio lo que nos

	Modelo Mujer				
	Con SC	Sin SC	ACP1	ACP2	Épsilon
Desempeño Promedio Top 5 %	54.13	42.52	37.67	39.21	39.11
Desempeño Promedio Top 3 Deciles	92.08	84.63	77.54	80.13	80.02
Área Bajo Curva ROC Promedio	0.9081	0.8655	0. 8323	0.8543	0.8402
# Atributos	17	154	17	69	69
Varianza Top 5 %	0.9787	4.4255	4.6250	4.5390	4.6521
Varianza Top 3 Deciles	0.0305	0.6971	0.8439	0.7892	0.7932
Varianza Cuva ROC	0.00000	0.00001	0.00025	0.00034	0.00043
% mejora en desempeño Top 5 %	27.3 %		–	–	–
% mejora en desempeño Top 3 Deciles	8.7 %		–	–	–
% mejora en curva ROC	4.9 %		–	–	–
% disminucin de # atributos	88.9 %		88.9 %	57.66 %	57.66 %

Tabla 4.2: Resultado de aplicar el algoritmo de selección de características a datos ENCOPREVENIMSS 2006 (Modelo Mujer).

puede indicar tanto estudios de prevención o por que el medico ha sospechado de alguna enfermedad, la quinta es parecida a la de ejercicio descarta una serie de enfermedades como son: obesidad, presión arterial, angina, colesterol alto, etc., y la persona opta por decir que ninguna enfermedad de estas padece, lo cual si lo pensamos, es muy buen atributo para discriminar, tanto si es cierto como si no, esto es, si no padece ninguna enfermedad esto nos puede hablar de una persona muy saludable, la cual efectivamente no padecerá diabetes pronto o al menos no hay indicios de esto, mientras mantenga este estatus, o bien puede hablar de un desconocimiento de salud propio, esto puede a su ves deteriorar mucho a la personas, al extremo de llevarlo a la enfermedad, la sexta habla sobre la genética, en este caso, si la madre ha transmitido estos genes propensos a la enfermedad o también puede hablar de un estilo de vida de la madre que puede reflejarse también en los hijos y la ultima es edad, sabemos que la edad esta muy relacionada con el padecimiento de la diabetes. Como vemos la selección de características hecha de forma automática por el algoritmos es bastante lógica y analizando una por una, podemos entender por que estas son las seleccionadas por ASC y a su vez corroborar que el algoritmo hace un buen trabajo discriminando atributos útiles.

Para el caso del modelo de mujer tabla 4.2, al aplicar ASC a su vista minable del modelo creado en cap. 3.6.1, para el diagnóstico de diabetes en mujeres, se obtienen los beneficios mostrados en la tabla 4.4, donde podemos ver que obtenemos una disminución del 88.9 % en el número de atributos utilizados para la creación del modelo, sin perder predictibilidad ni desempeño, al igual que en el caso anterior (modelo de hombre), hay una mejora en todas las métricas de la siguiente forma: 27.3 % para el Top 5 %, 8.7 % para los Top 3 Deciles y 4.9 % para la AUC, en la tabla 4.4 se muestra el resumen de las variables que

Atributo	Descripción
HPS712	Que tipo de ejercicio hace? Respuesta: Otro
HPS10	Conoce una manera para evitar el contagio de enfermedades sexuales?
HDE26	Algún medico/enfermera te ha dicho que tienes presión arterial alta?
HUS19	Se hizo exámenes de laboratorio?
HFR17	Padeces alguna de estas enfermedades? Respuesta: Ninguna
HAFM21	Su mamá padeció o padece diabetes?
CG_HACM	Edad

Tabla 4.3: Resumen de atributos seleccionado por el ASC para modelo hombre

fueron seleccionadas por el ASC, donde podemos observar otra vez que las elecciones de atributos hechas por el ASC son igual de lógicas que en el modelo para hombres, contempla los atributos de herencia familiar, información sobre salud en general, alimentación y ejercicio, así como el cuidado de la salud. Y también podemos observar que para ACP y épsilon se obtienen resultados similares a los encontrados en el caso de modelo de hombre.

Atributo	Descripción
ALIMENT	Ha recibido información de cmo cuidar su alimentación?
MPS61	En que institucion recibio informacion (sobre importancia del ejercicio)?
MPS74	Que tipo de ejercicio (voleibol)?
MPS83	En que institución recibió información (sobre temas de sexualidad)?
MPSRE225	Problemas con el peso Le hicieron alguna recomendación? opción:(no le recomendaron nada)
MSS19	se hizo los exámenes de laboratorio? La ultima ocasin
MFR7	Padece alguna de las siguiente enfermedades? (ninguna) entre ellas obesidad, colesterol
MFR8	Padece alguna de las siguiente enfermedades? (no sabe) entre ellas obesidad, colesterol
MAFP2	Algún familiar padece o padeció diabetes mellitus? (padre)
MAFM2	Algún familiar padece o padeció diabetes mellitus? (madre)
MAFM3	Algún familiar padece o padeció presión alta? (madre)
MAFM6	Algún familiar padece o padeció sobrepeso? (madre)
MAFH2	Algún familiar padece o padeció diabetes mellitus? (hermanos)
colesterol	Tiene colesterol?
CG_AGEM	Edad
CG_MSSA1	Número de veces que solicito atención medica ambulatoria ?
CG_DIFCENTIM	Medición de cintura en centímetros

Tabla 4.4: Resumen de atributos seleccionado por el ASC para modelo mujer

4.4.2. DxCG 97-99

Ahora aplicando el algoritmo de selección de características a los datos obtenidos para DCG (Diagnostic Cost Group), analizados en cap. 3.6.2 con los siguientes parámetros de configuración:

- Generaciones = 500
- Probabilidad Mutación = 0.01
- Probabilidad Cruza = 0.9
- Número Archivos Train/Test = 5

Para este caso se aplicaron 2 iteraciones del ASC sobre la vista minable del modelo DxCG, como podemos observa en la tabla 4.5, se obtuvieron disminuciones en el número de atributos en cada una de las iteraciones, en la primera se disminuye de 1160 a 404 y en la segunda de 404 a 82.

	Modelo sin SC	Modelo Con SC	
		Iteración 1	Iteración 2
Desempeño Promedio Top 5 %	33.92	36.26	35.73
Desempeño Promedio Top 3 Deciles	72.05	72.37	75.88
Área Bajo Curva ROC Promedio	0.7958	0.7929	0.8102
# Atributos	1160	404	82
Varianza Top 5 %	3.2540	2.1333	0.7680
Varianza Top 3 Deciles	2.2011	0.7646	0.3793
Varianza Cuva ROC	0.00006	0.00001	0.00000
% mejora en desempeño Top 5 %	x	6.8 %	5.3 %
% mejora en desempeño Top 3 Deciles	x	0.4 %	5.3 %
% mejora en curva ROC	x	0.3 %	1.8 %
% disminucin de # atributos	x	65.7 %	92.9 %

Tabla 4.5: Resultado de aplicar el algoritmo de selección de características a datos DxCG 97-99.

Además de la gran disminución en el número de atributos, después de las dos iteraciones, 92.9 %, podemos observar que no existe un gran aumento en los desempeños del modelo, pero si lo existe en la varianza de las diferentes metricas de desempeño, lo cual es también una parte muy importante en la creación de clasificadores. Esto nos garantiza que sobre multiples subconjuntos de datos o particiones que se puedan hacer del conjunto de entrenamiento, obtendremos resultados similares en los desempeños, lo cual evita depender de una buena selección de un subconjunto de datos, para obtener buenos modelo, esto lo podemos atribuir a que los atributos, los cuales generan mayor "ruido"fueron ya descartados por el ASC, quedando así un subconjunto de atributos con los cuales podemos obtener modelos congruentes y estables. La tabla 4.6 muestra la aplicación de ACP y épsilon para este modelo y vemos un comportamiento similar al de los modelos mostrados anteriormente.

	ACP1	ACP2	épsilon
Desempeño Promedio Top 5 %	29.22	31.63	30.93
Desempeño Promedio Top 3 Deciles	67.51	69.72	70.80
Área Bajo Curva ROC Promedio	0.7782	0.7812	0.7802
# Atributos	82	365	365
Varianza Top 5 %	4.0012	4.5323	4.6032
Varianza Top 3 Deciles	3.0121	2.9661	2.9321
Varianza Cuva ROC	0.00026	0.00011	0.00014
% mejora en desempeño Top 5 %	x	x	x
% mejora en desempeño Top 3 Deciles	x	x	x
% mejora en curva ROC	x	x	x
% disminucin de # atributos	92.9 %	68.5 %	68.5 %

Tabla 4.6: Resultado de aplicar ACP y épsilon como filtros para seleccionar caractersticas a datos DxCG 97-99.

Podemos considerar que los resultados obtenido con las técnicas de ACP y épsilon, se deben a que estas priorizan solamente la disminución de variables y no tienen dentro de su algoritmo ningún parámetro el cual intente mantener un desempeño igual o mejor al obtenido utilizando todas las variables, por lo tanto los resultados observados para los distintos modelos son comprensibles y lógicos, tomando en cuenta la naturaleza de estos algoritmos.

Capítulo 5

Naive Bayes Generalizado

5.1. Introducción [28]

El clasificador Naive Bayes (NBC), es uno de los algoritmos más utilizados en aprendizaje de máquina (machine learning) y minería de datos. Se han utilizado en múltiples y diferentes áreas de aplicación y ha demostrado ser notablemente robusto, tanto en aplicabilidad, como en rendimiento. De hecho, la robustez en su desempeño ha sido siempre un enigma, dada su fuerte suposición de la independencia entre las variables, lo cual ha dado lugar a muchos artículos, donde se trata de comprender y explicar, por que NBC parece ser injustificadamente exitoso.

También se han escrito muchos artículos relacionados con diferentes propuestas de generalización de NBC, para eludir la suposición de independencia entre las variables. Sin embargo, estas generalizaciones del NBC, son invariablemente más complicadas de implementar, y de mayor consumo de recursos (procesamiento) que el NBC. Por lo tanto, es importante desarrollar diagnósticos con los que, para un problema dado, se pueda medir cuando y sí el NBC dará lugar a un considerable error, con respecto a un clasificador más sofisticado, que intente tomar en cuenta las posibles correlaciones entre las variables,mientras que al mismo tiempo, desarrolla una profunda comprensión intuitiva y teórica de como y por que el NBC es tan robusto. En este trabajo, analizamos en que circunstancias se puede esperar que la Aproximación de Naive Bayes (NBA) y el NBC asociado, son subóptimos y desarrollar diagnósticos generales, con los que un problema puede ser examinado a priori, para determinar si el NBA es adecuado o si se requiere de una generalización más sofisticada.

En cuanto a explicar el funcionamiento robusto del NBC Domingos y Pazzani [29], han argumentado que es debido en gran parte a su aplicación en problemas de clasificación, donde los errores son contados con respecto a si la clasificación fue correcta (si/no), para un pronostico dado, y no si la probabilidad estimada correspondiente fue exacta. Por lo tanto, si pensamos en una lista ranqueada de las predicciones, donde el umbral de clasificación esta en el elemento n-esimo, entonces no importa si un elemento dado esta ranqueado como primero o (n - 1), sera asignado a la misma clase. Más material para apoyar esta hipótesis proviene de la obra de Frank et al. [30], quien mostró que el rendimiento del NBC es sustancialmente peor cuando se aplica a problemas de tipo regresión. La pura precision de clasificación , sin embargo, es solo una y única, medida global de rendimiento del clasificador y hay otros que pueden ser mas apropiados. Por ejemplo se ha demostrado [31], que el área bajo la curva ROC, es una medida mas completa que la pura exactitud de clasificación. A menudo de mayor interes son los scores de riesgo relativo, una medida de desempeño interesante es la exactitud del clasificador, con un conjunto especifico de

grupos de riesgo. Este es especialmente el caso en problemas tales como costos en salud, donde un grupo muy pequeño (1 % de la población), pero de muy alto riesgo, puede generar una gran parte de los costos en salud (30 % de los costos totales). En esta circunstancia es muy poco probable que un clasificador sea suficientemente preciso, que pueda colocar a alguien en un grupo tan pequeño. Más bien, lo que interesa, es el grado relativo riesgo de un paciente a otro [32]. Este tipo de razonamiento que va mas allá de pura clasificación, el donde el NBC puede ser más rigurosamente juzgado, esto es circunstancialmente apoyado por el hecho de que la NBA puede producir estimaciones de probabilidad pobres [33, 34], aunque en [35] se demostró que los modelos NBA pueden ser tan eficaces como las redes Bayesianas más generales para tareas de estimación de probabilidad.

Sin embargo, incluso en el caso de clasificación, como se ha señalado por Zhang [36][37], el argumento de Domingos y Pazzani no explica por que no es posible tener situaciones donde las estimaciones de probabilidad inexactas voltean la clasificación, lo que conduzca a un rendimiento inferior. Zhang ha propuesto que no se trata sólo de la presencia de dependencias entre los atributos, lo que afecta al rendimiento, si no la forma en la cual se distribuyen, entre las diferentes clases lo que juega un papel crucial en el desempeño del NBC, argumentando que el efecto de las dependencias se puede cancelar parcialmente entre las clases y además, las dependencias pueden potencialmente cancelarse entre diferentes subconjuntos de valores de características. La pregunta entonces es, si esto es posible, bajo cuales condiciones ocurrirá y si podemos cuantificarlo, y por lo tanto, predecir a priori cuando el uso de NBA puede ser inadecuado?.

En este capítulo, investigaremos cuantitativamente, y a detalle, la relación entre la dependencia del atributo y el error, mediante el análisis de un conjunto de problemas tipo, proporcionaremos herramientas para investigar el conjunto de dependencias entre atributos, y mostrar como afectan, tanto a las estimaciones de probabilidad, como a la precisión en clasificación. Lo que haremos será proporcionar diagnósticos estadísticos, los cuales permitan estimar tanto, por intuición, como por capacidad predictiva cuando el NBA puede fracasar. Como los conteos para la correlación de los atributos son cuestión de sesgo y no de varianza del modelo, como en [38], vamos a considerar "muestras infinitas" de distribuciones de probabilidad artificiales, elegidas con el fin de poder afinar el grado de correlación entre los diferentes atributos, sin tomar en cuenta los efectos de muestreos finitos. De hecho, es precisamente la existencia de error de muestreo en problemas reales, lo que se asocia con el desempeño superior de la NBA, a pesar de las dependencias entre los atributos [39].

La segunda cuestión más importante, gira entorno a como mejorar la NBA y NBC. Han habido muchas generalizaciones de la NBA [39] [40]. Algunos como Lazy Bayesian Rules [41], Super Parent TAN [42] and Hidden Naive Bayes [40], han demostrado tener muy buen rendimiento, con mejoras significativas a la NBA, pero a un costo computacional considerable. Un buen resumen de muchos de estos algoritmos, se puede encontrar en [41, 40]. Como no es el propósito de este trabajo introducir un "nuevo algoritmo competitivo", nos limitaremos a algunos comentarios generales. Todas estas generalizaciones buscan descubrir conjuntos de atributos, cuyos valores tengan dependencias tales que deberían o podrían combinarse entre si, o con la clase. En general, son de tal forma que la mejora asociada con la combinación de un conjunto de características, se juzga a posteriori a través del desempeño relativo del algoritmo con y sin esa combinación. Sin embargo como hay un gran número de posibles combinaciones de valores de atributos, las cuales deben considerarse, el proceso de selección de atributos debe ser intensivo, de tal forma que generalmente, los estudios se han limitado a considerar solo pares de atributos, realizando

una búsqueda exhaustiva solo de esas combinaciones.

La eficacia de estas generalizaciones de la NBA, son juzgadas generalmente por comparar el rendimiento del algoritmo propuesto en contra de la NBA, y potencialmente un conjunto elegido de otros algoritmos NBA, en un conjunto de problemas de prueba canónicos, mas frecuentemente que no fueron tomados del repositorio de UCI (Universidad de California Irvine). La eficacia del nuevo algoritmo, es inferida globalmente a través del conjunto de problemas ejemplo considerados. Sabemos por los teoremas de "No-Free Lunch" [43, 44], que ningún algoritmo es mejor a cualquier otro en todos los problemas existentes. La pregunta es: ¿podemos inferir a priori cuales algoritmos obtienen mejores resultados en un ejemplo de problema dado?. Esto es especialmente importante, si las mejoras en el rendimiento están dominadas por solo un pequeño número de instancias, también requiere una visión detallada de cómo y por qué una generalización dada supera a la NBA, en algunos conjuntos de datos y en otros no.

Un subproducto de la elaboración de las generalizaciones de la NBA, ha sido la construcción de diagnósticos, para determinar el grado de dependencia de los atributos, y por lo tanto, detectar cuales características deben ser combinadas potencialmente. El diagnóstico más utilizado, ha sido el de información condicional mutua [36, 38, 45]. Sin embargo, como Rish ha señalado, esta medida no se correlaciona bien con el desempeño de la NBA, argumentando que un mejor predictor de la precisión, es la perdida de información que sufren los atributos, con respecto a la clase, cuando se asume un modelo NB. Lo que se requiere es una medida de la dependencia del atributo, la cual se relacione directamente con la aparición del error correspondiente en la NBA y, además, cómo estos errores se combinan para producir un error global, para un determinado conjunto de características o clasificador.

5.2. Comparando Clasificadores

5.2.1. Aproximación Naive Bayes

Tratando de entender en cuales circunstancias podríamos esperar que la NBA y NBC "fallaran", vamos a comenzar con el teorema de Bayes

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})} \tag{5.1}$$

para una clase C y un vector de N características $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, donde $P(C)$ es la probabilidad anterior, $P(\mathbf{X}|C)$ es la función de probabilidad dados los datos $\mathbf{X}$ y $P(C|\mathbf{X})$ la probabilidad posterior.

Desafortunadamente cuando $\mathbf{X}$ es de alta dimensionalidad, hay demasiadas probabilidades diferentes $\hat{P}(C|\mathbf{X})$ para estimar. Relacionado con esto está el hecho de que $N_{C\mathbf{X}}$ (el número de elementos en $\mathbf{X}$ y C), es generalmente tan pequeño que las estimaciones estadísticas $\hat{P}(C|\mathbf{X})$ de $P(C|\mathbf{X})$, no son confiables debido al gran error de la muestra. [1] Usar el teorema de Bayes no mejora el problema, dado que las estimaciones estadísticas de $\hat{P}(\mathbf{X}|C)$ sufren del mismo problema. Sin embargo si hay independencia estadística de las X_i en la clase C, entonces $P(\mathbf{X}|C) = \prod_{i=1}^{N} P(X_i|C)$, donde $P(X_i|C)$ es la probabilidad marginal condicional para X_i dado C. Generalmente, este no es el caso, sin embargo se puede hacer la suposición de que es aproximadamente verdadera, tomando

[1] De hecho, para un conjunto suficientemente grande de características discriminatorias de las cuales cada combinación es única, tendremos $N_{C\mathbf{X}} = 0,\ 1$ con la mayoría de las combinaciones siendo cero.

$P_{NB}(\mathbf{X}|C) = \prod_{i=1}^{N} P(X_i|C)$ y aproximando (5.1) as

$$P_{NB}(C|\mathbf{X}) = \frac{\prod_{i=1}^{N} P(X_i|C)P(C)}{(\prod_{i=1}^{N} P(X_i|C)P(C) + P(\mathbf{X}|\bar{C})P(\bar{C}))} \tag{5.2}$$

donde $\bar{C}$ es el complemento de C. Por supuesto, si tuviéramos que calcular $P(C|\mathbf{X})$ utilizando (5.2), habría que estimar también $P(\mathbf{X}|\bar{C})$, lo cual presenta los mismos problemas como la estimación de $P(\mathbf{X}|C)$. La misma aproximación naive se puede utilizar en este caso también, escribiendo $P(\mathbf{X}|\bar{C}) = \prod_{i=1}^{N} P(X_i|\bar{C})$. Note que en este caso la NBA, ha sido no solo aplicada a $P(\mathbf{X}|C)$, si no también a $P(\mathbf{X}|\bar{C})$. En otras palabra que las X_i son independientes cuando condicionan a cada clase. Como $P(\mathbf{X}) = P(\mathbf{X}|C)P(C) + P(\mathbf{X}|\bar{C})P(\bar{C})$, esto implica que

$$P_{NB}(\mathbf{X}) = \prod_{i=1}^{N} (P(X_i|C)P(C) + P(X_i|\bar{C})P(\bar{C})) \tag{5.3}$$

lo cual puede ser considerado como la NBA para $P(\mathbf{X})$.

En lugar de construir $P(C|\mathbf{X})$ directamente, a través de la ecuación 5.1), usualmente una función score, $S(\mathbf{X})$ que es una función monótona de $P(C|\mathbf{X})$ en sí, es construida considerando la razón de posibilidades de la clase C y otra clase, por lo general su complemento, $\bar{C}$.

$$S(\mathbf{X}) = \log \frac{P(C|\mathbf{X})}{P(\bar{C}|\mathbf{X})} = \log \frac{P(C)}{P(\bar{C})} + \log \frac{P(\mathbf{X}|C)}{P(\mathbf{X}|\bar{C})}$$

La NBA a esta función de score, $S_{NB}(\mathbf{X})$, la cual es una función monótona de $P_{NB}(C|\mathbf{X})$, aunque no necesariamente una función monótona de $P(C|\mathbf{X})$ en si, esta dada por

$$S_{NB}(\mathbf{X}) \quad = \quad \log \frac{P(C)}{P(\bar{C})} + \sum_{i=1}^{N} \log \frac{P(X_i|C)}{P(X_i|\bar{C})} = \log \frac{P(C)}{P(\bar{C})} + \sum_{i=1}^{N} S(X_i) \tag{5.4}$$

donde, una vez mas, la NBA se ha aplicado a las funciones de probabilidad para ambos C y $\bar{C}$.

Como una simple suma esta forma de la aproximación es transparente. Otra ventaja de esta forma, es que no es necesario tener en mano $P(\mathbf{X})$, dado que la idea de una función de score es solo discriminar entre las clases C y $\bar{C}$. Esta es una tarea diferente de la estimación de la probabilidad $P(C|\mathbf{X})$ directamente. Como un clasificador (5.4) es talque si $S(\mathbf{X}) > 0$ entonces el ejemplar definido por $\mathbf{X}$ se asigna a la clase C y si $S(\mathbf{X}) < 0$ al complemento $\bar{C}$.

5.2.2. Aproximación Naive Bayes Generalizado

La NBA y NBC se basan en una factorización máxima de la función de probabilidad, que aparece en la parte derecha del teorema de Bayes. Esencialmente todas las generalizaciones de la NBA que han sido consideradas están asociadas con introducir dependencias entre los atributos de tal forma que buscan una factorización alternativa para las funciones de probabilidad. Con el fin de contar con un marco teórico concreto, en el cual examinar la validez de la NBA, el marco de las redes bayesianas [45], y en particular, el clasificador semi-Naive bayesiano [46] son particularmente apropiados.

Para definir generalizaciones de la NBA y NBC, asociados con factorizaciones alternativas, se debe introducir primero una notacion de "esquema" basada en probabilidades

marginales del tipo usado en algoritmos genéticos [47, 48], donde para cualquier vector $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, una probabilidad marginal $P(X_{i_1} X_{i_2} \ldots X_{i_m}|C)$ puede escribirse $P(*^{i_1-1} X_{i_1} *^{i_2-i_1} X_{i_2} \ldots X_{i_m}|C)$, donde $*^n$ significa $*$ repetido n veces y un $*$ en la ith posición X_i significa que X_i ha sido marginalizado. El orden del esquema es sólo el número de variables no marginalizadas. Asi, por ejemplo, los posibles marginales de $P(X_1X_2X_3|C)$ son los tres de orden 2 $P(X_1X_2 * |C)$, $P(X_1 * X_3|C)$ y $P(*X_2X_3|C)$, los tres de oren 1 $P(X_1 * *|C)$, $P(*X_2 * |C)$, $P(* * X_3|C)$; y finalmente de orden cero $P(* * *|C) = P(C)$. La NBA usa solo esquemas orden 1. Entonces podemos denotar una combinación arbitraria del valor m-atributo por esquema $\xi = (\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m})$, de orden $m < N$.

Con esta notación en mano, podemos definir la Aproximación Generalizada de Bayes (GBA) y el correspondiente Clasificador Generalizado de Bayes (GBC), a través de las generalizaciones de las ecuaciones (5.2) y (5.4), que correspondan a factorizaciones alternativas de las funciones de probabilidad. Un elemento importante en este caso, sin embargo, es que a diferencia de la NBA, la GBA no es única, ya que hay un gran número de posibles factorizaciones para un conjunto especifico de valores de atributos dado. Para N atributos hay B_N particiones, donde B_N es el número de Bell. Peor aún, este número no es para cada valor de un atributo y hay $\mathcal{R} = \prod_{i=1}^{n} a(i)$ combinaciones de valores de atributos, donde $a(i)$ es la cardinalidad del ith atributo. Aunque el mismo esquema puede aparecer en múltiples conjuntos de atributos, obviamente, una buena búsqueda heurística, como un algoritmo genético, es necesaria para poder muestrear adecuadamente este vasto espacio.

Por supuesto, tanto en las estimaciones de probabilidad, como en los clasificadores la pregunta es: ¿de todas las posibles factorizaciones cual es óptima y como definimos óptima? Por ejemplo, para el caso de tres variables, de las 4 posibles factorizaciones ¿cual da la mejor aproximación a $P(X_1X_2X_3|C)$?. Gran parte de este capítulo se ocupara de la cuestión de como definir mejores factorizaciones usando diagnósticos. Vamos a denotar el GBA asociado con una factorización por un conjunto de N_ξ esquemas $\xi^{(i)} = \cup_{\alpha=1}^{N_\xi} \xi^\alpha$. Tenga en cuenta que cualquier factorización debe corresponder a una partición del conjunto de valores de atributos N. Por lo tanto, para cualquier vector dado $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, cada X_i debe ser miembro de uno y sólo un esquema $\xi^\alpha \in \xi^{(i)}$. Por ejemplo, considere el caso anterior de 3 atributos X_1, X_2 y X_3 y la función asociada de probabilidad $P(X_1X_2X_3|C)$. Hay 4 posibles factorizaciones: $P(X_1X_2X_3|C) = P(X_1 * *|C)P(*X_2 * |C)P(* * X_3|C)$, $P(X_1X_2X_3|C) = P(X_1 * *|C)P(*X_2X_3|C)$, $P(X_1X_2X_3|C) = P(X_1 * X_3|C)P(*X_2 * |C)$, $P(X_1X_2X_3|C) = P(X_1X_2 * |C)P(* * X_3|C)$, y análogamente para $P(X_1X_2X_3|\bar{C})$. Por lo tanto, tenemos la GBA para las funciones de probabilidad

$$P_{GB}(\mathbf{X}|C) = P(\xi^{(i)}|C) = \prod_{\alpha=1}^{N^C_{\xi^{(i)}}} P(\xi^\alpha|C) \tag{5.5}$$

y

$$P_{GB}(\mathbf{X}|\bar{C}) = P(\xi^{(j)}|\bar{C}) = \prod_{\alpha=1}^{N^{\bar{C}}_{\xi^{(j)}}} P(\xi^\alpha|\bar{C}) \tag{5.6}$$

donde $N^C_{\xi^{(i)}}$ es el número de marginales independientes utilizadas en la GBA, para la función de probabilidad para las clase C y $N^{\bar{C}}_{\xi^{(j)}}$, es el número de marginales independientes utilizadas en el GBA, para la función de probabilidad para $\bar{C}$. En la NBA $N^C_{\xi^{(i)}} = N^{\bar{C}}_{\xi^{(j)}} = N$. Tenga en cuenta que en este nivel de generalidad no hacemos una suposición a priori

de que la factorización óptima de la función de probabilidad de C, es la misma que para $\bar{C}$, aunque generalmente se asume esto. En otras palabras los conjuntos de esquemas $\xi^{(i)}$ y $\xi^{(j)}$ no son necesariamente los mismos.

Para la probabilidad posterior tenemos

$$P_{GB}(C|\mathbf{X}) = P(C|\xi^{(i)}) = \frac{\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C)P(C)}{(\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C)P(C) + \prod_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} P(\xi^\alpha|\bar{C})P(\bar{C}))} \tag{5.7}$$

y finalmente, para la función de score

$$S_{GB}(\mathbf{X}) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_{\xi^{(i)}}^C} S^C(\xi^\alpha) - \sum_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} S^{\bar{C}}(\xi^\alpha) \tag{5.8}$$

donde definimos

$$S^C(\xi^{(i)}) = \sum_{\alpha=1}^{N_{\xi^{(i)}}^C} \ln P(\xi^\alpha|C) \qquad S^{\bar{C}}(\xi^{(j)}) = \sum_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} \ln P(\xi^\alpha|\bar{C}) \tag{5.9}$$

como las contribuciones al score global de la probabilidad de la clase y su complemento respectivamente.

Si hacemos la suposición simplista de que la factorización óptima de las funciones de probabilidad para C y $\bar{C}$ son las mismas entonces (5.8) se simplifica a

$$S_{GB}(\mathbf{X}) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_\xi} S(\xi^\alpha) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_\xi} \ln \frac{P(\xi^\alpha|C)}{P(\xi^\alpha|\bar{C})} \tag{5.10}$$

la cual es una generalización natural de la función de score en la NBA.

El GBC es entonces de tal forma, que un vector de atributos $\mathbf{X}$ pertenece a la clase C, si $\ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_\xi^C} \ln P(\xi^\alpha|C) > \sum_{\alpha=1}^{N_\xi^{\bar{C}}} \ln P(\xi^\alpha|\bar{C})$ y a la clase $\bar{C}$ en caso contrario.

5.2.3. La Diferencia

Cualquier diferencia entre la NBA y la GBA, tiene su origen en la existencia de correlaciones entre el atributo X_i y C y $\bar{C}$. En términos de sesgo del modelo, la factorización más apropiada de las funciones de probabilidad, debería ser aquella, la cual respete mas la existencia de tales correlaciones. Para una factorización dada, podemos determinar las diferencias entre la GBA y la NBA de las ecuaciones (5.5), (5.6), (5.7) y (5.8). Para las probabilidades tenemos

$$\begin{aligned} \Delta_{P_{GB}}(\mathbf{X}|C) &= P(\xi^{(i)}|C) - P_{NB}(\mathbf{X}|C) \\ &= \prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C) - \prod_{i=1}^{N} P(X_i|C) \end{aligned} \tag{5.11}$$

$$\begin{aligned} \Delta_{P_{GB}}(\mathbf{X}|\bar{C}) &= P(\xi^{(j)}|\bar{C}) - P_{NB}(\mathbf{X}|C) \\ &= \prod_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} P(\xi^\alpha|\bar{C}) - \prod_{i=1}^{N} P(X_i|C) \end{aligned} \tag{5.12}$$

Una propiedad importante de las diferencias (5.11) y (5.12) es que satisfacen la ecuación

$$\sum_{\mathbf{X}} \Delta_{P_{GB}}(\mathbf{X}|C) = \sum_{\mathbf{X}} \Delta_{P_{GB}}(\mathbf{X}|\bar{C}) = 0 \tag{5.13}$$

En otras palabras, las diferencias entre la GBA y NBA no pueden ser todas del mismo signo. Como veremos mas adelante, esto juega un papel importante en la comprensión de bajo cuales circunstancias la NBA es buena.

Para la probabilidad posterior tenemos

$$\begin{aligned} \Delta_{P_{GB}}(C|\mathbf{X}) &= \frac{\prod_{\alpha=1}^{N^C_{\xi^{(i)}}} P(\xi^\alpha|C)P(C)}{(\prod_{\alpha=1}^{N^C_{\xi^{(i)}}} P(\xi^\alpha|C)P(C) + \prod_{\alpha=1}^{N^{\bar{C}}_{\xi^{(j)}}} P(\xi^\alpha|\bar{C})P(\bar{C}))} \\ &- \frac{\prod_{i=1}^{N} P(X_i|C)P(C)}{\prod_{i=1}^{N}(P(X_i|C)P(C) + P(X_i|\bar{C})P(\bar{C}))} \end{aligned} \tag{5.14}$$

y, finalmente, para la función de score,

$$\begin{aligned} \Delta_{S_{GB}}(C|\mathbf{X}) &= \sum_{\alpha=1}^{N^C_{\xi^{(i)}}} \ln P(\xi^\alpha|C) - \sum_{\alpha=1}^{N^{\bar{C}}_{\xi^{(j)}}} \ln P(\xi^\alpha|\bar{C}) - \sum_{i=1}^{N} \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \\ &= \sum_{\alpha=1}^{N^C_{\xi^{(i)}}} \ln \frac{P(\xi^\alpha|C)}{P_{NB}(\xi^\alpha|C)} - \sum_{\alpha=1}^{N^{\bar{C}}_{\xi^{(j)}}} \ln \frac{P(\xi^\alpha|\bar{C})}{P_{NB}(\xi^\alpha|\bar{C})} \end{aligned} \tag{5.15}$$

donde $P_{NB}(\xi^\alpha|C) = \prod_{i=1}^{m} P(\xi_i^\alpha|C)$, m es el número de atributos, ξ_i^α, en el esquema de atributos ξ^α. En el caso de factorizaciones idénticas para C y $\bar{C}$

$$\Delta_{S_{GB}}(C|\mathbf{X}) = \sum_{\alpha=1}^{N_\xi} \ln \frac{P(\xi^\alpha|C)}{P(\xi^\alpha|\bar{C})} - \sum_{i=1}^{N} \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \tag{5.16}$$

5.2.4. Medidas de Desempeño

En la determinación de las ventajas relativas de la GBA frente al NBA requerimos una o más medidas de desempeño para juzgarlas. Aquí, consideraremos varias. En este capítulo, como se subrayo, la idea es entender los errores generados por los sesgos intrínsecos derivados de las diferentes aproximaciones. Por lo tanto, para determinar el impacto de las correlaciones en la validez de la NBA, y la mejora de la GBA, se puede y podríamos argumentar, se debe considerar en primer lugar un escenario de muestra infinita, donde los efectos de muestras finitas pueden ser ignorados. Esto último puede entonces ser considerado como un elemento secundario. Por lo tanto, en este capítulo vamos a considerar diferentes distribuciones de probabilidad definidas en un entorno con un número pequeño de atributos. La ventaja es que entonces tenemos en mano las distribuciones de probabilidad exactas y, por lo tanto, podemos medir los errores tanto en la NBA como en GBA, en relación con la distribución exacta $P_e(C|\mathbf{X})$. Explícitamente, consideraremos distribuciones definidas por las funciones de probabilidad $P_e(\mathbf{X}|C)$ y $P_e(\mathbf{X}|\bar{C})$, donde $\mathbf{X}$ es un vector N-dimensional que representa N atributos binarios $X_i = 0,\ 1$. Estas funciones de probabilidad se especificarán a fin de modelar diferentes grados de correlación entre los distintos atributos y, por lo tanto, nos permite entender mejor como la NBA y GBA actúan en función de estas correlaciones.

Como tenemos en mano las distribuciones exactas, consideraremos directamente como medida de desempeño el error en la distribución de probabilidad posterior $\Delta_i(C|\mathbf{X}) = P_e(C|\mathbf{X}) - P_i(C|\mathbf{X})$, donde i denota la aproximación correspondiente - NBA o GBA. En segundo lugar vamos a considerar la precisión de la clasificación, teniendo en cuenta la sensibilidad $\mathcal{S}_i$ de cada clasificador, $S_i(\mathbf{X})$, la cual corresponde al número o fracción de combinaciones de atributos clasificados correctamente. Podemos, de hecho, aplicar la misma lógica más allá de clasificación estándar pura por considerar como dos clasificadores clasifican con respecto a un umbral de score dado. Por ejemplo podemos definir un umbral de score, por, digamos $S_i = S_i^*$ para definir una categoría $S(X_1X_2) > S^*$. Un ejemplo podría ser el decile más alto de valores de score. Esto es particularmente útil en contextos donde $P(C)$ es muy pequeña, de modo que cualquier clasificador tendería a asignar cualquier ejemplar a $\bar{C}$. Entonces, si $\mathcal{S}(X_1X_2) = signo(S_i(X_1X_2) - S_j^*) - signo(S_j(X_1X_2) - S_j^*) = 0$, entonces ambos clasificadores, $S_i(\mathbf{X})$ y $S_j(\mathbf{X})$, asignan el ejemplar a las misma categoría, por ejemplo, al top 10 % de los valores de score. Para nuestro conjunto de distribuciones de probabilidad exactas, este nos permite probar el rendimiento de nuestros clasificadores contra el clasificador exacto $S_e(\mathbf{X})$.

En tercer lugar, vamos a considerar el ranqueo relativo de todo el conjunto de combinaciones de atributos en términos de la función de score correspondiente $\{S_i(1), S_i(2), \ldots, S_i(2^n)\}$, donde $S_i(1) \geq S_i(2) \geq \ldots \geq S_i(2^n)$. Entonces consideraremos la distancia

$$D_{ij} = \left(\sum_{m=1}^{2^n} (r_i(m) - r_j(m))^2 \right)^{1/2} \tag{5.17}$$

donde $r_i(m)$ es el ranqueo relativo de la combinación de atributos m del clasificador S_i y de manera similar $r_j(m)$ para el clasificador $S_j(m)$. Si $i = e$, corresponde al clasificador exacto, entonces la ecuación (5.17) mide que tan "correcto" es el ranqueo del clasificador relativo con el exacto.

5.3. Lidiando con Correlaciones

Las diferencias entre la NBA y GBA en el cap. 5.2.3, dependen en lo particular de la factorización elegida y esta factorización esta compuesta de componentes - esquemas. En la determinación de las ventajas relativas de la NBA y GBA hay entonces 2 preguntas básicas. En primer lugar, cuales criterios se deben utilizar para determinar los atributos que deberían considerarse conjuntamente en lugar de forma independiente. Esto está relacionado con la cuestión de cómo la diferencia global entre la NBA y GBA esta compuesta de las diferencias entre los componentes. En segundo lugar, una vez que hemos determinado cuales atributos se combinan, debemos preguntarnos como la NBA debe ser modificada. Consideraremos primero como identificar conjuntos de atributos que deben ser puestos juntos, empezando con el ejemplo sencillo de solo 2 atributos.

5.3.1. Dos Atributos

Tomando 2 atributos, X_1 y X_2, tratándolos como independientes, potencialmente conduce a errores en $P(\mathbf{X}|C)$, $P(\mathbf{X}|\bar{C})$ y $P(\mathbf{X})$, y por lo tanto en $P(C|\mathbf{X})$ y $S(\mathbf{X})$. Denotando

estos errores como $\delta(X_1X_2|C)$, $\delta(X_1X_2|\bar{C})$ y $\delta(X_1X_2)$ tenemos

$$\begin{aligned}
\delta(X_1X_2|C) &= P(X_1X_2|C) - P_{NB}(X_1X_2|C) \\
&= P(X_1X_2|C) - P(X_1|C)P(X_2|C) \qquad (5.18) \\
\delta(X_1X_2|\bar{C}) &= P(X_1X_2|\bar{C}) - P_{NB}(X_1X_2|\bar{C}) \\
&= P(X_1X_2|\bar{C}) - P(X_1|\bar{C})P(X_2|\bar{C}) \qquad (5.19) \\
\delta(X_1X_2) &= P(X_1X_2) - P_{NB}(X_1X_2) \\
&= \delta(X_1X_2|C)P(C) + \delta(X_1X_2|\bar{C})P(\bar{C}) \qquad (5.20)
\end{aligned}$$

lo cual satisface

$$\sum_{X_1X_2} \delta(X_1X_2|C) = \sum_{X_1X_2} \delta(X_1X_2|\bar{C}) = \sum_{X_1X_2} \delta(X_1X_2) = 0 \qquad (5.21)$$

lo que significa que los errores en la funciones de probabilidad, no pueden ser del mismo signo para todas las combinaciones de valor del atributo X_1X_2. Una de estas, al menos, debe tener un signo diferente del resto. De hecho para el caso simple de atributos binarios tenemos $\delta(X_1X_2|C) = \delta(\bar{X}_1\bar{X}_2|C) = -\delta(X_1\bar{X}_2|C) = -\delta(X_1\bar{X}_2|C)$, donde $\bar{X}_i$ es el bit complemento de X_i, lo que implica que solo hay dos errores independientes, los cuales tienen la misma magnitud y signo opuesto. Tenga en cuenta que el error en $P(X_1X_2)$ es solo un promedio ponderado de los errores en las funciones de probabilidad, para la clase C y su complemento. También podemos normalizar cualquiera de estos términos de error $\delta \to \delta'$; por ejemplo, dividiendo por la NBA. Esto es, $\delta(X_1X_2|C) \to \delta'(X_1X_2|C) = \delta(X_1X_2|C)/P_{NB}(X_1X_2|C)$.

Estos errores implican un error correspondiente en la probabilidad posterior

$$\begin{aligned}
\delta(C|X_1X_2) &= \left(\frac{P(X_1X_2|C)P(C)}{P(X_1X_2)} - \frac{P_{NB}(X_1X_2|C)P(C)}{P_{NB}(X_1X_2)}\right) \qquad (5.22) \\
&= \frac{\delta(X_1X_2|C)P_{NB}(\bar{C}|X_1X_2)P(C) - \delta(X_1X_2|\bar{C})P_{NB}(C|X_1X_2)P(\bar{C})}{P(X_1X_2)}
\end{aligned}$$

Las cantidades $\delta(X_1X_2|C)$ y $\delta(X_1X_2|\bar{C})$ ofrecen una descripción competa de los errores de la NBA para el caso de 2 variables. Curiosamente, podemos ver como $\delta(C|X_1X_2)$ involucra la diferencia de los errores en las dos funciones de probabilidad, es posible tener errores grandes en este, sin que esto necesariamente conduzca a un error significativo en la probabilidad posterior.

Con respecto a la función de score (5.4), hay dos posibles fuentes de error en $P(X_1X_2|C)$ y en $P(X_1X_2|\bar{C})$. El score, teniendo en cuenta la correlación es

$$\begin{aligned}
S(X_1X_2) &= S^C(X_1X_2) - S^{\bar{C}}(X_1X_2) \\
&= \ln\frac{P(C)}{P(\bar{C})} + \ln P(X_1X_2|C) - \ln P(X_1X_2|\bar{C}) \\
&= \ln\frac{P(C)}{P(\bar{C})} + \ln\frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})} \qquad (5.23)
\end{aligned}$$

Retomando que en el cap. 5.2.2 indicamos que las factorizaciones óptimas de la funciones de probabilidad para C y $\bar{C}$ pueden ser distintas, es conveniente introducir medidas separadas para los errores en las funciones de score correspondiente

$$\begin{aligned}
\delta_s(X_1X_2|C) &= \ln\left(1 + \frac{\delta(X_1X_2|C)}{P_{NB}(X_1X_2|C)}\right) \qquad (5.24) \\
\delta_s(X_1X_2|\bar{C}) &= \ln\left(1 + \frac{\delta(X_1X_2|\bar{C})}{P_{NB}(X_1X_2|\bar{C})}\right) \qquad (5.25)
\end{aligned}$$

Una consecuencia de (5.21) es que $\delta_s(X_1X_2|C)$ no puede tener el mismo signo para todos los X_1X_2, dando de este modo la base por la cual, las desviaciones de la NBA pueden cancelarse entre las diferentes combinaciones de atributos, cuando consideramos el caso de más de dos atributos.

Comparando con la ecuación (5.4) la diferencia entre la GBA y NBA que resulte de correlaciones entre X_1 y X_2 en C o $\bar{C}$ esta dada por

$$\begin{aligned}\delta_s(C|X_1X_2) &= \ln\left(\frac{1+\frac{\delta(X_1X_2|C)}{P_{NB}(X_1X_2|C)}}{1+\frac{\delta(X_1X_2|\bar{C})}{P_{NB}(X_1X_2|\bar{C})}}\right) \\ &= S(X_1X_2) - S_{NB}(X_1X_2) = S(X_1X_2) - S(X_1) - S(X_2)\end{aligned} \tag{5.26}$$

Así vemos que el error en el score es una medida de la no-linealidad de la contribución de las variables al score. Una vez más, podemos ver que el signo de los errores en las probabilidades, es crucial para determinar cuando es más probable que los errores en clasificación ocurran. En particular, podemos ver por qué la clasificación puede ser muy robusta, incluso en la presencia de correlaciones sustanciales. Si la correlación es similar entre C y $\bar{C}$, entonces el error en el score global, y por lo tanto, en el criterio de clasificación, puede ser relativamente pequeño. Por ejemplo, si los errores para las probabilidades de C y $\bar{C}$ son $X\,\%$ de sus respectivos valores NB, para cualquier X, entonces el error resultante en la función de score es cero y ambos clasificadores tendrán el mismo rendimiento independientemente de lo fuerte que las correlaciones son.

5.3.2. Caso General: Más de Dos Atributos

Para el caso de dos atributos hay solo una factorización posible. Para mas de dos atributos sin embargo, hay dos problemas relacionados para confrontar: ¿cuantas combinaciones de atributos aparecen en una factorización dada y ¿cuales atributos aparecerán en una combinación de atributo dada?. Dicho de otra forma, usando nuestra notación esquema: ¿como muchos esquemas diferentes aparecen en una factorización dada y después, ¿cuales atributos aparecen en un esquema dado?.

En cuanto al error en la NBA para una combinación de atributo dada - esquema - el análisis de dos atributos generaliza tranquila y fácilmente. Usando nuestra notación de esquema, errores (5.18), (5.19) y (5.20) tienen generalizaciones simples

$$\begin{aligned}\delta(\xi|C) &= P(\xi|C) - P_{NB}(\xi|C) \\ &= P(\xi|C) - \prod_{i=1}^{m} P(\xi_i|C) \qquad (5.27) \\ \delta(\xi|\bar{C}) &= P(\xi|\bar{C}) - P_{NB}(\xi|\bar{C}) \\ &= P(\xi|\bar{C}) - \prod_{i=1}^{m} P(\xi_i|\bar{C}) \qquad (5.28) \\ \delta(\xi) &= P(\xi) - P_{NB}(\xi) \\ &= \delta(\xi|C)P(C) + \delta(\xi|\bar{C})P(\bar{C}) \qquad (5.29)\end{aligned}$$

donde m es el número de atributos en el esquema ξ. En analogía con la ecuación (5.21) tenemos

$$\sum_{\xi}\delta(\xi|C) = \sum_{\xi}\delta(\xi|\bar{C}) = \sum_{\xi}\delta(\xi) = 0 \tag{5.30}$$

Para el error en la probabilidad posterior la generalización es

$$\begin{aligned}\delta(C|\xi) &= \left(\frac{P(\xi|C)P(C)}{P(\xi)} - \frac{P_{NB}(\xi|C)P(C)}{P_{NB}(\xi)}\right) \\ &= \frac{(\delta(\xi|C)P_{NB}(\bar{C}|\xi)P(C) - \delta(\xi|\bar{C})P_{NB}(C|\xi)P(\bar{C}))}{P(\xi)}\end{aligned} \tag{5.31}$$

Por último, para la función de score, a diferencia del caso de dos atributos, donde las factorizaciones de $P(X_1X_2|C)$ y $P(X_1X_2|\bar{C})$ son, por definición la misma, aquí es apropiado considerar por separado los errores en las contribuciones al score de las funciones de probabilidad para C y $\bar{C}$. A partir de la ecuación (5.9) podemos entonces definir

$$\delta_s(\xi|C) = \ln\left(1 + \frac{\delta(\xi|C)}{P_{NB}(\xi|C)}\right) \tag{5.32}$$

$$\delta_s(\xi|\bar{C}) = \ln\left(1 + \frac{\delta(\xi|\bar{C})}{P_{NB}(\xi|\bar{C})}\right) \tag{5.33}$$

Si asumimos que el mismo esquema aparece tanto en la factorización de $P(\mathbf{X}|C)$ y $P(\mathbf{X}|\bar{C})$ entonces el error es

$$\delta_s(C|\xi) = \delta_s(\xi|C) - \delta_s(\xi|\bar{C}) = S(\xi) - \sum_{i=1}^{m} S(\xi_i) \tag{5.34}$$

Como en el caso de dos atributos, las restricciones (5.30) implican que no todos los errores $\delta_s(\xi|C)$ y $\delta_s(\xi|\bar{C})$ pueden ser del mismo signo, lo que conduce a la posibilidad de cancelación de errores entre distintos esquemas.

5.3.3. Caso General: Más de Dos Esquemas

Las funciones de error anteriores, son útiles para determinar el impacto de las dependencias en un subconjunto $\xi \subset \mathbf{X}$, de valores de atributos. Sin embargo, cuando hay múltiples conjuntos, no esta claro cómo los errores en diferentes subconjuntos se pueden combinar para influir la diferencia total entre NBA y GBA como se ilustra en las ecuaciones (5.11), (5.12), (5.14) y (5.15). Como era de esperarse, la diferencia entre las funciones de probabilidad y las probabilidades posteriores para el GBA y NBA no parecen ser funciones simples de los errores $\delta(\xi|C)$. Sin embargo, para la función de score podemos escribir

$$\begin{aligned}\Delta_{S_{GB}}(C|\mathbf{X}) &= \sum_{\alpha=1}^{N_\xi^C} \ln P(\xi^\alpha|C) - \sum_{\alpha=1}^{N_\xi^{\bar{C}}} \ln P(\xi^\alpha|\bar{C}) - \sum_{i=1}^{N} \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \\ &= \sum_{\alpha=1}^{N_\xi^C} \ln \frac{P(\xi^\alpha|C)}{P_{NB}(\xi^\alpha|C)} - \sum_{\alpha=1}^{N_\xi^{\bar{C}}} \ln \frac{P(\xi^\alpha|\bar{C})}{P_{NB}(\xi^\alpha|\bar{C})} \\ &= \sum_{\alpha=1}^{N_\xi^C} \ln\left(1 + \frac{\delta(\xi^\alpha|C)}{P_{NB}(\xi^\alpha|C)}\right) - \sum_{\alpha=1}^{N_\xi^{\bar{C}}} \ln\left(1 + \frac{\delta(\xi^\alpha|\bar{C})}{P_{NB}(\xi^\alpha|\bar{C})}\right) \\ &= \sum_{\alpha=1}^{N_\xi^C} \delta_s(\xi^\alpha|C) - \sum_{\alpha=1}^{N_\xi^{\bar{C}}} \delta_s(\xi^\alpha|\bar{C})\end{aligned} \tag{5.35}$$

En (5.35) podemos ahora empezar a ver cómo las restricciones (5.30), pueden desempeñar un papel en cancelación de error relativo a la NBA. En el error asociado con C, como para cada esquema, ξ^α, existe al menos una combinación característica particular, $\xi^\alpha_{i_1 i_2 \ldots i_m}$, con

un error $\delta_s(\xi^\alpha|C)$, que tiene un signo diferente a los demas, entonces un cierto grado de error de cancelación es inevitable entre esquemas diferentes, cuando se consideran diferentes combinaciones de atributos para esos esquemas.

En el caso donde las factorizaciones son las mismas, la ecuación (5.35) se reduce aun mas,

$$\Delta_{S_{GB}}(C|\mathbf{X}) = \sum_{\alpha=1}^{N_\xi} \delta_s(C|\xi^\alpha) \tag{5.36}$$

Las ecuaciones (5.35) y (5.15), tienen una importante enseñanza. Al igual que para un solo esquema se pueden tener cancelaciones en los errores entre C y $\bar{C}$, por ejemplo, cancelaciones intra-esquemas, así que, cuando consideramos múltiples esquemas, pueden haber cancelaciones no solo entre C y $\bar{C}$, si no también entre los diferentes esquemas, es decir, cancelaciones inter-esquemas. Lo que es más, los errores de probabilidad para los componentes de las factorizaciones de C y $\bar{C}$, son los que determinan el error global de la NBA.

5.4. ¿Que Diferencia Hace?

Para determinar el impacto de las correlaciones para la validez de la NBA, es importante entender esto desde el punto de vista de modelo sesgado. Como hemos argumentado, este deberia ser considerado, en primer lugar, en un escenario de muestra infinita, donde los efectos de muestras finitas pueden ser ignorados. Este último puede entonces ser considerado como un elemento secundario, aunque importante para controlar aplicaciones reales. En el apéndice A se introduce un conjunto de 12 diferentes pares de ejemplares de distribuciones de probabilidad con valores binarios de sus elementos. Las distribuciones se caracterizan únicamente especificando los valores de las probabilidades $P(X_1X_2|C)$ y $P(X_1X_2|\bar{C})$ para $X_i = 0, 1$ para cada distribución.

Estas distribuciones de probabilidad se han construido para mostrar las diferentes estructuras de correlación que pueden ocurrir. Por ejemplo, las distribuciones 1 a 3 todas usan el mismo conjunto de probabilidades para C y $\bar{C}$ y solo se diferencian en como las probabilidades de $\bar{C}$ son asignadas a las diferentes combinaciones de valores de los elementos. Las tres distribuciones presentan fuertes correlaciones entre los elementos. Como lo veremos, no obstante, estas correlaciones afectan de manera muy distinta la validez de la NBA, a pesar de las similitudes. La distribución 4 muestra solo correlaciones débiles entre los elementos de ambas probabilidades. Las distribuciones 5 a 7 muestran correlaciones moderadas pero con diferentes características. Por ejemplo, las distribuciones 6 y 7 tienen correlaciones en las probabilidades que son de igual magnitud, difieren solo en el signo entre C y $\bar{C}$. La distribución 8 muestra fuertes correlaciones en la probabilidad para C, pero correlaciones débiles para $\bar{C}$. La distribución 9 es la inversa de la distribución 7 con C y $\bar{C}$ intercambiadas. La distribución 10 es el análogo a distribución 9 como la distribución 2 es a la 1, es decir, las probabilidades de C son las mismas, pero las probabilidades de $\bar{C}$ se han permutado. Finalmente, las distribuciones 11 y 12 otras dos que exhiben fuertes correlaciones en todas las funciones de probabilidad, pero difieren en como las correlaciones se distribuyen entre los elementos. En todos estos casos la probabilidad de la clase fue tomada como $P(C) = 0.6$.

Tendremos en cuenta los errores a dos niveles: en primer lugar, en la estimación de la probabilidad posterior $P(C|X_1X_2)$; y en segundo lugar, como un clasificador, donde consideramos la diferencia entre el clasificador exacto $S_e(\mathbf{X}) = \ln(P(C|X_1X_2)/P(\bar{C}|X_1X_2)) +$

$\ln(P(C)/P(\bar{C}))$ y la NBC ecuación (5.4). Tendremos en cuenta el porcentaje de errores entre los clasificadores exactos y las probabilidades posteriores y su contra-parte NBA, así como la función de distancia (5.17). En la tabla A.1 vemos el resultado del cálculo de la probabilidad posterior exacta para las doce distribuciones de probabilidad, la cual especificamos utilizando los valores de las funciones de probabilidad $P(X_1X_2|C)$, la NBA, la probabilidad posterior y el porcentaje de diferencia comparada con la expresión exacta. Hay cuatro combinaciones de elementos $X_1 = 0,\ 1,\ X_2 = 0,\ 1$ para cada clase $C = 1$, $\bar{C} = 0$ y viceversa. También en esta tabla están los scores para cada clasificador y su NBA, y el porcentaje de diferencia entre ellos.

¿Qué podemos deducir de estos resultados?. Claramente el desempeño de la NBA es muy heterogéneo en las diferentes distribuciones de probabilidad, con errores absolutos promedio sobre una distribución dada de entre 10 % y 100 % relativos a la probabilidad posterior exacta, y en el caso de las diferencias en score, errores entre 10 % y 5000 %. En el caso de la estimación de las probabilidades posteriores de las distribuciones, donde el error NB es mayor son 11, 3, 1, 8 y 10 y menor en 4, 7, 12, 9 y 2.

¿Cuáles son los factores que distinguen el mejor/peor desemepeño?. En la tabla A.1 podemos ver también los errores en las funciones de probabilidad (5.18) y (5.19), y los errores en las funciones de score (5.24) and (5.25). Podemos primeramente observar el hecho de que los errores en las funciones de probabilidad, en si mismas, no son necesariamente buenos indicadores de error en la estimación de las probabilidades posteriores o en la clasificación. Esto es consistente con nuestras observaciones en el cap. 5.3.1. Lo más importante es el signo relativo de los errores para C y $\bar{C}$. Las distribuciones con los errores más grandes en la probabilidad posterior o score - 11, 3, 1, 8 y 10 - todos tienen signos diferentes entre $\delta(X_1X_2|C)$ y $\delta(X_1X_2|\bar{C})$, para cada valor de X_1 y X_2. Por otro lado, cuatro de las cinco distribuciones con mínimo error - 2, 7, 9 y 12 - no tienen ninguna diferencia de signo entre $\delta(X_1X_2|C)$ y $\delta(X_1X_2|\bar{C})$, para cada valor X_1 y X_2. La distribución de error mínimo, 4, tienen diferencias de signos, pero en este caso la magnitud de los errores es muy pequeña $\approx 10^{-2}$. Así, vemos claramente la asociación entre los signos contrarios o complementarios de los errores en las funciones de probabilidad para la clase y su complemento como un indicador de la validez relativa de la NBA. En la misma tabla podemos ver la misma historia a nivel de la función de score. Comparando las tablas 10 y 11 podemos observar que existe una correlación entre los errores en la probabilidad posterior, el score y en la precisión de la clasificación y distancia. Vamos a investigar esto mas a fondo en la siguiente sección.

5.5. Diagnóstico para Cuando la NBA es Valida

Al considerar la validez de la NBA tenemos que preguntarnos: ¿La NBA de que? y también, ¿Con respecto a que punto de referencia?. Para responder a la NBA de que - dos cantidades importantes para el cálculo son las probabilidades posteriores y los scores de los clasificadores. Para la evaluación comparativa, obviamente, no podemos usar como punto de referencia a la respuesta exacta en problemas del mundo real, por la tanto debemos usar como punto de referencia otro algoritmo. Aquí, la evaluación comparativa será con el GBA. Generalmente, la evaluación de un algoritmo, tal como la NBA o GBA, proviene de una métrica de desempeño, sobre un conjunto de problemas de prueba. En otras palabras, la elección del algoritmo se determina a posteriori, después de probar el algoritmo en un conjunto de problemas. Como se ha subrayado, nuestro objetivo es inferir a priori si un tipo de algoritmo GBA será mejor que el NBA. Hemos visto que las diferencias entre las dos aproximaciones se derivan de 2 factorizaciones diferentes de las funciones de probabilidad,

para la clase de interés y su complemento. También hemos visto que podemos inferir errores potenciales al nivel de los factores individuales - esquemas -, los cuales son combinaciones de los elementos. Esto proporciona una gran simplificación, lo cual significa que existe la posibilidad de inferir, cual aproximación será mejor para un problema dado antes de aplicar el algoritmo completo. En otras palabras, hay la posibilidad de tener diagnósticos para determinar cual algoritmo será potencialmente mejor antes de ejecutar los algoritmos.

5.5.1. Diagnósticos para Cuando Combinar Elementos

Entonces, ¿Cual es un buen diagnóstico para determinar la precisión relativa de las aproximaciones? En primer lugar, regresemos al caso de dos elementos. Hemos encontrado que errores significativos en las funciones de probabilidad son una condición necesaria, pero no suficiente para tener errores significativos en la estimación de las probabilidades posteriores o en el desempeño de la clasificación. Con esto en mente, proponemos (5.18) y (5.19), o las funciones asociadas (5.24) y (5.25), como medidas directas de cuando es necesario combinar los elementos X_1X_2 en el contexto de una probabilidad dada. Específicamente introducimos

$$\Delta_C(X_1X_2) = \delta(X_1X_2|C)/P_{NB}(X_1X_2|C) \tag{5.37}$$
$$\Delta_{\bar{C}}(X_1X_2) = \delta(X_1X_2|\bar{C})/P_{NB}(X_1X_2|\bar{C}) \tag{5.38}$$
$$\Delta_s^C(X_1X_2) = \delta_s(X_1X_2|C)/S_{NB}^C(X_1X_2) \tag{5.39}$$
$$\Delta_s^{\bar{C}}(X_1X_2) = \delta_s(X_1X_2|\bar{C})/S_{NB}^{\bar{C}}(X_1X_2) \tag{5.40}$$

o sus equivalentes desnormalizadas, como medidas de cuando potencialmente combinar variables en las probabilidades de C y $\bar{C}$ respectivamente. Podemos entonces preguntarnos hasta qué grado errores significativos en estas funciones conducen a errores significativos en las probabilidades posteriores $P(C|X_1X_2)$, o en el desempeño de clasificación, o en cualquier otra métrica de desempeño que elijamos. Examinamos esto para las 12 distribuciones de probabilidad introducidas en la sección 5.4. En las figuras 5.1 y 5.2 vemos gráficos del error en la probabilidad posterior exacta $P(C|X_1X_2)$, $(P(C|X_1X_2) - P_{NB}(C|X_1X_2))/P_{NB}(C|X_1X_2)$, como una función de $\Delta_C(X_1X_2)$ y $\Delta_{\bar{C}}(X_1X_2)$, para las 12 distribuciones del apéndice A.

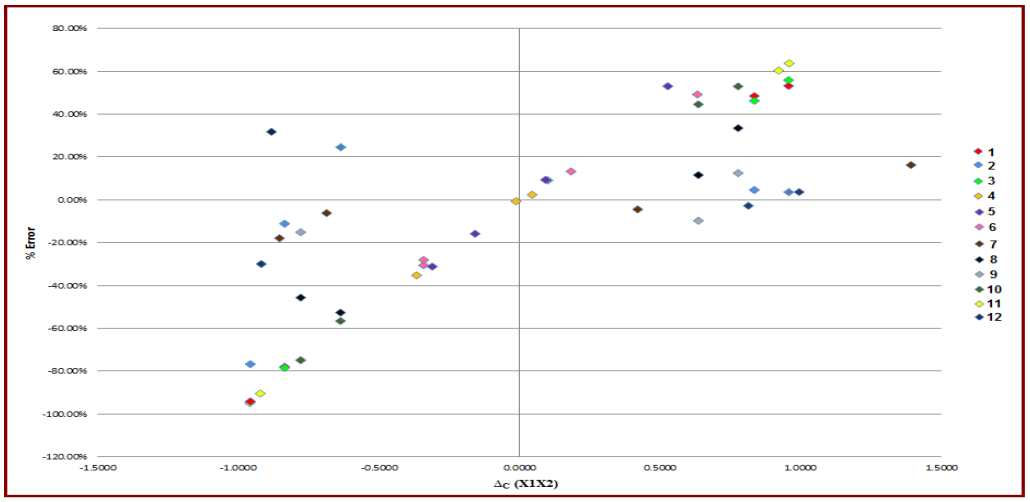


Figura 5.1: Gráfica de % de error en la NBA para la probabilidad posterior $P(C|X_1X_2)$, como función de $\Delta_C(X_1X_2)$, para las 12 distribuciones de probabilidad del apéndice A

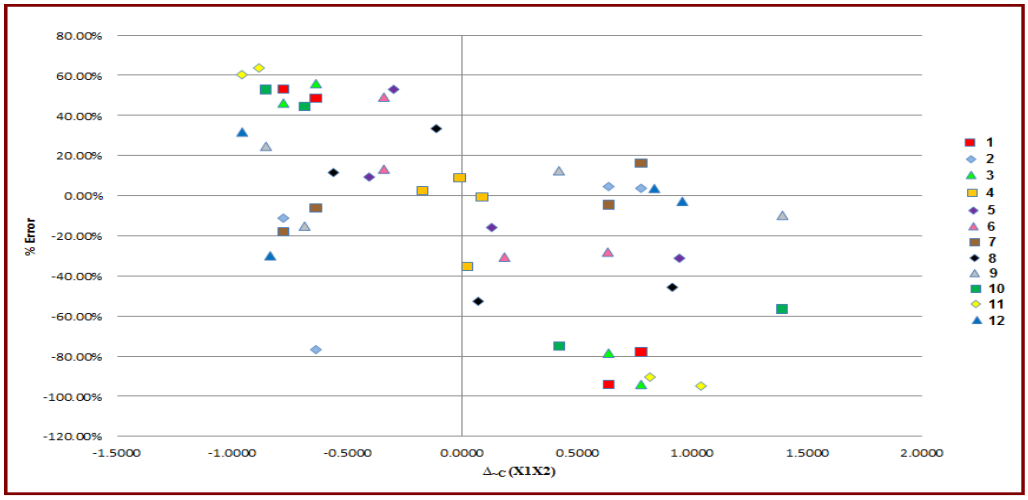


Figura 5.2: Gráfica de % de error en la NBA para la probabilidad posterior $P(C|X_1X_2)$, como función de $\Delta_{\bar{C}}(X_1X_2)$, para las 12 distribuciones de probabilidad del apéndice A

En los gráficos hemos diferenciado las distintas distribuciones, en donde, a partir del análisis de la seccíon 5.4, podemos distinguir las 5 distribuciones con mayor error, las 5 con menos error y las 2 intermedias - "neutrales". Vemos claramente el hecho de que los errores significativos en las funciones de probabilidad C o $\bar{C}$ por separado no es suficiente para predecir errores en la probabilidad posterior de la distribución correspondiente. De las cinco distribuciones con errores más pequeños - 2, 4, 7, 9 y 12 - sólo una, 4, tiene errores en las probabilidades, las cuales son pequeñas en magnitud. Las otras cuatro están asociadas con distribuciones donde los errores en las probabilidades individuales son grandes pero del mismo signo para C y $\bar{C}$, lo que conduce a una cancelación parcial entre los dos en sus contribuciones a las distribuciones de probabilidad posteriores. En distinción las distribuciones con el mayor error en la probabilidad posterior son tales que tienen grandes errores en la probabilidad, pero de signos opuesto entre C y $\bar{C}$, esto conduce a un refuerzo de los errores individuales.

Dado entonces que la diferencia entre la GBA y la NBA depende de los signos y magnitudes relativas de los errores en las funciones de probabilidad, proponemos como nuevos diagnósticos en el caso de 2 variables

$$\Delta(X_1X_2) = \left(\frac{\delta(X_1X_2|C)}{P_{NB}(X_1X_2|C)} - \frac{\delta(X_1X_2|\bar{C})}{P_{NB}(X_1X_2|\bar{C})} \right) \tag{5.41}$$

$$\Delta_s(X_1X_2) = \left(\frac{\delta_s(X_1X_2|C) - \delta_s(X_1X_2|\bar{C})}{S_{NB}(X_1X_2|C)} \right) \tag{5.42}$$

Tomaremos estos diagnósticos como una medida de la diferencia entre la GBA y la NBA, para una determinada función de combinación de dos variables. Por lo tanto, si $\Delta(X_1X_2)$ o $\Delta_s(X_1X_2)$ es grande, entonces se espera que la NBA tendrá relativamente un pobre rendimiento para predecir, digamos, la probabilidad posterior $P(C|X_1X_2)$.

En la figura 5.3 podemos ver el gráfico del error en la probabilidad posterior como función de $\Delta(X_1X_2)$. Como se puede ver, la correlación es impresionante - coeficiente de correlación 0.90, versus 0.77 y -0.59 para Δ_C y $\Delta_{\bar{C}}$ respectivamente - muestra que $\Delta(X_1X_2)$, en este entorno simplificado, es muy buen indicador de la validez de la NBA para el cálculo de probabilidades a posterior. Por ejemplo, si se requiere una exactitud de

20 % para las probabilidades posteriores de la NBA, entonces de la gráfica podemos ver que se requiere un Δ en el rango $[-0.08, 0.18]$. Por lo tanto, la NBA es suficiente cuando $-0.08 < \Delta < 0.18$.

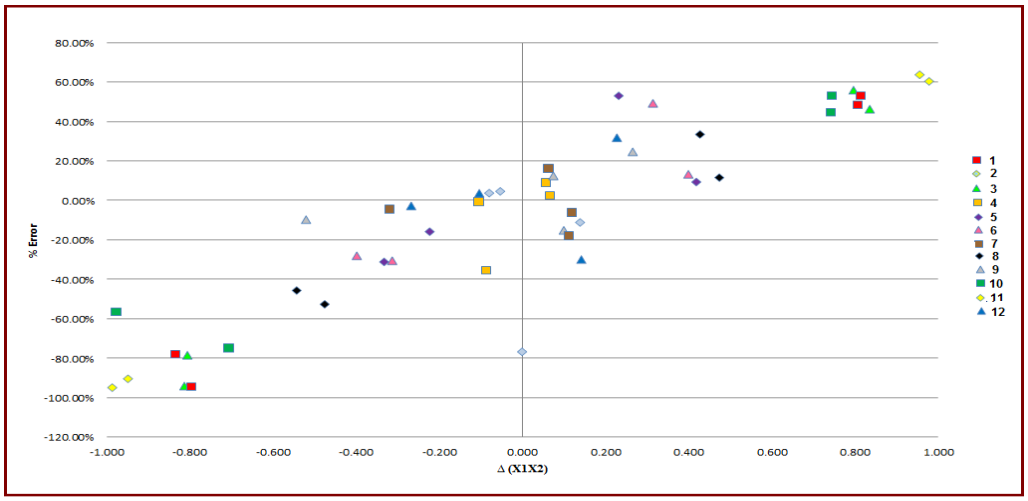


Figura 5.3: Gráfico de % error en la NBA de la probabilidad posterior $P(C|X_1X_2)$, como función de $\Delta(X_1X_2)$, para las 12 distribuciones de probabilidad del Apéndice A

En el caso de más de dos variables en un esquema, proponemos la ecuación (5.27) y (5.28) como diagnósticos para cuando el error en la función de probabilidad es suficiente, para justificar el uso de la aproximación GNB para ese esquema. Al igual que con el caso de dos variables, sin embargo, errores significativos en las funciones de probabilidad no son una condición suficiente para errores significativos en la distribución posterior o en la exactitud de la clasificación. De ahí que, uno puede estar tentado a tomar

$$\Delta(\xi) = \left(\frac{\delta(\xi|C)}{P_{NB}(\xi|C)} - \frac{\delta(\xi|\bar{C})}{P_{NB}(\xi|\bar{C})} \right) \tag{5.43}$$

$$\Delta_s(\xi) = \left(\frac{\delta_s(\xi|C) - \delta_s(\xi|\bar{C})}{S_{NB}(\xi|\bar{C})} \right) \tag{5.44}$$

como un diagnóstico natural para cuando los elementos se deben combinar en un esquema de más de dos variables. Sin embargo mas allá del caso de dos elementos, debemos hacer frente a la posibilidad de que la factorización óptima de las probabilidades para C y $\bar{C}$ pueden ser diferentes. En este caso, proponemos como diagnóstico

$$\Delta_C(\xi) = \frac{\delta(\xi|C)}{P_{NB}(\xi|C)} \tag{5.45}$$

$$\Delta_{\bar{C}}(\xi) = \frac{\delta(\xi|\bar{C})}{P_{NB}(\xi|\bar{C})} \tag{5.46}$$

$$\Delta_s^C(\xi) = \frac{\delta_s(\xi|C)}{S_{NB}^C(\xi)} \tag{5.47}$$

$$\Delta_s^{\bar{C}}(\xi) = \frac{\delta_s(\xi|\bar{C})}{S_{NB}^{\bar{C}}(\xi)} \tag{5.48}$$

o sus equivalentes desnormalizados. Así tenemos un dilema, podemos identificar medidas para cuando las funciones de probabilidad, o contribuciones al score, para un esquema determinado no deben ser factorizadas, pero esto no es suficiente para garantizar errores significativos en las probabilidades posteriores, o el desempeño en la clasificación. Por

otro lado, podemos identificar medidas, ecuaciones (5.41), (5.42), (5.43) y (5.44), que hablan directamente de errores, pero no necesariamente esta asociado con una factorización simétrica, no necesariamente óptimas, de las probabilidades $P(\mathbf{X}|C)$ y $P(\mathbf{X}|C)$.

Cuando las factorizaciones de probabilidad no son simétricas, la discución anterior indica que la validez de la NBA es una cuestión "global" en lugar de una "local"; es decir, para una factorización no-simétrica, la validez deberá ser declarada después de calcular la contribución y los errores correspondientes de todos los factores juntos. Por otro lado, para una factorización simétrica, la evaluación de los errores se puede partir en un conjunto de componentes separados, y así es como la generalización de la NBA se presento en la introducción. Consideraremos estos dos casos en mayor profundidad en la próxima sección, comenzando con el caso simétrico.

Primero, sin embargo, en lugar de un análisis para cada elemento individual de la combinación de variables, como en problemas del mundo real puede haber muchas combinaciones de valores de los elementos. Probablemente estamos mas interesados en la validez de la NBA cuando promediamos sobre ellos , y como, $\Delta(X_1X_2)$ puede cambiar de signo de un conjunto de valores de elementos a otro, como una medida de la validez de la NBA sobre el conjunto entero de posibles valores de elementos podemos tomar

$$D_2 = \sum_{i,j} |\Delta(X_iX_j)| \qquad \text{or} \qquad D_{2s} = \sum_{i,j} |\Delta_s(X_iX_j)| \tag{5.49}$$

En la figura 5.4 vemos un gráfico del error absoluto promedio en la probabilidad posterior contra el valor promedio D2, sobre las combinaciones de los valores de elementos 00, 01, 10, 11, y el valor promedio de $\delta(X_1X_2|C)/P_{NB}(X_1X_2|C)$ sobre estas combinaciones, para cada una de las 12 distribuciones de probabilidad mostradas en el apéndice A. Vemos claramente el alto grado de correlación para Δ, con el coeficiente de correlación 0.97 (el coeficiente de correlación correspondiente usando δ es 0.43), esto demuestra una vez mas, al menos a este nivel simple, su valor como un predictor de cuando la NBA no funciona. Se ve que los grandes errores en la función de probabilidad para la clase de interés no se traducen necesariamente en grandes errores en la probabilidad posterior. Esto también indica la importancia de considerar $\bar{C}$ en entender el efecto de la NBA en el cálculo de la probabilidad posterior.

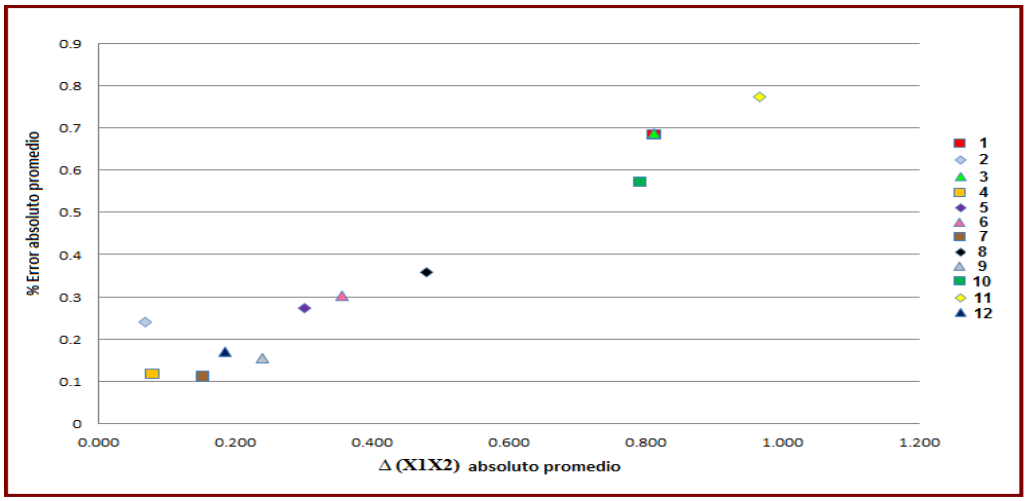


Figura 5.4: Gráfica del error absoluto promedio en $P(C|X_1X_2)$, como una función del valor promedio de D_2, para las 12 distribuciones de probabilidad del Apéndice A.

Además del impacto de la NBA en el cálculo de la probabilidad posterior, esta la pregunta de como esto impacta la exactitud de la clasificación. Podríamos, de hecho, utilizar la ecuación (5.41) como un diagnóstico. Sin embargo, también podemos utilizar (5.42), la cual muestra un alto grado de correlación con (5.41). En el caso de sólo dos elementos, si tomamos el score exacto como nuestro objetivo, entonces la ecuación (5.42) es algo tautológica. Sin embargo si consideramos la exactitud en la clasificación como nuestro objetivo, podemos graficar la sensibilidad de la NBA contra la clasificación exacta como función de (5.42) como se ve en la figura 5.5.

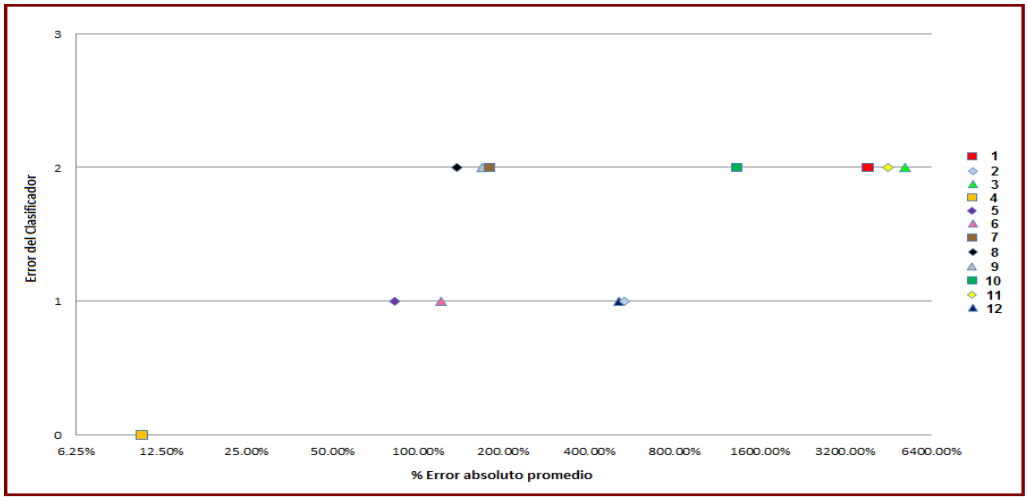


Figura 5.5: Gráfica del error absoluto promedio en $S_{NB}(X_1X_2)$, como una función del número promedio de errores de clasificación, para las 12 distribuciones del Apéndice A.

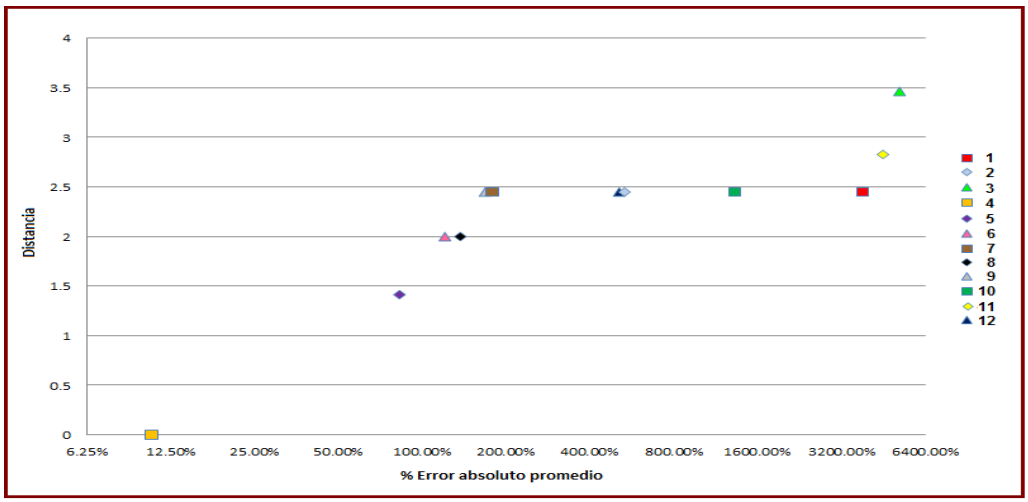


Figura 5.6: Gráfica del error absoluto promedio en $S_{NB}(X_1X_2)$, como una función del valor promedio de la distancia D, para las 12 distribuciones del Apéndice A.

Aunque no existe una clara tendencia también hay un alto grado de dispersión. La razón de esto es que una clasificación errónea no depende de la diferencia en magnitud entre $S(X_1X_2)$ y $S_{NB}(X_1X_2)$ para un X_1X_2, sino más bien, sólo si hay una diferencia en signo entre ellos, para tener en cuenta esto, podemos introducir el diagnóstico

$$\mathcal{S}(X_1X_2) = signo(S(X_1X_2)) - signo(S_{NB}(X_1X_2)) \tag{5.50}$$

Si $\mathcal{S}(X_1X_2) = 0$, entonces el GBC y NBC colocaran el ejemplar X_1X_2 en la misma clase. Si $\mathcal{S}(X_1X_2) \neq 0$ por otro lado el GBC Y NBC colocaran el ejemplar en clases diferentes. De esto podemos inferir claramente la robustez del NBC, dado que es sólo cuando el clasificador exacto y el NBC están en desacuerdo, cuando puede haber diferencias entre ellos. Esto ocurrirá principalmente cerca del umbral de score, S^*, genéricamente $S^* = 0$, lo cual marca el limite entre una clase y otra. En otras palabras, S y S_{NB} pueden ser sustancialmente diferentes y aún así estar de acuerdo en la clase asignada. Esta es la logica del argumento de [49] sobre la solidez del NBC.

El mismo pensamiento se puede aplicar más allá de la pura clasificación estándar, considerando como los dos clasificadores hacen su predicción (clasifican) con respecto a cualquier umbral de score dado. Por ejemplo, podemos definir un umbral de score, por digamos, $S_{NB} = S^*_{NB}$ para definir una categoría $S(X_1X_2) > S^*$. Un ejemplo sería el "top" decile de los valores de score más altos. Esto es particularmente útil en contextos donde $P(C)$ es muy pequeña, por lo que cualquier clasificador tendería a asignar cualquier ejemplar a $\bar{C}$. Entonces si $\mathcal{S}(X_1X_2) = signo(S(X_1X_2) - S^*_{NB}) - signo(S_{NB}(X_1X_2) - S^*_{NB}) = 0$ ambos clasificadores asignaran el ejemplar a la misma categoría, esto es, al "top" 10 % de los valores de score. También podemos ver, en la figura 5.6, la relación entre el desempeño del clasificador en términos de la métrica mas sensible de la distancia (5.17). Podemos ver que, efectivamente, esta métrica dependen más sensitivamente del error con las distribuciones de menor error, estando asociado esto con un mucho mejor promedio de desempeño que las distribuciones de mayor error.

Un diagnóstico complementario es la relación de la varianza en el score exacto y la varianza en el score NB , el valor promedio de esta relación es 429 para las 5 distribuciones de mayor error y 29 para las de menor error. Esto refleja el hecho de que hay mucho más dispersión en las probabilidades posteriores exactas que en sus aproximaciones NB y que esta dispersión se asocia particularmente con correlación entre los elementos, en comparación con las contribuciones de los elementos individuales en si mismos. Para la distribución 4, la NBA funciona muy bien, y podemos claramente ver porque. La varianza en los scores NB es 5.94, mucho mas grande que otras distribuciones, y muy similar a la varianza del exacto. Esta varianza es la materia prima sobre la cual la validez de la NBA descansa. Cuanto mayor sea esta varianza, y mas similar es a la varianza del exacto, menos probable es que los errores del muestreo den lugar a un "ranqueo" erróneo del NBC en relación con el exacto.

5.6. Dependencia de la Factorización en el Desempeño de la GBA versus la NBA

Nos concentraremos, a partir de este punto, sobre el efecto de las aproximaciones GNB y NB en términos de los correspondientes clasificadores $S_{GNB}(\mathbf{X})$ y $S_{NB}(\mathbf{X})$, en lugar del cálculo de las probabilidades posteriores. La razón de esto es doble, primeramente todos los trabajos sobre la NBA y GBA en su vasta mayoría son en términos de clasificación y, en segundo lugar, el análisis de los errores es mucho más simple debido a su naturaleza puramente aditiva en las funciones de score, como puede verse en las ecuaciones (5.15) y (5.35). Analizando la ecuación (5.35) mas profundamente, dos observaciones fundamentales son pertinentes: en primer lugar, como se ve claramente, el error en la función de score para un esquema ξ dado puede ser pequeño, a pesar de que los errores en los componentes de las funciones de probabilidad sean grandes, debido a una cancelación entre las diferencias $\delta_s(C|\xi)$ y $\delta_s(\bar{C}|\xi)$. En segundo lugar, la diferencia global en los clasificadores GBC y NBC pueden ser pequeñas debido a las cancelaciones entre las diferencias $\delta_s(C|\xi)$ y/o $\delta_s(\bar{C}|\xi)$

con los, $\delta_s(\bar{C}|\xi')$ y $\delta_s(\bar{C}|\xi')$, de los otros esquemas ξ'. Podemos pensar en estos diferentes tipos de cancelación como intra e inter esquemas.

Para desarrollar mayor intuición en este tema, podemos recurrir a la filosofía del caso de dos-variables, donde nos planteamos distribuciones especificas para las funciones de probabilidad. Ahora, sin embargo, estamos interesados en más de dos elementos. Vamos a considerar en esta sección distribuciones de tres-elementos $P(X_1X_2X_3|C)$, antes de pasar en secciones posteriores a distribuciones de cuatro, seis y ocho-elementos $P(X_1X_2X_3X_4|C)$, $P(X_1X_2X_3X_4X_5X_6|C)$ y $P(X_1X_2X_3X_4X_5X_6X_7X_8|C)$, todas con expresiones análogas para las probabilidades de $\bar{C}$. Por supuesto, se pueden considerar correlaciones en las probabilidades exactas de tal manera que no poseen ninguna factorización exacta en absoluto. Sin embargo, cualquier problema del mundo real por lo menos tendrá factorizaciones aproximadas. Lo que es más, en una implementación algorítmica de la GBA, donde un vasto espacio de posibles factorizaciones pueden ser buscadas, es natural concentrarse en las correlaciones entre pares de elementos ya que estos tienden a ser las combinaciones con el mayor tamaño de muestra y por lo tanto de mayor significado estadístico. Por lo tanto, vamos a utilizar las distribuciones de probabilidad del Apéndice A, para la creación de distribuciones con más de dos elementos mediante la concatenación de las distribuciones de dos-elementos que tenemos construidas, donde vamos a suponer que no existen dependencias entre las diferentes distribuciones de dos-elementos en si mismas. De esta manera no tendremos en cuenta las posibles correlaciones entre mas de dos elementos. La razón de esto es: primero, como se ha mencionado, en cualquier problema del mundo real las correlaciones binarias tenderán a ser lo más importante, sólo desde el punto de vista del significado estadístico, ya que el tamaño de la muestra para la coincidencia de cualquier par de elementos dado será siempre mayor que para la coincidencia de tres o mas elementos que contienen ese par de elementos, y en segundo lugar, el análisis siguiente se generaliza de manera muy sencilla a mas de dos elementos. Por último, todas nuestras observaciones cualitativas sobre la validez de la NBA son independientes del número de elementos correlacionados.

Vamos a proceder considerando diferentes escenarios posibles: primero, podemos considerar que las correlaciones en las distribuciones de probabilidad subyacentes de las probabilidades para C y $\bar{C}$ pueden ser simétricas, es decir, aparecen las mismas combinaciones de elementos en las dos probabilidades, o asimétricas, es decir, que los elementos correlacionados combinados en las probabilidades para C y $\bar{C}$ son distintos; segundo, en la aplicación de la GBA podemos elegir una factorización de las probabilidades que es simétrica, es decir, la misma para ambos, o asimétrica, es decir, distinta para ambos. Observaremos, naturalmente, que la factorización óptima, en función de nuestras métricas de desempeño, son las que mejor respetan las correlaciones en las distribuciones de probabilidad subyacente.

5.6.1. Correlaciones y Factorizaciones Simétricas - Tres Elementos

Vamos a comenzar con el caso donde las correlaciones en las probabilidades para C y $\bar{C}$ están asociadas con los mismos elementos y por otra parte, las factorizaciones de las probabilidades de C y $\bar{C}$ son simétricas en que los elementos combinados en la GBA son los mismos para ambas probabilidades. De hecho, tales factorizaciones simétricas están en el corazón de las generalizaciones de la NBA que han sido consideradas en la literatura. En lenguaje de esquemas, las factorizaciones simétricas son tales que las particiones de esquemas, $\xi^{(i)}$, de ambas probabilidades son iguales. Vamos a componer la distribución exacta de la probabilidad para ser un producto

$$P_e(\mathbf{X}|C) = P(X_1X_2X_3|C) = P(X_1X_2|C)P(X_3|C) \tag{5.51}$$

donde cada $P(X_1X_2|C)$ se puede elegir de una distribución independiente, por ejemplo, de nuestro conjunto de 12 distribuciones. En este caso, la probabilidad es producto de un esquema de orden dos y un esquema de orden uno. Análogamente, en este caso simétrico

$$P_e(\mathbf{X}|\bar{C}) = P(X_1X_2X_3|\bar{C}) = P(X_1X_2|\bar{C})P(X_3|\bar{C}) \tag{5.52}$$

con una función score/clasificador exacto

$$\begin{aligned} S_e(X_1X_2X_3) &= \ln \frac{P(X_1X_2X_3|C)}{P(X_1X_2X_3|\bar{C})} \\ &= \ln \frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})} + \ln \frac{P(X_3|C)}{P(X_3|\bar{C})} \end{aligned} \tag{5.53}$$

Por lo tanto, tomamos las correlaciones en las distribuciones de probabilidad subyacentes para ser iguales en ambas probabilidades.

Ahora queremos determinar los méritos relativos de la GBA y NBA en este problema. En primer lugar, a diferencia de la NBA, la cual es única, para la GBA hay muchas posibles factorizaciones distintas. Para el caso presente de 3 elementos binarios, denotamos la NBA por la partición de esquema $\xi^{(0)} = (\xi_1 = X_1, \xi_2 = X_2, \xi_3 = X_3)$. Las 3 posibles particiones de dos-esquemas son: $\xi^{(1)} = (\xi_1 = X_1X_2, \xi_2 = X_3)$, $\xi^{(2)} = (\xi_1 = X_1X_3, \xi_2 = X_2)$ y $\xi^{(3)} = (\xi_1 = X_2X_3, \xi_2 = X_1)$. El único esquema de orden 3 $\xi^{(4)} = (\xi_1 = X_1X_2X_3)$ corresponde a la distribución de probabilidad exacta en si misma. Como, por construcción, no existen dependencias entre los esquemas $\xi_1 = X_1X_2$ y $\xi_2 = X_3$, la GBA, si esta basada en los dos esquemas, $\xi_1 = X_1X_2$ y $\xi_2 = X_3$, debería ser exacta. En otras palabras, donde el superindice e sobre P^e_{GB} significa que esta particular factorización de la GBA es exacta ya que respeta la estructura de la correlación de la distribución de probabilidad exacta. Las factorizaciones de la GBA que vamos a considerar aquí son $(\xi^{(0)}, \xi^{(0)})$, $(\xi^{(1)}, \xi^{(1)})$, $(\xi^{(2)}, \xi^{(2)})$ y $(\xi^{(3)}, \xi^{(3)})$ las cual son todas simétricas, en las que los elementos combinados son los mismos en ambas probabilidades C y $\bar{C}$.

En la tabla 5.1 vemos el desempeño de estas diferentes factorizaciones en relación con el clasificador exacto y también la NBA, $(\xi^{(0)}, \xi^{(0)})$, para los casos en que las correlaciones en la distribución de probabilidad subyacentes son fuertes, s (usamos distribución de probabilidad 1 del Apéndice A) o débiles, w (donde utilizamos distribución 4 del Apéndice A). Como medidas de desempeño se considera la sensitividad de los diferentes clasificadores y la función de distancia, la ecuación (5.17). Consideraremos los casos ss, ww, sw y ws para la fuerza de las correlaciones en las probabilidades para C y $\bar{C}$ respectivamente. Por lo tanto, ss se refiere a una fuerte correlación en ambas probabilidades (distribución 1 para las probabilidades de C y $\bar{C}$), mientras ws corresponde a una débil correlación en la probabilidad de C (distribución 4), pero una fuerte correlación en la probabilidad de $\bar{C}$ (distribución 1) y viceversa para sw. Por último, vamos a considerar las distribuciones con correlaciones fuertes ss', las cuales corresponden a la distribución 2 para ambas probabilidades C and $\bar{C}$. La distinción importante entre las distribuciones 1 y 2 es que, aunque ambas están asociadas con correlaciones fuertes en las probabilidades para C y $\bar{C}$, la distribución 1 resulta en un reforzamiento de los errores entre las probabilidades para C y $\bar{C}$, mientras la distribución 2 esta asociada con una cancelación intra-esquema.

Lo que podemos observar es lo siguiente: para el caso en el que existen fuertes correlaciones en una o ambas funciones de probabilidad, la GBA es mejor o igual que la NBA en todos los casos, independientemente de la factorización. Para la factorización $\xi^{(1)}$ tanto para C y $\bar{C}$ es estrictamente mejor. Es comprensible, como en este caso la factorización

	SS		WW		WS		SW		S S'	
F	S	D	S	D	S	D	S	D	S	D
00	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
11	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
22	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
33	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898

Tabla 5.1: Medición de Desempeño para factorizaciones simétricas de la GBA en el problema de 3 elementos con correlaciones simétricas en las funciones de probabilidad.F es Factorización, S es Sensitividad y D es Distancia.

capta con precisión la estructura de la correlación exacta de la distribución de probabilidad exacta subyacente. Para las factorizaciones $\xi^{(2)}$ y $\xi^{(3)}$, el desempeño de la GBA es equivalente a de la NBA, por que no hay correlaciones entre los elementos en los pares X_1X_3 or X_2X_3. Por ejemplo, para $\xi^{(2)}$, el cual considera $P(\xi_1|C) = P(X_1X_3|C)$, tenemos $P(X_1X_3|C) = P(X_1|C)P(X_3|C)$, lo cual es equivalente a la NBA. Tenga en cuenta también que el desempeño de la NBA en el caso ss' es sustancialmente mejor que para la distribución ss, mostrando como la calidad de la NBA es mejor cuando hay un efecto de cancelación entre los grandes error de la probabilidades. También podemos concluir que no hay un costo relativo en el desempeño de la NBA por una elección inapropiada de la factorización simétrica 22 o 33, pero tampoco hay una mejora.

5.6.2. Correlaciones Simétricas, Factorizaciones Asimétricas - Tres Elementos

Hemos observado que no es un requisito de una generalización de la NBA que la factorización de las funciones de probabilidad de C y $\bar{C}$ sean las mismas. Por lo tanto, en esta sección consideraremos el caso en el cual la estructura de las correlaciones de las probabilidades para $P(\mathbf{X}|C)$ y $P(\mathbf{X}|\bar{C})$ son las mismas, pero las factorizaciones de estas probabilidades son distintas, es decir, las factorizaciones son asimétricas. Para tres elementos utilizamos de nuevo las distribuciones (5.51) y (5.52). A diferencia del caso anterior, consideramos factorizaciones donde distintos pares de elementos son combinados entre las probabilidades para C y $\bar{C}$. Por lo tanto, consideramos las particiones de esquemas: $\xi_C^{(i)}, \xi_{\bar{C}}^{(j)}$ con $i,\ j = 0, 1, 2, 3$ y $i \neq j$. Por ejemplo, la partición $\xi_C^{(0)}, \xi_{\bar{C}}^{(1)}$ corresponde a la NBA para la probabilidad de C y la GBA para la probabilidad de $\bar{C}$ con los elementos X_1X_2 combinados, mientras la partición $\xi_C^{(2)}, \xi_{\bar{C}}^{(1)}$ corresponde a la GBA donde la probabilidad de C tiene los elementos X_1X_3 combinados y la probabilidad de $\bar{C}$ tiene los elementos X_1X_2 combinados. Dado que las correlaciones en las distribuciones de probabilidad de las funciones de probabilidad son simétricas, mientras las factorizaciones de la GBA son consideradas asimétricas, solo podremos obtener la factorización óptima de ya sea la probabilidad de C o $\bar{C}$ pero no de ambas al mismo tiempo.

Al igual que en la sección anterior examinamos el desempeño relativo de estas diferentes factorizaciones con respecto al clasificador exacto y también a la NBA para los casos donde las correlaciones en las distribuciones de probabilidad subyacentes son fuertes o débiles, utilizando otra vez como medidas de desempeño la exactitud de la clasificación de los diferentes clasificadores, la función de distancia, ecuación (5.17), y también la diferencia relativa del score con la NBA y el clasificador exacto. Consideramos de nuevo los casos ss, ww, sw, ws y ss' de la sección anterior para la fuerza de las correlaciones en las

probabilidades para C y $\bar{C}$ respectivamente. Una vez más, utilizamos la distribución 1 para s en ss, ws y sw y distribución 4 para w en ww, ws y sw. Para ss' usamos distribución 2.

En la tabla 5.2 vemos el resultado. En 114 de los 120 casos posibles, la GBA asimétrica, donde solo una probabilidad combina elementos, es igual o superior a la NBA con respecto a las dos métricas - distancia y exactitud en la clasificación. En el caso donde la factorización captura la correlación subyacente en una de las probabilidades, es decir, $i = 1$ o $j = 1$ en $\xi_C^{(i)}, \xi_{\bar{C}}^{(j)}$, vemos lo siguiente: la GBA es bastante mejor que la NBA en todos los casos para la distribución ss para ambas métricas; para la distribución ws es bastante mejor en los casos 01, 21 y 31 donde la GBA captura las correlaciones fuertes en la probabilidad para $\bar{C}$; para la distribución sw es bastante mejor en los casos 10, 12 y 13, donde la GBA cuenta para las correlaciones fuertes en las probabilidades para C. Finalmente, para el caso ss', observamos que la GBA es bastante mejor para los casos 10, 12 y 13 pero bastante peor para 01, 21 y 31. También debemos señalar que para la distribución sw la GBA es ligeramente peor que la NBA para las factorizaciones 01, 21 y 31 correspondientes a una factorización de las probabilidades fuertemente correlacionadas para C pero una combinación de los elementos X_1X_2 para las probabilidades débilmente correlacionadas para $\bar{C}$.

	SS		WW		WS		SW		S S'	
F	S	D	S	D	S	D	S	D	S	D
00	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
01	0.750	2.000	1.000	0.000	1.000	2.828	0.625	7.348	0.375	10.29
02	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
03	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
10	1.000	1.414	1.000	0.000	0.875	4.242	1.000	0.000	0.875	2.449
12	1.000	1.414	1.000	0.000	0.875	4.242	1.000	0.000	0.875	2.449
13	1.000	1.414	1.000	0.000	0.875	4.242	1.000	0.000	0.875	2.449
20	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
21	0.750	2.000	1.000	0.000	1.000	2.828	0.625	7.348	0.375	10.29
23	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
30	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
31	0.750	2.000	1.000	1.414	1.000	2.828	0.625	7.348	0.375	10.29
32	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898

Tabla 5.2: Medidas de desempeño para factorizaciones asimétricas de la GBA en problema de tres-elementos con correlaciones simétricas en las funciones de probabilidad, donde F es Factorización, S es Sensitividad y D es Distancia y añadimos la NBA para facilitar la comparación.

En resumen: para las situaciones en las que se mezclan las correlaciones ws y sw, las factorizaciones donde la GBA supera a la NBA son aquellas en los que los elementos combinados son precisamente aquellos en los que la probabilidad muestra fuertes correlaciones. Podemos concluir entonces de este ejemplo de tres elementos, que una factorización de la GBA combinando elementos en una probabilidad, la cual presenta fuertes correlaciones en los elementos, lleva a un mejor desempeño que la NBA. Incluso en el caso donde las correlaciones son débiles - la distribución ww - una factorización GBA que captura la estructura correcta de la correlación de una de las probabilidades, conduce a mejores resultados en términos de la métrica más sensitiva de la distancia. Por otro lado, para el caso fuertemen-

te correlacionado ss', donde hay cancelaciones entre los errores en las probabilidades para C y $\bar{C}$, combinando elementos para una sola de las probabilidades conduce a resultados mixtos.

5.6.3. Correlaciones Asimétricas, Todas las Factorizaciones - Tres Elementos

Por último, consideramos el caso donde las correlaciones en las dos probabilidades son asimétricas con

$$P_e(\mathbf{X}|C) = P(X_1X_2X_3|C) = P(X_1X_2|C)P(X_3|C) \tag{5.54}$$

y

$$P_e(\mathbf{X}|\bar{C}) = P(X_1X_2X_3|\bar{C}) = P(X_1X_3|\bar{C})P(X_2|\bar{C}) \tag{5.55}$$

con una función score/clasificador exacta

$$\begin{aligned} S_e(X_1X_2X_3) &= \ln \frac{P(X_1X_2X_3|C)}{P(X_1X_2X_3|\bar{C})} \\ &= \ln P(X_1X_2|C) + \ln P(X_3|C) - \ln P(X_1X_3|\bar{C}) - \ln P(X_2|\bar{C}) \end{aligned} \tag{5.56}$$

Entonces, ¿Como la GBA y NBA se comportan en este caso?. Una vez mas, hay tres posibles factorizaciones: $\xi^{(1)} = (\xi_1 = X_1X_2, \xi_2 = X_3)$, $\xi^{(2)} = (\xi_1 = X_1X_3, \xi_2 = X_2)$ y $\xi^{(3)} = (\xi_1 = X_2X_3, \xi_2 = X_1)$. Sin embargo, distinto al caso simétrico, aquí la factorización óptima de la GBA es diferente para las dos probabilidades. Para $P(X_1X_2X_3|C)$, la factorización $\xi^{(1)}$ será óptima, mientras para $P(X_1X_2X_3|\bar{C})$ la factorización $\xi^{(2)}$. Como en las dos secciones previas, consideramos los casos ss, ww, sw, ws y ss' para la fuerza en las correlaciones en las probabilidades para C and $\bar{C}$ respectivamente. Otra vez, usamos la distribución 1 para s en ss, ws y sw y distribución 4 para w en ww, ws y sw. Para ss' utilizamos la distribución 2. La diferencia ahora es que s en ss está asociada con la distribución 1 para $P(X_1X_2|C)$ y $P(X_1X_3|\bar{C})$, $P(X_1X_2|C)$ para s en sw y $P(X_1X_3|\bar{C})$ para s en ws. Del mismo modo para w.

En la tabla 5.3 vemos el desempeño relativo de las diferentes factorizaciones con respecto al clasificador exacto y la NBA (factorization 00) para el caso ss, ww, sw, ws y ss'.

La primera observación inmediata es que la factorización 12, correspondiente a $\xi^{(1)}$ para $P(X_1X_2X_3|C)$ y $\xi^{(2)}$ para $P(X_1X_2X_3|\bar{C})$, tiene un desempeño óptimo, con un 100 % de exactitud en la clasificación y "ranqueo" perfecto. En general, vemos que la GBA para la mayoría de las factorizaciones es mejor o igual a la NBA con respecto a ambas métricas. Para la distribución ss, la GBA es bastante mejor o igual a la NBA para todas las factorizaciones y ambas medidas de desempeño. Además, es bastante mejor en estos casos, $1*$ y $*2$, donde uno de los pares de elementos combinado captura la correlación subyacente en las distribuciones de probabilidad; $1*$ captura la fuerte correlación en $P(X_1X_2|C)$ y $*2$ la fuerte correlación subyacente en $P(X_1X_3|C)$. Curiosamente, para el caso ww, la NBA es óptima con respecto a ambas: sensitividad y distancia. Esto a diferencia del caso de correlaciones simétricas. En las distribuciones mixtas, ws and sw, vemos el mismo patrón como para el caso de correlaciones simétricas, es decir, la GBA es bastante mejor para sw en los casos 10, 11, 12 y 13, donde la GBA captura las correlaciones fuertes la probabilidad para C; para las distribuciones ws es bastante mejor en los casos 02, 12, 22 y 32, donde la GBA contabiliza las correlaciones fuertes en la probabilidad para $\bar{C}$. Para la distribución ss', curiosamente, en ningún caso es la GBA peor que la NBA en términos de la métrica distancia. Para la métrica clasificación es peor para 02, 22 y 32. Estos son precisamente los casos donde los elementos combinan la probabilidad fuertemente correlacionada para $\bar{C}$, pero no para la probabilidad fuertemente correlacionada en C.

	SS		WW		WS		SW		S S'	
F	S	D	S	D	S	D	S	D	S	D
00	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
01	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
02	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
03	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
10	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.162
11	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.62
12	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
13	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.262
20	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
21	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
22	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
23	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
30	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
31	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
32	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
33	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348

Tabla 5.3: Medidas de desempeño para todas las factorizaciones del GBA, en el problema de tres-elementos con correlaciones asimétricas en las funciones de probabilidad, donde F es Factorización, S es Sensitividad y D es Distancia y agregamos la NBA para fácil comparación.

5.7. Análisis de Errores - Eligiendo la Factorización Correcta

En lo descrito anteriormente hemos mostrado, como el desempeño relativo de la GBA, cuando se compara a la NBA, fue sensible a la factorización usada para la GBA, relativo a la distribución de las correlaciones inherentes en las distribuciones de probabilidad subyacentes. Sin embargo, para determinar la eficacia relativa de la GBA tuvimos que considerar exhaustivamente cada factorización. Esto no es posible en problemas del mundo real con muchas variables, de ahí, la importancia de un diagnóstico a priori. Hemos propuesto anteriormente como diagnósticos de error, las estadísticas $\Delta_C(X_iX_j)$, $\Delta_{\bar{C}}(X_iX_j)$ y $\Delta(X_iX_j)$. Ahora vamos a considerar la distribución de los errores asociados con todas las combinaciones de dos-elementos usando estas estadísticas, para mostrar como estos indican cuando y por que la GBA debe utilizarse. Vamos a trabajar primero en el contexto del problema de tres-elementos considerando primero correlaciones simétricas y después asimétricas en las distribuciones de probabilidad subyacentes.

Consideremos primero correlaciones simétricas, donde no hay ninguna correlación entre los elementos X_1 y X_2 en ambas probabilidades C y $\bar{C}$. Un análisis de los errores para cada par de valores de elementos usando nuestras estadísticas $\Delta_C(X_iX_j)$, $\Delta_{\bar{C}}(X_iX_j)$ y $\Delta(X_iX_j)$ produce varias observaciones notables. En primer lugar, que $\Delta_C(X_iX_j) = \Delta_{\bar{C}}(X_iX_j) = \Delta(X_iX_j) = 0$ para $X_iX_j = X_1X_3$ o $X_iX_j = X_2X_3$, para los cinco tipos de distribución de correlaciones fuertes ss, ww, sw, ws y ss'. En otras palabras, nuestros diagnósticos pueden claramente identificar combinaciones de elementos donde no hay correlaciones, que pudieran ser bien aproximadas por la NBA. Los únicos valores distintos de cero de nuestros diagnósticos están asociados con la combinación de los elementos X_1X_2.

Allí, sus valores dependen del grado de correlación en las distribuciones de probabilidad subyacentes. Para la distribución ss, donde hay fuertes correlaciones en las probabilidades de C y $\bar{C}$, ambas $\Delta_C(X_1X_2)$ y $\Delta_{\bar{C}}(X_1X_2)$ son grandes, entre 64 %-96 % de las probabilidades correspondientes en la NBA, mientras el error en $\Delta(X_1X_2)$ es aproximadamente el doble debido al reforzamiento de los errores de C y $\bar{C}$. Para la distribución ww, por otro lado, cuando hay correlaciones débiles en las probabilidades de C y $\bar{C}$, ambos $\Delta_C(X_1X_2)$ y $\Delta_{\bar{C}}(X_1X_2)$ son relativamente pequeños, sólo alrededor del 1 %-36 % de las probabilidades correspondientes NBA.

Los errores en $\Delta(X_1X_2)$ son solo del orden de 10 %-39 % a pesar del hecho de que no hay ninguna cancelación entre los errores de probabilidad para C y $\bar{C}$. Para los casos, sw y ws los errores están dominados por las correspondientes distribuciones fuertemente correlacionadas. Así, para sw, donde la probabilidad de C muestra una fuerte correlación entre los elementos, el error de la probabilidad en C es 83 %-95 %, mientras que de $\bar{C}$ es solo 1 %-17 %. El error resultante $\Delta(X_1X_2)$, esta por lo tanto, dominado por el error de la probabilidad en C. Un resultado similar ocurre para la distribución ws, donde ahora el error en la probabilidad para $\bar{C}$ domina el error global. Para la distribución ss', aunque los errores en $\Delta_C(X_1X_2)$ y $\Delta_{\bar{C}}(X_1X_2)$ son grandes, 83 %-95 % en C y 64 %-78 % en $\bar{C}$, el error en $\Delta(X_1X_2)$ es relativamente pequeño, 6 %-54 %, debido a la cancelación de errores entre las dos probabilidades.

Además de considerar nuestro diagnóstico de errores para cada combinación de elementos, se puede considerar también el error absoluto promedio sobre todas las combinaciones de elementos por pares y para cada tipo de correlación. En la tabla 5.4 vemos el resultado para correlaciones simétricas. Como era de esperar, dado que las correlaciones en las distribuciones de probabilidad subyacentes están asociadas con el elemento par X_1X_2, el error promedio para un elemento par X_iX_j con $i \neq 1$ y $j \neq 2$ es cero. Podemos observar para la combinación ss que los errores en ambas probabilidades son grandes sin cancelación, mientras para la combinación ww ambos errores son pequeños. Para las distribuciones mixtas el error esta dominado por la probabilidad asociada con la correlación fuerte, mientras en el caso de ss' vemos el error promedio en Δ es mucho mas pequeño que el error en las probabilidades correspondientes.

Ahora vamos a considerar el análisis de errores en el caso donde las correlaciones en las distribuciones de probabilidad subyacentes para las funciones de probabilidad son asimétricas, con una correlación entre X_1X_2 para C y entre X_1X_3 para $\bar{C}$. Consideramos esto como un caso de estudio importante ya que la mayoría de las generalizaciones de la NBA en la literatura implícitamente asumen que la estructura de correlación en el problema subyacente es simétrica, así como las factorizaciones correspondientes de las funciones de probabilidad se eligen para ser simétricas. No vemos ninguna razón por la cual los problemas del mundo real deben exhibir estructuras de correlación que son intrínsecamente simétricas entre las clases. Una diferencia importante para el caso simétrico es que, como se esperaba, la Δ_{ij} en el caso ss' no presenta ninguna cancelación, como las correlaciones fuertes están asociadas a las probabilidades con distintos conjuntos de elementos.

5.7.1. Relacionando los Errores a la GBA - Tres Elementos

Vemos entonces que nuestros diagnósticos indican claramente cual combinación de elementos debería potencialmente ser utilizada junta en lugar de forma independiente, y por lo tanto cual factorización de la GBA es mas adecuada. También vemos que podría ser necesario combinar elementos en solo una de las probabilidades un lugar de ambas. Con el fin de determinar si los elementos X_iX_j deben combinarse juntos y, además, si

Correlación Combinación	Elementos Considerados	Promedio Δ_C	Promedio $\Delta_{\bar{C}}$	Promedio Δ
SS	12	0.372	0.440	0.810
SS	13	0.000	0.000	0.000
SS	23	0.000	0.000	0.000
WW	12	0.037	0.042	0.079
WW	13	0.000	0.000	0.000
WW	23	0.000	0.000	0.000
WS	12	0.029	0.420	0.390
WS	13	0.000	0.000	0.000
WS	23	0.000	0.000	0.000
SW	12	0.370	0.033	0.340
SW	13	0.000	0.000	0.000
SW	23	0.000	0.000	0.000
S S'	12	0.372	0.440	0.068
S S'	13	0.000	0.000	0.000
S S'	23	0.000	0.000	0.000

Tabla 5.4: Error absoluto promedio sobre todas las combinaciones de valores de elementos para cada tipo de correlación y factorización.

deberían utilizarse combinadas para ambas probabilidades C y $\bar{C}$ o solo en una o ninguna, implica el establecimiento de un umbral para los errores encima del cual los elementos deben combinarse. En este ajuste, el umbral natural de población infinita es cero ya que no hay errores de muestreo afectando la Δ, la cuales son entonces una medida pura del error debido al sesgo del modelo. Vamos a considerar el uso de Δ_C para determinar cuando para determinar cuando combinar elementos en la probabilidad para C, $\Delta_{\bar{C}}$ para combinar elementos en $\bar{C}$ y Δ para determinar si es probable ver cancelación o reforzamiento de errores. En esta configuración de tres-elementos consideramos el desempeño del clasificador como función de Δ, donde el desempeño es medido en términos de nuestra función de distancia y exactitud del clasificador. En la figuras 5.7 y 5.8 vemos graficas de error promedio versus distancia promedio y exactitud del clasificador respectivamente. Para cada distribución de correlación ww, ws, sw ss' y ss mostramos resultados para ambos: el promedio de error con signo normalizado y el error absoluto promedio. La diferencia entre los dos errores es una métrica del grado de cancelación en los errores entre C and $\bar{C}$.

Para ambas medidas de desempeño vemos una clara correlación entre el error promedio y el desempeño, como se esperaba, la distribución ww tiene el mejor desempeño y ss el peor. Para las distribuciones ws, sw y, en particular la distribución ss', la diferencia entre los errores con signo promedio y el absoluto, una vez mas, indica que errores grandes en las probabilidades individuales no son suficientes para predecir el desempeño, pero muestra que nuestros diagnósticos se correlacionan con el desempeño.

5.8. Cancelaciones entre Correlaciones en Diferentes Combinaciones de Elementos

Por un análisis exhaustivo y sistemático de los casos de dos y tres elementos, pudimos entender casi la totalidad de los elementos principales - tipo de correlación (simétrica,

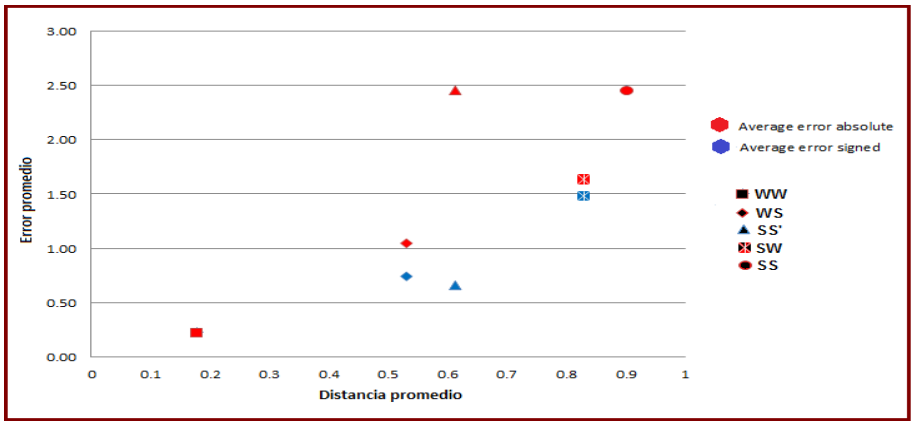


Figura 5.7: Gráfica de distancia promedio como función del error promedio para factorización simétrica de la GBA, en problema tres-elementos con correlaciones simétricas en las funciones de probabilidad.

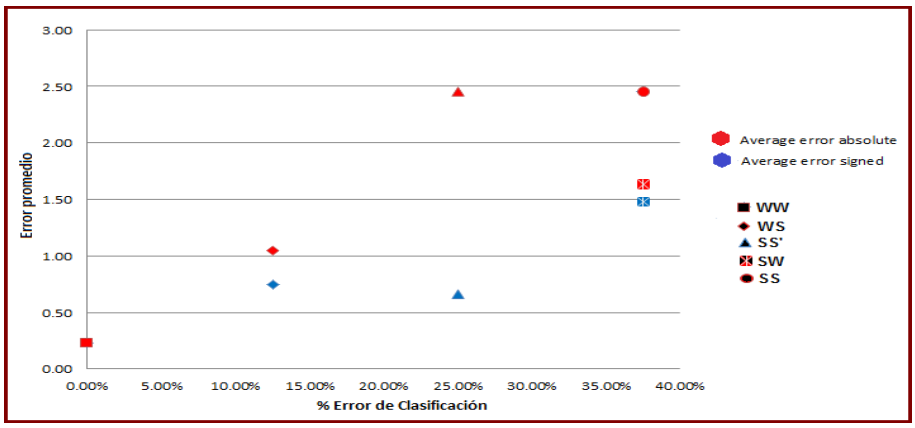


Figura 5.8: Gráfica de error en clasificación como función del error promedio para factorizaciónes simétricas de la GBA, en problema de tres-elementos con correlaciones simétricas en las funciones de probabilidad.

asimétrica), fuerza en la correlación (fuerte, débil) y "correlación correlacion" (reforzamiento, cancelación) - que explican la diferencia relativa en desempeño entre la GBA y NBA. Además, vimos que nuestros diagnósticos se correlacionan muy bien con las diferencias, permitiéndonos predecir a priori cuando se puede esperar que la NBA falle y cual factorización de la GBA deberia ser usada. El anánlisis también nos permite entender porque la NBA actúa de forma tan robusta y versátil a pesar de su fuerte suposición de independencia entre las variables. Esencialmente el único elemento faltante cuando pasa a más de tres variables es el reforzamiento o cancelación de error entre las diferentes combinaciones de elementos, al nivel de cada probabilidad individual para C o $\bar{C}$ en comparación con el reforzamiento o cancelación entre los errores en C y $\bar{C}$.

5.8.1. Cancelaciones para Cuatro Elementos

La ilustración más simple de la cancelación entre diferentes combinaciones de elementos se produce con dos combinaciones de dos variables. Consideremos concatenaciones de dos distribuciones de dos-elementos tomadas del Apéndice A. Explícitamente, consideramos una distribución de probabilidad $P_e(\mathbf{X}|C)$ para cuatro elementos, X_1, X_2, X_3, y X_4, de la forma

$$P_e(\mathbf{X}|C) = P(X_1X_2X_3X_4|C) = P(X_1X_2|C)P(X_3X_4|C) \tag{5.57}$$

$$P_e(\mathbf{X}|\bar{C}) = P(X_1X_2X_3X_4|\bar{C}) = P(X_1X_2|\bar{C})P(X_3X_4|\bar{C}) \tag{5.58}$$

en la que cada $P(X_iX_{i+1}|C)$ se puede elegir de una distribución independiente; por ejemplo, de nuestro conjunto de doce distribuciones. Tenga en cuenta que la estructura de correlación es simétrica en las probabilidades para C y $\bar{C}$. Restringiremos la atención a correlaciones simétricas con el propósito de estudiar el reforzamiento o cancelación de errores entre conjuntos de elementos, no hay nada nuevo en la versión asimétrica. En este caso hay dos esquemas de orden-dos. Como, por construcción no hay dependencias entre los elementos X_iX_j para $ij = 13, 14, 23, 24$ la partición del esquema $\xi^{(e)} = (\xi_1, \xi_2)$ con $\xi_1 = X_1X_2$ y $\xi_2 = X_3X_4$, será exacta, y por lo tanto, la GBA basada en estos esquemas debería ser una aproximación exacta. En otras palabras,

$$\begin{aligned} P(X_1X_2X_3X_4|C) &\equiv P(X_1X_2|C)P(X_3X_4|C) \\ &= P(\xi^{(1)}|C) = P(\xi_1|C)P(\xi_2|C) \\ &\equiv P^e_{GB}(X_1X_2X_3X_4|C) \end{aligned} \tag{5.59}$$

y análogamente para $P(X_1X_2X_3X_4|\bar{C})$. En contraste, la NBA dada para la función de probabilidad es

$$P(\mathbf{X}|C) = P(X_1|C)P(X_2|C)P(X_3|C)P(X_4|C) \tag{5.60}$$

mientras que el error en la función de score relativo a la factorización óptima (exacto), es dado por

$$\begin{aligned} \Delta_{S_{GB}}(C|X_1X_2X_3X_4) &= \delta_s(\xi_1|C) + \delta_s(\xi_2|C) - \delta_s(\xi_1|\bar{C}) - \delta_s(\xi_2|\bar{C}) \\ &= \delta_s(C|\xi_1) + \delta_s(C|\xi_2) \\ &\equiv \ln\left(\frac{P(X_1X_2|C)}{P(X_1|C)P(X_2|C)}\right) - \ln\left(\frac{P(X_1X_2|\bar{C})}{P(X_1|\bar{C})P(X_2|\bar{C})}\right) \\ &+ \ln\left(\frac{P(X_3X_4|C)}{P(X_3|C)P(X_4|C)}\right) - \ln\left(\frac{P(X_3X_4|\bar{C})}{P(X_3|\bar{C})P(X_4|\bar{C})}\right) \end{aligned} \tag{5.61}$$

Ahora podemos empezar a experimentar en cómo diferentes grados de dependencia entre los elementos pueden afectar la validez general de la NBA y, lo más importante aquí, considerar como los errores pueden cancelarse entre diferentes combinaciones de elementos. Tendremos en cuenta diferentes concatenaciones de las distribuciones de probabilidad del Apéndice A, como discutimos en el sección 5.3.1. Vimos ahí, en el caso de dos variables, que para las distribuciones 1, 3, 8, 10 y 11 la NBA fue particularmente mala, debido a las fuertes dependencias entre los atributos, mientras para las distribuciones 2, 4, 7, 9 y 12 la NBA era mejor. Sin embargo, también hemos visto que existen diferentes razones por las cuales la NBA podría funcionar bien en los ejemplos de dos-elementos. En primer lugar, como en la distribución 4, que las correlaciones fueron débiles en ambas probabilidades para C y $\bar{C}$, y alternativamente, como en la distribucíon 2, las correlaciones fueron fuertes en las probabilidades para C y $\bar{C}$ pero habían cancelaciones en los errores entre ellos. Por lo tanto, como bloques de construcción para las concatenaciones usaremos la distribución 4 denotada como W, la cual tiene pequeños errores en la NBA para ambas probabilidades; distribución 1, S, la cual exhibe grandes errores en ambas probabilidades y de signo opuesto; distribución 2, W', la cual tiene grandes errores en ambas probabilidades, pero con cancelaciones entre ellos, y por último, la distribución 1, pero donde C y $\bar{C}$ han sido intercambiadas, S'. Este último artificio tiene el efecto de dar errores para cada combinación de elementos específicos de diferente signo, al correspondiente error de distribución 1. Por lo cual, WW es una concatenación de la distribución 4, SW de las distribuciones 1 y 4; SS de la distribución 1; WW' de las distribuciones 4 y 2, $W'W'$ de la distribución 2 y SS' de la distribución 1 con la distribución 1 donde C y $\bar{C}$ están invertidas.

Con esto en mano, podemos ahora examinar como los errores se cancelan tanto en nivel intra e inter-esquema. En la tabla 5.5 vemos los diferentes errores en las probabilidades para C y $\bar{C}$ para cada esquema $\xi_1 = X_1X_2;\ \xi_2 = X_3X_4$ para la distribución SS. La característica más notable es que para cualquier esquema, los signos de los errores varían entre las diferentes combinaciones de elementos 11, 10, 01, 00. De hecho, como se dijo anteriormente, para elementos binarios los errores en las funciones de probabilidad para X_1X_2 y $\bar{X}_1\bar{X}_2$ son los mismos y opuestos a los de $\bar{X}_1X_2$ y $X_1\bar{X}_2$. Aunque para los elementos de mayor cardinalidad y para esquemas de más de dos elementos la situación es más complicada, sigue siendo cierto que los errores para diferentes valores de elementos no pueden ser todos del mismo signo, y por lo tanto, cuando combinaciones de valores de elementos son considerados en diferentes esquemas es inevitable que habrá cancelaciones.

Este fenómeno puede ser visto claramente en la tabla 5.5: Considerando $\delta_s(C|\xi) = \delta_s(C|X_1X_2) + \delta_s(C|X_3X_4)$, 8 configuraciones, $X_1X_2X_1X_2$ y $X_1X_2\bar{X}_1\bar{X}_2$ conducen a un error mejorado mientras otras 8, $X_1X_2X_1\bar{X}_2$ y $X_1X_2\bar{X}_1X_2$, están asociadas con una cancelación de errores. Lo mismo es cierto para $\delta_s(\bar{C}|\xi)$. Por ejemplo, 1111, 1010 etc. están asociados con cancelación de errores.

Variables	$\Delta S_C(X_1X_2)$	$\Delta S_{\bar{C}}(X_1X_2)$	$\Delta S_C(X_3X_4)$	$\Delta S_{\bar{C}}(X_3X_4)$
1111	-3.15	0.49	-3.15	0.49
1110	-3.15	0.49	0.67	-1.50
1101	-3.15	0.49	0.61	-1.01
1100	-3.15	0.49	-1.80	0.58
1011	0.67	-1.50	-3.15	0.49
1010	0.67	-1.50	0.67	-1.50
1001	0.67	-1.50	0.61	-1.01
1000	0.67	-1.50	-1.80	0.58
0111	0.61	-1.01	-3.15	0.49
0110	0.61	-1.01	0.67	-1.50
0101	0.61	-1.01	0.61	-1.01
0100	0.61	-1.01	-1.80	0.58
0011	-1.80	0.58	-3.15	0.49
0010	-1.80	0.58	0.67	-1.50
0001	-1.80	0.58	0.61	-1.01
0000	-1.80	0.58	-1.80	0.58

Tabla 5.5: Errores para distribución de cuatro-elementos SS.

Este patrón de mejora y cancelación de error es igualmente valido para cualquier distribución de cuatro-elementos binaria. Por lo tanto, con el fin de analizar las diferentes posibilidades de cancelaciones entre las cuatro diferentes funciones de probabilidad, en este problema de cuatro-elementos, consideramos las siguientes cantidades: $\Delta S_i,\ i = 1,\ 2 = \delta_s(C|\xi_i) = \delta_s(\xi_i|C) - \delta_s(\xi_i|\bar{C})$ es la suma de los errores con signo para cada combinación de elementos, mientras $|\Delta S_i| = |\delta_s(\xi_i|C)| + |\delta_s(\xi_i|\bar{C})|$ es la suma de los errores absolutos en las dos probabilidades. Del mismo modo, $\Delta S_C = \delta_s(\xi_1|C) + \delta_s(\xi_2|C)$ es la suma de los errores con signo para las probabilidades de C sumados a través de los dos esquemas ξ_1 y ξ_2. $\Delta S_{\bar{C}} = \delta_s(\xi_1|\bar{C}) + \delta_s(\xi_2|\bar{C})$ es la cantidad análoga para las probabilidades de $\bar{C}$. Finalmente, $\Delta S_{total} = \delta_s(\xi_1|C) - \delta_s(\xi_1|\bar{C}) + \delta_s(\xi_2|C) - \delta_s(\xi_2|\bar{C})$ es el error con signo para el conjunto completo de elementos, mientras $|\Delta S_{total}| = |\delta_s(\xi_1|C)| + |\delta_s(\xi_1|\bar{C})| + |\delta_s(\xi_2|C)| + |\delta_s(\xi_2|\bar{C})|$

es la suma de los errores absolutos a traves de todas las cuatro funciones de probabilidad. Estas cuatro cantidades, $\Delta S_i,\ i = 1,\ 2,\ C,\ \bar{C}$, nos permiten analizar todas las posibles cancelaciones de error, tanto intra-esquema, es decir, entre las probabilidades para C y $\bar{C}$ para los mismos esquemas, e inter-esquema, es decir, entre probabilidades para la misma clase pero en diferentes esquemas.

La tabla 5.6 muestra los valores absolutos de estos diferentes diagnósticos promediados a lo largo de las 16 diferentes combinaciones de elementos $1111, 1110, \ldots, 0000$ para cada concatenación de distribución de probabilidad. Considere primero las distribuciones homogéneas WW y SS: En ambos casos $\Delta S_i = |\Delta S_i|,\ i = 1, 2$, lo que indica que no existen cancelaciones entre los errores en las probabilidades para C y $\bar{C}$ en un esquema dado. En otras palabras, los errores en las probabilidades para C y $\bar{C}$ se refuerzan mutuamente en lugar de cancelarse. Por otra parte, en ambos casos, $\Delta S_i < |\Delta S_i|,\ i = C, \bar{C}$, lo cual indica que hay cancelaciones entre esquemas, para ambas funciones de probabilidad, como se ilustro anteriormente para el caso SS. En el caso de la distribución SW, una vez mas no hay cancelaciones entre los errores para C y $\bar{C}$ en un esquema determinado pero hay entre los esquemas. Para las distribuciones WW', SS' y WW', las tres contienen distribuciones, S' o W', donde hay una cancelación de errores entre las probabilidades para C y $\bar{C}$ dentro de un esquema dado, es decir, que $\Delta S_i < |\Delta S_i|$ para $i = 2$ para WW' y SS', o ambos en el caso de $W'W'$.

Además, hay también cancelaciones entre esquemas para las tres distribuciones. Al comparar ΔS_{total} y $|\Delta S_{total}|$ podemos ver la magnitud de la cancelación del error global para la función de score. Por mucho las reducciones mayores están asociadas con estas distribuciones, WW', SS' y $W'W'$, donde hay cancelaciones en ambos niveles intra e inter -esquema. La menor cancelación es para la distribución SW. Esto es debido al hecho de que es una concatenación de una distribución, S, con errores grandes, con otra, W, con errores pequeños. Sin embargo, aunque el mayor error de cancelación esta asociado con estas distribuciones donde hay ambas cancelaciones intra e inter -esquema no corresponden necesariamente a estas distribuciones, donde la diferencia absoluta promedio entre el score NB y el score exacto es alto. Por el contrario, las mayores diferencias en score están asociadas con aquellas distribuciones donde el score NB es pequeño. Como se mencionó anteriormente, la NBA tiene materia prima inadecuada con la cual trabajar.

	ΔS_1	$\lvert\Delta S_1\rvert$	ΔS_2	$\lvert\Delta S_2\rvert$	ΔS_C	$\lvert\Delta S_C\rvert$	$\Delta S_{\bar{C}}$	$\lvert\Delta S_{\bar{C}}\rvert$	ΔS_T	$\lvert\Delta S_T\rvert$	S_{GNB}	S_{NB}	$\lvert\%\rvert$
WW	0.228	0.228	0.228	0.228	0.263	0.304	0.121	0.152	0.323	0.456	2.511	2.305	50
SW	2.453	2.453	0.228	0.228	1.558	1.709	0.896	0.972	2.453	2.681	2.757	2.091	121
SS	2.453	2.453	2.453	2.453	2.476	3.115	1.258	1.792	3.010	4.907	3.008	0.136	23956
WW'	0.228	0.228	0.662	2.453	1.558	1.709	0.896	0.972	0.754	2.681	2.288	2.091	37
SS'	2.453	2.453	2.403	2.403	1.859	2.403	1.867	2.453	2.681	4.857	2.731	0.134	1667
W'W'	0.662	2.453	0.662	2.403	2.476	3.115	1.258	1.792	1.218	4.907	1.216	0.136	6595

Tabla 5.6: Error promedio para diferentes distribuciones de cuatro-elementos.

5.8.2. Cancelaciones para mas de Cuatro Elementos

Podemos concatenar las distribuciones de dos-elementos tantas veces como se necesite para ver como cambian las cosas, en función del número de elementos y como una función

de la mezcla de correlaciones. Por ejemplo, para seis y ocho elementos tomaremos las distribuciones de probabilidad para las funciones de probabilidad como

$$\begin{aligned} P_e(\mathbf{X}|C) &= P(X_1X_2X_3X_4X_5X_6|C) \\ &= P(X_1X_2|C)P(X_3X_4|C)P(X_5X_6|C) \end{aligned} \tag{5.62}$$

y

$$\begin{aligned} P_e(\mathbf{X}|C) &= P(X_1X_2X_3X_4X_5X_6X_7X_8|C) \\ &= P(X_1X_2|C)P(X_3X_4|C)P(X_5X_6|C)P(X_7X_8|C) \end{aligned} \tag{5.63}$$

con expresiones análogas para $\bar{C}$, de modo que la estructura de la correlación es simétrica en las probabilidades para C and $\bar{C}$.

Para el caso de seis-elementos hay tres esquemas de orden-dos, $\xi_1 = X_1X_2$, $\xi_2 = X_3X_4$, $\xi_3 = X_5X_6$, y para el caso de ocho-elementos cuatro, $\xi_1 = X_1X_2$, $\xi_2 = X_3X_4$, $\xi_3 = X_5X_6$ y $\xi_4 = X_7X_8$ sin dependencias en cualquiera de los casos entre los elementos que son concatenados en diferentes bloques de dos-elementos. La partición del esquema $\xi^e = (\xi_1, \xi_2, \xi_3)$ y $\xi^e = (\xi_1, \xi_2, \xi_3, \xi_4)$, con $\xi_1 = X_1X_2$, $\xi_2 = X_3X_4$, $\xi_3 = X_5X_6$, $\xi_4 = X_7X_8$ será exacta como lo será la GBA basada en esta factorización, entonces

$$\begin{aligned} P(X_1 \dots X_6|C) &\equiv P(X_1X_2|C)P(X_3X_4|C)P(X_5X_6|C) \\ &= P(\xi^{(1)}|C) = P(\xi_1|C)P(\xi_2|C)P(\xi_3|C) \\ &\equiv P^e_{GB}(X_1X_2X_3X_4X_5X_6|C) \end{aligned} \tag{5.64}$$

y

$$\begin{aligned} P(X_1 \dots X_8|C) &\equiv P(X_1X_2|C)P(X_3X_4|C)P(X_5X_6|C)P(X_7X_8|C) \\ &= P(\xi^{(1)}|C) = P(\xi_1|C)P(\xi_2|C)P(\xi_3|C)P(\xi_4|C) \\ &\equiv P^e_{GB}(X_1X_2X_3X_4X_5X_6X_7X_8|C) \end{aligned} \tag{5.65}$$

donde P^e_{GB} es la GBA exacta. En contraste, la NBA dada para la función de probabilidad será

$$P(X_1 \dots X_6|C) = \prod_{i=1}^{6} P(X_i|C) \qquad P(X_1 \dots X_8|C) = \prod_{i=1}^{8} P(X_i|C)$$

Los errores correspondientes en la función de score son

$$\begin{aligned} \Delta_{S_{GB}}(C|X_1 \dots X_6) &= \delta_s(C|\xi_1) + \delta_s(C|\xi_2) + \delta_s(C|\xi_3) \\ &= \ln\left(\frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})}\right) - \ln\left(\frac{P(X_1|C)P(X_2|C)}{P(X_1|\bar{C})P(X_2|\bar{C})}\right) \\ &+ \ln\left(\frac{P(X_3X_4|C)}{P(X_3X_4|\bar{C})}\right) - \ln\left(\frac{P(X_3|C)P(X_4|C)}{P(X_3\bar{C})P(X_4|\bar{C})}\right) \\ &+ \ln\left(\frac{P(X_5X_6|C)}{P(X_5X_6|\bar{C})}\right) - \ln\left(\frac{P(X_5|C)P(X_6|C)}{P(X_5|\bar{C})P(X_6|\bar{C})}\right) \end{aligned}$$

	ΔS_C	$\lvert\Delta S_C\rvert$	$\Delta S_{\bar{C}}$	$\lvert\Delta S_{\bar{C}}\rvert$	ΔS_{total}	$\lvert\Delta S_{total}\rvert$	Score NBG	Score NB	$\lvert\%\rvert$
WWW	0.366	0.470	0.158	0.235	0.410	0.706	3.435	3.237	15.90
SWW	1.625	1.921	0.925	1.082	2.533	3.003	3.557	2.379	506.64
SSW	2.598	3.372	1.311	1.928	3.139	5.300	3.702	2.158	154.54
SSS	3.338	4.823	1.559	2.774	3.838	7.598	3.861	0.175	5708.10
WW'W	1.625	1.921	0.925	1.082	0.863	3.003	2.867	2.379	156.64
SS'S	2.772	4.140	2.177	3.457	3.799	7.598	3.799	0.175	4471.94
WWWW	0.436	0.607	0.179	0.304	0.464	0.912	3.691	3.377	114.07
SWWW	1.601	2.013	0.896	1.124	2.453	3.137	3.879	3.136	110.82
SSWW	2.558	3.419	1.282	1.944	3.076	5.363	4.084	2.306	942.91
SSSW	3.296	4.824	1.530	2.764	3.738	7.588	4.296	2.091	185.62
SSSS	4.021	6.230	1.888	3.584	4.514	9.814	4.511	0.195	8574.92
WW'WW'	2.558	3.419	1.282	1.944	1.351	5.363	3.057	2.306	453.01
SS'SS'	2.993	4.907	2.993	4.907	3.093	9.814	4.023	0.193	2281.02

Tabla 5.7: Error promedio para las diferentes distribuciones seis y ocho -elementos.

$$
\begin{aligned}
\Delta_{S_{GB}}(C|X_1 \ldots X_8) &= \delta_s(C|\xi_1) + \delta_s(C|\xi_2) + \delta_s(C|\xi_3) + \delta_s(C|\xi_4) \\
&= \ln\left(\frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})}\right) - \ln\left(\frac{P(X_1|C)P(X_2|C)}{P(X_1|\bar{C})P(X_2|\bar{C})}\right) \\
&+ \ln\left(\frac{P(X_3X_4|C)}{P(X_3X_4|\bar{C})}\right) - \ln\left(\frac{P(X_3|C)P(X_4|C)}{P(X_3\bar{C})P(X_4|\bar{C})}\right) \\
&+ \ln\left(\frac{P(X_5X_6|C)}{P(X_5X_6|\bar{C})}\right) - \ln\left(\frac{P(X_5|C)P(X_6|C)}{P(X_5|\bar{C})P(X_6|\bar{C})}\right) \\
&+ \ln\left(\frac{P(X_7X_8|C)}{P(X_7X_8|\bar{C})}\right) - \ln\left(\frac{P(X_7|C)P(X_8|C)}{P(X_7|\bar{C})P(X_8|\bar{C})}\right)
\end{aligned}
$$

Tenga en cuenta que en estos ejemplos, hemos construido todas las dependencias entre las diferentes variables explícitamente. En este caso, se eligió como la GBA los esquemas $\xi_1 = X_1X_2$ y $\xi_2 = X_3X_4$, para cuatro elementos, $\xi_1 = X_1X_2$, $\xi_2 = X_3X_4$, $\xi_3 = X_5X_6$ para seis elementos y $\xi_1 = X_1X_2$, $\xi_2 = X_3X_4$, $\xi_3 = X_5X_6$ y $\xi_4 = X_7X_8$ para ocho elementos hemos elegido una GBA que coincide con la factorización exacta de la función de probabilidad y, por lo tanto, la GBA es exacta. La implicación de esto es que el error de la NBA es sólo la suma de los errores asociados con las cuatro dependencias de orden-dos que surgen de los cuatro esquemas ξ_1, ξ_2, ξ_3 y ξ_4.

Consideraremos concatenaciones de las distribuciones W, W', S y S' de la sección 5.8.1. Explícitamente, para seis elementos vamos a considerar las distribuciones WWW, SWW, SSW, SSS, $WW'W$ y $SS'S$ y para ocho elementos las distribuciones $WWWW$, $SWWW$, $SSWW$, $SSSW$, $SSSS$, $WW'WW'$ y $SS'SS'$. En la tabla 5.7, para cada distribución vemos nuestras medidas de error Δ_C, $\Delta_{\bar{C}}$, Δ_{total}, $|\Delta_C|$, $|\Delta_{\bar{C}}|$ y $|\Delta_{total}|$, así como los scores GNB y NB.

Aquí vemos manifestarse el fenómeno de cancelación de errores a ambos niveles intra e inter -esquema. Como una concatención de cuatro distribuciones esta asociada con ocho elementos, el análisis de correlaciones y errores esta a nivel de los muchos ejemplos del mundo real, como puede ser encontrado en el repositorio de datos de la UCI. Tenga en cuenta que los ejemplos con el mayor grado de cancelación, son aquellos con elementos

pares asociados con las distribuciones fuertemente correlacionadas W' y S', que exhiben importantes cancelaciones de error entre las probabilidades para C y $\bar{C}$, dando lugar a cancelaciones de hasta el 75 % del error absoluto, con la principal contribución viniendo de cancelaciones de los errores en Δ_C y $\Delta_{\bar{C}}$. Las distribuciones simétricas WWW, SSS, $WWWW$ y $SSSS$ ilustran muy bien la existencia de cancelaciones a través de diferentes combinaciones de valores de elementos dado que, como se subrayo antes, no todos pueden tener el mismo signo en el error. Por otra parte, este tipo de cancelación aumenta a medida que el número de variables crece. De hecho, vemos que, incluso si hay muy fuertes correlaciones con los errores que se refuerzan entre las probabilidades de C y $\bar{C}$, es decir, sin cancelaciones intra-esquemas, la cancelación del error global debido a la cancelación inter-esquemas es mas del 50 % del error absoluto y esta reducción es casi la misma para el caso débilmente correlacionado $WWWW$.

Estos efectos se pueden ver resumidos en la figura 5.9, para nuestras 19 distribuciones de probabilidad concatenadas. Podemos ver que para un tipo de correlación dada - (WW, WWW, $WWWW$), (SS, SSS, $SSSS$) - el grado relativo de cancelación es una función creciente del número de variables. Esta concatenación repetida de la misma distribución muestra y aísla el efecto de la cancelación de errores intra-esquema, entre las diferentes combinaciones de valores de elementos en diferentes módulos, debido al hecho de que la función de error no puede ser del mismo signo sobre todas las combinaciones de valores de los elementos. También vemos la cancelación mejorada para las distribuciones con módulos W' y S' debido a la adición de cancelación intra-esquema.

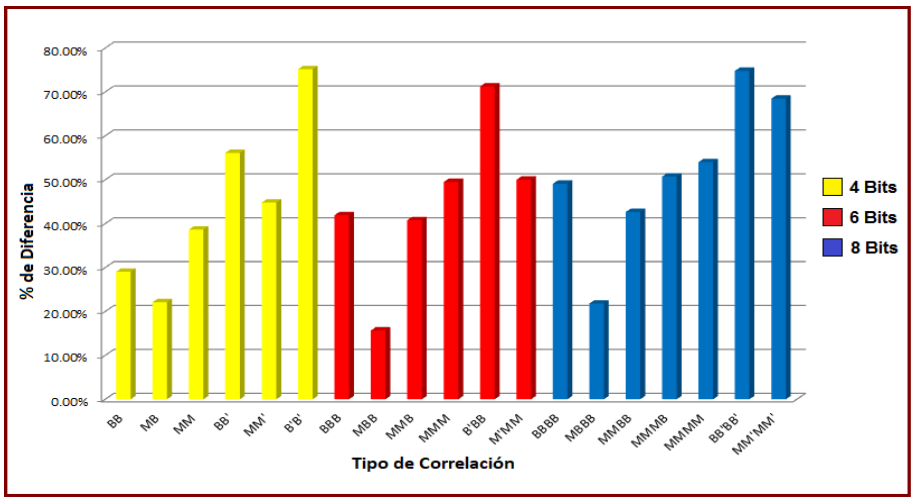


Figura 5.9: Gráfica de la cancelación relativa de score vs tipo de correlación.

5.8.3. Desempeño como Función del Número de Variables y Grado de Correlación

Entones, ¿cómo nuestras medidas de error se relacionan con el desempeño en este caso multi-variable? En las figuras 5.10, 5.11, 5.12 y 5.13 vemos el desempeño de la NBA como función de nuestras medidas de error en score con signo y absoluto, ΔS_{total} y $|\Delta S_{total}|$, promediados sobre todas las combinaciones de valores de elementos para las distribuciones de cuatro, seis y ocho elementos consideradas en la secciones previas. La característica mas notable de las cuatro gráficas es el buen grado de correlación entre la medida del error y la correspondiente medida del desempeño, con R^2 variando de 0.61 a 0.76, muestra

claramente que nuestros diagnósticos de errores predicen el desempeño relativo. También podemos observar en estos gráficos, que las correlaciones entre los puntos asociados con un número fijo de elementos son mas fuertes que cuando se considera el conjunto completo de distribuciones. Esto es reflejado en los valores de R^2 que van de $0.73 - 0.99$.

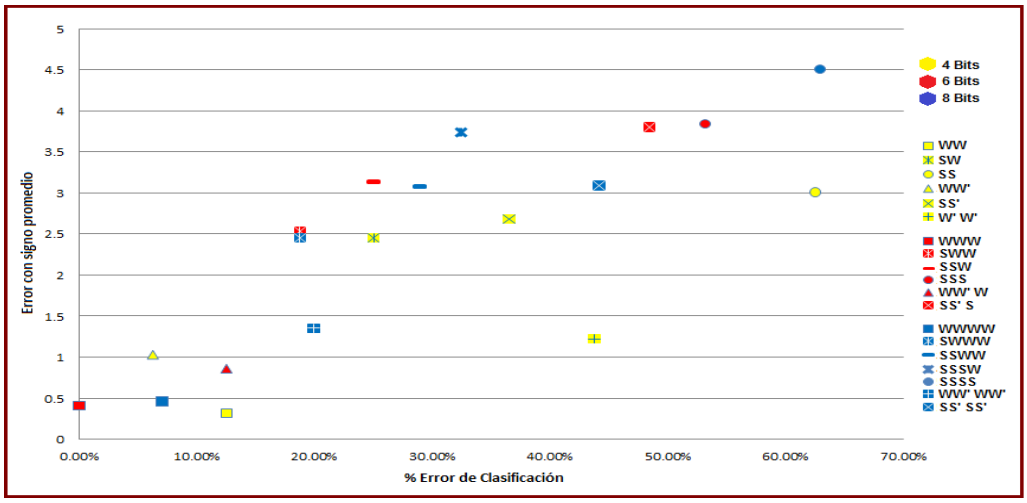


Figura 5.10: Gráfica de error de clasificación vs ΔS_{total}, para las diferentes distribuciones cuatro, seis y ocho - variables.

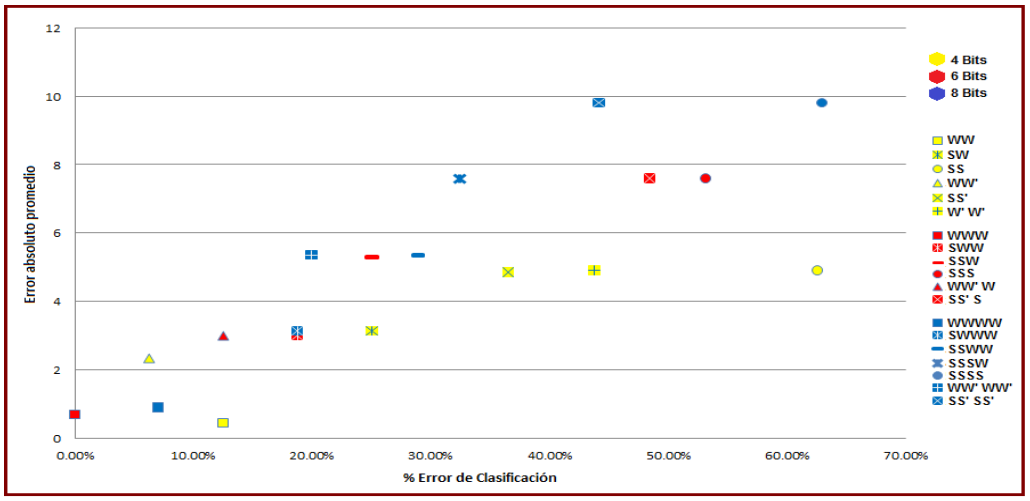


Figura 5.11: Gráfica de error de clasificación vs $|\Delta S_{total}|$, para las diferentes distribuciones cuatro, seis y ocho - variables.

Hemos utilizado dos métricas de desempeño diferentes, en cierto sentido, estas representan dos medidas diferentes de desempeño del clasificador. Como destaca Domingos y Pazzani [49], la naturaleza de la clasificación de todo o nada debe explicar algo de la solidez de la NBA. De hecho, podemos confirmar esto muy bien con el presente análisis. En la figura 5.14, vemos el gráfico de la tasa de falsos positivos contra la distancia para las 19 distribuciones concatenadas que hemos considerado. Podemos ver fácilmente el bajo grado de correlación entre las dos medidas con una $R^2 = 0.32$. ¿Por que la diferencia? Bueno la medida de distancia es un diagnóstico que determina la similitud entre el "ranqueo" de

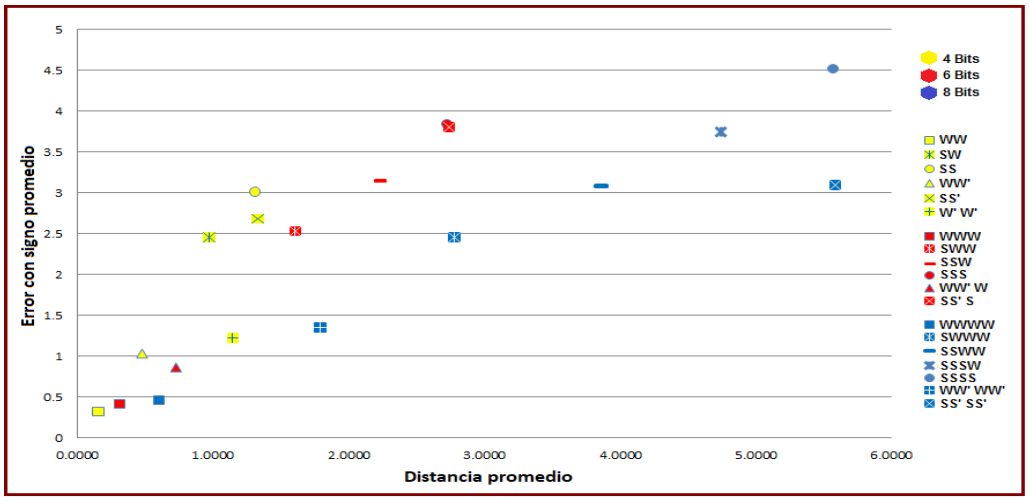


Figura 5.12: Gráfica de distancia vs ΔS_{total}, para las diferentes distribuciones cuatro, seis y ocho - variables.

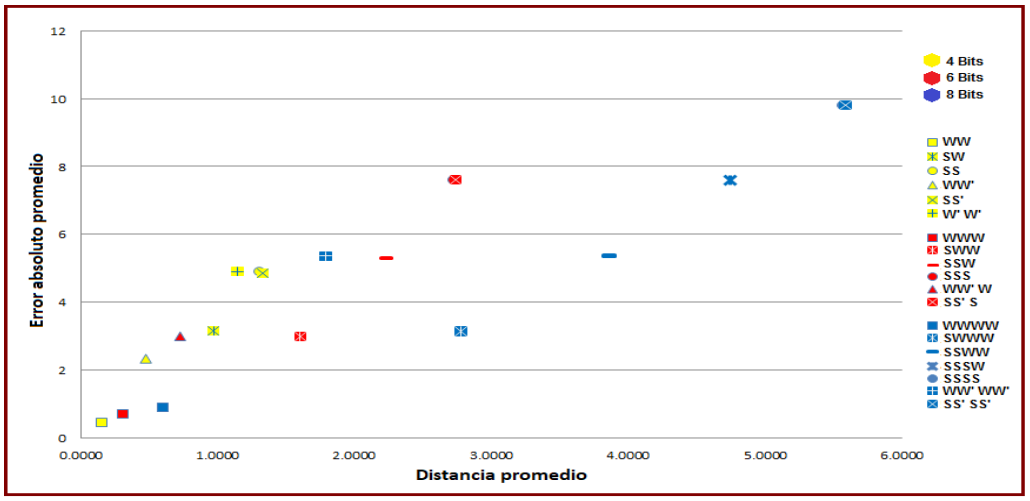


Figura 5.13: Gráfica de distancia vs $|\Delta S_{total}|$, para las diferentes distribuciones cuatro, seis y ocho - variables.

la NBA y GBA (en este caso, exacta) y es global sobre todo el conjunto de predicciones. Por otra parte, el desempeño de la clasificación es particularmente sensible al score NB en las proximidades del umbral de score, es decir, $S_{NBA} = 0$. Esto significa que incluso los errores grandes son relativamente poco importantes si el score NB esta lejos del umbral y, por el contrario, el efecto de los errores pequeños puede ser significativamente amplificado en las proximidades del umbral.

En la figuras 5.16 y 5.14 vemos la relación entre el error de score relativo, es decir, relativo al score NB, contra el error de clasificación (tasa de falsos positivos) y la distancia. Podemos ver claramente la fuerte relación entre el error de score relativo y la tasa de falsos positivos. Esto confirma la importancia de la proximidad del umbral de score $S_{NB} = 0$, donde se esperaría que el error relativo fuera más grande y, por lo tanto, mayor sensibilidad a errores de clasificación. Por el contrario, podemos ver la correlación relativamente débil entre el error de score relativo y la métrica de distancia en la figura 5.16. Esto muestra que las dos métricas son sensibles a bastantes diferentes caracteristicas de la función de error.

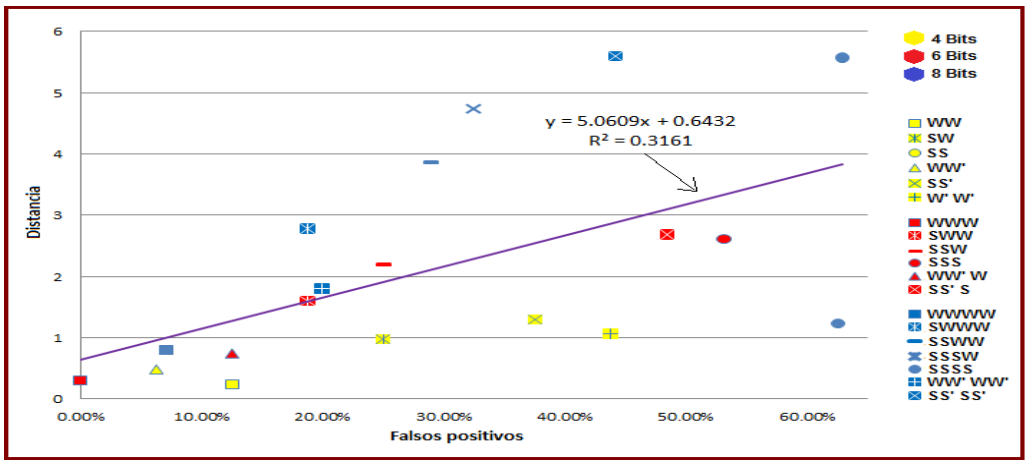


Figura 5.14: Gráfica que muestra la distancia contra la tasa de falsos positivos para las cuatro diferentes distribuciones cuatro, seis y ocho - variables.

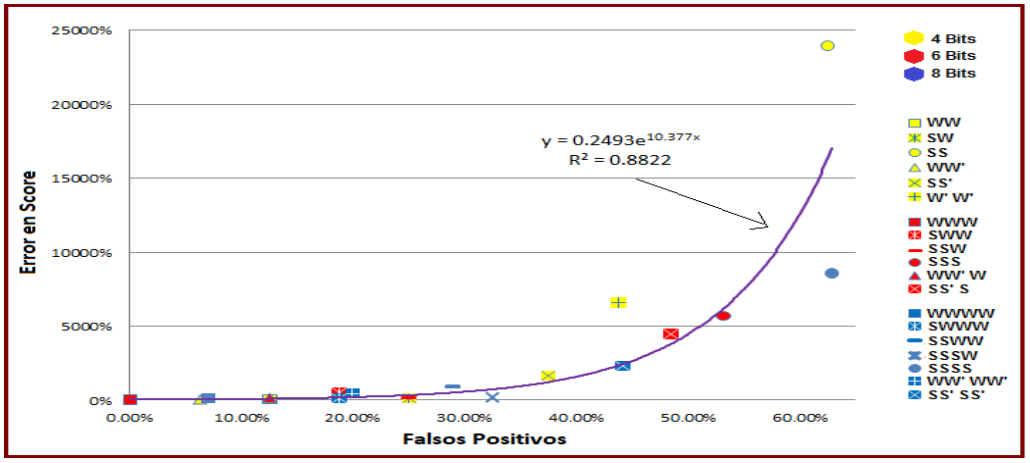


Figura 5.15: Gráfica que muestra el error de score relativo contra la tasa de falsos positivos para las cuatro diferentes distribuciones cuatro, seis y ocho - variables.

5.9. Aplicación sobre datos reales

En esta sección haremos una breve demostración de como los conocimiento obtenidos de los análisis previos pueden ser aplicados a problemas con datos reales. Hemos visto como la medición de los errores locales en las probabilidades ecuación (5.27) y para los scores (5.33), pueden ser usados para determinar cuales atributos deben ser combinados y, por lo tanto, como construir una apropiada factorización para el NBG. Sin embargo, estos diagnósticos están todos asociados con probabilidades y, por lo tanto, independientes del tamaño de la muestra. Por ejemplo, para 2 atributos, si $P(X_1X_2|C) = N_{CX_1X_2}/N_C = 0.3$, $P(X_1|C) = N_{CX_1}/N_C = 0.4$ y $P(X_2|C) = N_{CX_2}/N_C = 0.4$, debemos considerar la posibilidad de que el error $\delta(X_1X_2|C) = 0.14$ no es estadísticamente significativo si $N_{X_1X_2}$, N_C, N_{X_1} o N_{X_2} son pequeños. Para determinar el grado de significancia estadística de los

Figura 5.16: Gráfica que muestra el error de score relativo contra la distancia para las cuatro diferentes distribuciones cuatro, seis y ocho - variables.

errores (5.27) usaremos la siguiente prueba binomial

$$\varepsilon(\xi|\mathcal{C}) \quad = \quad \frac{N_{\mathcal{C}}\delta(\xi|\mathcal{C})}{\sqrt{N_{\mathcal{C}}P_{NB}(\xi|\mathcal{C})(1-P_{NB}(\xi|\mathcal{C}))}} \tag{5.66}$$

donde, como en todo el documento $\mathcal{C} = C$ o $\bar{C}$. Dado que $\delta(\xi|\mathcal{C}) = (P(\xi|\mathcal{C}) - P_{NB}(\xi|\mathcal{C})$ la prueba se toma como hipótesis nula de que no hay correlaciones entre los atributos de ξ. Por lo tanto, la prueba determina el grado en que la actual observación $P(\xi|\mathcal{C})$ es inconsistente con la hipótesis nula. El error en la probabilidad será tomado como estadísticamente significativo si ε excede algún umbral de la hipótesis de prueba. Por ejemplo, en el caso donde la distribución binomial puede ser aproximada por una distribución normal $\varepsilon = 1.96$ correspondería al intervalo de confianza del 95 % de que el error no se produce por casualidad relativo a la hipótesis nula del Naive Bayes. En el caso donde las distribuciones no son bien aproximadas por una distribución normal, la hipótesis de prueba puede usar una aproximación más sofisticada, usando por ejemplo los intervalos de Wilson. Otro problema que se presenta en las muestras finitas es la posibilidad de tener $N_{\mathcal{C}\xi} = 0$. Esto implica que $P(\xi|\mathcal{C}) = 0$ lo cual dará lugar a contribuciones al score infinitas para el esquema correspondiente. Para evitar esto algún tipo de suavizamiento debe ser usado como puede ser correción de Laplace o m-estimados.

Vamos a utilizar como criterio los parámetros $\varepsilon(\xi|C) > 2$ y $\varepsilon(\xi|\bar{C}) > 2$ para determinar los conjuntos de atributos que deben ser usados combinados para todas las combinaciones de valores de atributos posibles. Consideraremos dos algoritmos de combinación diferentes: uno donde los atributos se combinan de forma independiente en ambas probabilidades - el NBG asimétrico y el otro donde los atributos se combinan juntos - el NBG simétrico. Por ejemplo en el primer caso si $\varepsilon(X_1X_2|C) = 3.1$ y $\varepsilon(X_1X_2|\bar{C}) = 0.9$, entonces los atributos X_1 y X_2 se combinarían solo en la probabilidad de C y no en la $\bar{C}$. Por otro lado, en el caso simétrico estos serán combinados en ambas probabilidades. Puede ocurrir que un determinado atributo califique para ser miembro de más de una combinación de atributos. Por ejemplo, para 3 tributos, si $\varepsilon(X_1X_2|C) = 3.1$, $\varepsilon(X_1X_3|C) = 2.4$ y $\varepsilon(X_2X_3|C) = -0.2$, entonces el atributo X_1 califica para ser combinado con ambos X_2 y X_3. Si $\varepsilon(X_1X_2X_3|C) = 1.8$, o si se restringe a solo combinaciones binarias, entonces X_1 solo puede ser combinado con un solo atributo para un vector de atributos dado. En este caso elegimos la combinación de atributos con el valor más alto de ε. En el presente ejemplo, esto significaría que X_1 es combinado con X_2, por que $\varepsilon(X_1X_2|C) > \varepsilon(X_1X_3|C)$. En el caso donde diferentes combinaciones den como resultado el mismo valor de ε entonces

el valor de δ_s es utilizado para romper el empate. Para simplificar, consideramos solo esquemas binarios, es decir, se combinaron solo hasta dos atributos. Además de la sencillez, la otra razón previamente discutida para esto es que las muestras de dos atributos son inevitablemente más grandes que las muestras para tres atributos y por lo tanto, con todo lo demás, conduce a valores más altos de ε.

Se consideraron 20 bases de datos del repositorio de UCI, las cuales se pueden ver en la tabla 5.8. Por simplicidad consideramos cada problema como un problema de dos clases. En los problemas multi-clase se escogió el caso donde la clase era la de muestra más pequeña y en los casos con clases binarias, se escogió la clase especificada en el repositorio de UCI. Los atributos con valores numéricos se discretizaron dividiendo en un número de intervalos fijos, los cuales fueron elegidos de tal forma que cada uno de estos intervalos contenía aproximadamente el mismo número de elementos. 10 intervalos fueron escogido por default. Sin embargo, en el caso de bases de datos con pocos elementos consideramos un número menor de intervalos. Para cada base de datos se realizo un sub-muestreo eligiendo al azar con una división 70/30 entrenamiento/prueba repetida 20 veces. Tenga en cuenta que no se considero ninguna sintonización de la GBA. Por ejemplo, ningun algoritmo de selección de características fue utilizado. Los atributos en el esquema ξ se combinan cuando $|\varepsilon(\xi)| > 2$. Como métricas para medir desempeño utilizamos error de clasificación y AUC. Comparamos nuestras dos aproximaciones GBA - simétrico y asimétrico contra el NBA implementado en WEKA, así como con otros 3 clasificadores de los mas utilizados y disponibles en WEKA: AODE, WAODE y HNB. Cada clasificador fue ejecutado en exactamente los mismos conjuntos de datos entrenamiento/prueba para cada una de las 20 corridas. Se utilizo el suavizado de Laplace contenido por defecto en WEKA, donde $(N_{CX}/N_C) \to (N_{CX} + 1/2)/(N_C + 1)$. Los resultados pueden ser vistos en las tablas 5.8 y 5.9 para error de clasificación y AUC respectivamente. Comparamos los errores y AUC de los diferentes clasificadores, donde se promedio el error sobre las 20 diferentes ejecuciones. También utilizamos una prueba binomial para determinar el significado estadístico de la diferencia en desempeño de los 5 clasificadores mejorados, tomando como hipótesis nula el NBC de WEKA. Consideramos que la diferencia en desempeño era significativa si estaba al nivel $p < 0.05$. Las celdas sombreadas en color azul son aquellas donde hubo una mejora significativa en el desempeño del clasificador mejorado relativo con al NBC y en rojo para los casos donde NBC es significativamente mejor. Los sombreados en amarillo corresponden a diferencias estadísticamente no significativas considerando las 20 ejecuciones. Como las comparaciones de múltiples pares pueden ser problemáticas decidimos utilizar también la prueba de ranqueo de Wilcoxon [50], comparando cada clasificador mejorado contra el NBC. Mostramos la correspondiente estadística Z definida como

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{5.67}$$

donde N es el número de conjuntos de datos, $T = min(R_+, R_-)$ y R_+ es la suma de ranqueos para los conjuntos de datos en los cuales el NBC supera al GBA y R_- es la suma de ranqueos en el caso contrario. La hipótesis nula de que los algoritmos tienen igual desempeño puede ser rechazada con un nivel de confianza del 95 % si $z < -1.96$. En términos de error de clasificación podemos ver que la prueba de ranqueos de Wilcoxon muestra que todas las versiones NBG son significativamente mejores comparadas con NBC. Por otra parte, en términos de AUC vemos que, usando la prueba de ranqueos de Wilcoxon, solo el clasificador AODE muestra una mejora en el desempeño significativamente estadística relativa al NBC. El fuerte desempeño del NBC como un algoritmo de ranqueo ha sido ampliamente demostrado [51, 52], donde se ha demostrado que en términos de ranqueo el NBA es mejor o equivalente a C4.4.

Dominio	Atributos	Casos	Error NB	Error GNB_S	Error GNB_A	Error AODE	Error WAODE	Error HNB
mushroom	22	8000	4.75 %	0.13 %	0.16 %	0.04 %	0.01 %	0.04 %
pendigits	17	10992	5.01 %	1.55 %	2.05 %	0.74 %	0.44 %	1.07 %
segment	20	2310	12.74 %	6.35 %	6.39 %	2.81 %	2.18 %	2.61 %
vehicle	19	846	13.34 %	5.38 %	6.30 %	3.94 %	3.59 %	2.54 %
anneal	39	898	2.39 %	0.72 %	0.78 %	1.76 %	0.85 %	0.74 %
chess (kr-kp)	37	3169	12.04 %	7.48 %	7.71 %	8.84 %	6.13 %	7.41 %
hypothyroid	26	3163	3.60 %	2.22 %	2.43 %	2.40 %	1.84 %	2.01 %
letter recognition	17	20000	1.79 %	1.36 %	1.46 %	1.05 %	1.12 %	1.09 %
satellite	37	6435	15.12 %	13.13 %	13.49 %	12.22 %	15.53 %	14.38 %
adult	15	48842	18.17 %	15.94 %	15.96 %	15.65 %	15.23 %	15.96 %
house-votes	17	435	10.08 %	8.73 %	8.73 %	6.17 %	5.59 %	6.44 %
tic-tac-toe	10	958	30.17 %	29.27 %	30.00 %	25.99 %	27.26 %	23.47 %
ionosphere	35	351	10.65 %	10.48 %	10.52 %	8.22 %	6.32 %	7.27 %
credit crx	16	690	13.31 %	16.71 %	16.16 %	12.44 %	13.59 %	13.67 %
hepatitis	20	155	19.36 %	16.41 %	15.98 %	17.55 %	18.08 %	17.22 %
cancer (bcw)	10	699	2.84 %	3.16 %	2.99 %	3.26 %	3.79 %	4.43 %
statlog (heart)	14	270	15.00 %	17.59 %	17.41 %	15.31 %	16.91 %	17.41 %
post-operative	9	90	32.09 %	28.33 %	28.33 %	32.27 %	34.85 %	36.16 %
liver (bupa)	7	345	33.86 %	37.96 %	36.75 %	34.92 %	35.90 %	36.14 %
wine	14	178	2.81 %	1.79 %	0.85 %	1.03 %	2.44 %	1.13 %
Wilcoxon Test Z				-2.35	-2.48	-3.43	-2.48	-2.37

Tabla 5.8: Error para NB, GNB_s, GNB_a, AODE, WAODE y HNB para los 20 conjuntos de datos de UCI.

Podemos, por supuesto, analizar estos resultados desde la perspectiva de "diseño de algoritmos que den un mejor desempeño a lo largo de un amplio conjunto de dominio de problemas", comparando nuestros NBG simétrico y asimétrico a los 3 clasificadores establecidos AODE, WAODE y HNB. Como se ha subrayado, nuestro objetivo en este trabajo no fue diseñar un nuevo clasificador. Tampoco era para demostrar el conocimiento de, y diagnósticos para, los errores inherentes en el NBA pueden ser utilizados para identificar cuales tributos deben ser combinados y en estas combinaciones incluido tener un clasificador de mejor desempeño, aunque eso es, de hecho, un resultado de esta investigación. Nuestro objetivo era demostrar que nuestros diagnósticos podían predecir a priori cual clasificado: NBC o GBC funcionaria mejor y en cuales problemas.

Lo que estos diagnósticos hacen para predecir se manifiesta en el alto grado de correlación entre la métrica de error elegida y el desempeño del clasificador. Explicitamente para los conjuntos de datos de UCI consideramos la relación entre el promedio del valor absoluto del error con signo ΔS_{total}, promediado sobre todos los vectores atributos en el conjunto de entrenamiento vs la diferencia relativa en el error de clasificación entre el NBC y el GBC, donde para calcular el ΔS_{total} para un vector de atributos dado incluimos solo combinaciones con errores estadísticamente significativos. Los valores grandes de error

Dominio	Atributos	Casos	AUC NB	AUC GNB_S	AUC GNB_A	AUC AODE	AUC WAODE	AUC HNB
mushroom	22	8000	99.76 %	99.87 %	99.86 %	100.00 %	100.00 %	100.00 %
pendigits	17	10992	97.68 %	99.44 %	99.28 %	99.96 %	99.96 %	99.93 %
segment	20	2310	94.66 %	98.12 %	98.30 %	99.47 %	99.69 %	99.62 %
vehicle	19	846	93.59 %	97.60 %	97.41 %	98.86 %	99.03 %	99.31 %
anneal	39	898	99.66 %	99.48 %	99.42 %	99.78 %	99.92 %	99.95 %
chess (kr-kp)	26	3163	95.35 %	98.11 %	98.03 %	97.29 %	98.50 %	98.25 %
hypothyroid	17	20000	98.66 %	98.76 %	98.61 %	98.69 %	98.86 %	98.82 %
letter recognition	37	6435	97.95 %	98.81 %	98.78 %	99.41 %	99.67 %	99.75 %
satellite	15	48842	92.39 %	92.89 %	93.05 %	93.36 %	93.73 %	94.32 %
adult	17	435	90.00 %	90.53 %	90.56 %	90.92 %	90.94 %	89.24 %
house-votes	10	958	97.41 %	96.17 %	96.17 %	98.72 %	98.74 %	98.69 %
tic-tac-toe	35	351	73.38 %	74.35 %	73.07 %	82.09 %	79.54 %	84.67 %
ionosphere	16	690	95.10 %	92.55 %	93.00 %	97.98 %	98.11 %	97.86 %
credit crx	20	155	93.25 %	89.52 %	89.63 %	93.46 %	92.90 %	92.74 %
hepatitis	10	699	87.43 %	83.67 %	84.41 %	88.00 %	85.46 %	86.64 %
cancer (bcw)	14	270	99.25 %	97.79 %	97.78 %	99.92 %	99.09 %	99.11 %
statlog (heart)	9	90	90.98 %	87.60 %	87.86 %	90.93 %	90.04 %	90.31 %
post-operative	7	345	35.50 %	38.88 %	39.00 %	28.51 %	31.48 %	28.81 %
liver (bupa)	14	178	70.57 %	62.69 %	65.94 %	69.02 %	67.46 %	67.58 %
wine	37	3169	99.88 %	97.26 %	97.26 %	99.90 %	99.88 %	99.92 %
Wilcoxon Test Z				-0.28	-0.34	-2.32	-1.65	-1.41

Tabla 5.9: AUC para NB, GNB_s, GNB_a, AODE, WAODE y HNB para los 20 conjuntos de datos de UCI.

corresponden a aquellos conjuntos de problemas que muestran correlaciones significativas cuando se promedian sobre el conjunto de entrenamiento completo y, por lo tanto, se podría esperar que el NBC es menos efectivo. En la tabla 5.10 presentamos un resumen de estas correlaciones para los 20 conjuntos de datos de UCI y las 19 distribuciones artificiales consideradas en el cap. 5.8.2. Para los conjuntos de datos de UCI tomamos como medidas de desempeño la diferencia relativa en error entre el GNB y el NBC, mientras para las distribuciones artificiales tomamos el error en si mismo, en este caso la GNB es exacta por construcción. En la tabla 5.10 se observa que los coeficientes de correlación de Pearson para nuestro GBC simétrico y asimétrico son mas pequeños en los conjuntos de datos de UCI de las que se presentaban en las distribuciones artificiales. Sin embargo todas las correlaciones mostradas son estadísticamente significativas con un nivel del 95 % de confianza, dando una clara evidencia de que nuestra medida de error es predictiva del desempeño del GBC contra el NBC. Curiosamente, nuestra medida de error es también un buen predictor del desempeño para los clasificadores AODE, WAODE y HNB. Dado que nuestros GBCs simétricos y asimétricos hechos a la medida fueron diseñados para tomar en cuenta errores surgidos por tomar en cuenta las correlaciones significativas de los atributos combinados en alternativa a la factorización de las probabilidades, tal vez no sea sorprendente que los desempeños de los clasificadores GNB_s y GNB_a relativos al NBC son mayores cuanto mayor es la magnitud de las correlaciones de los atributos. Sin embargo, es gratificante observar que la relación de desempeño de los clasificadores

AODE, WAODE y HNB esta también altamente correlacionada con nuestros diagnósticos de error. Lo cual esta ligado al hecho de que todos los clasificadores que consideramos están tratando de tomar en cuenta las dependencias entre los atributos mediante cambiar al máximo el criterio de factorización de NB, simplemente lo hacen de diferentes maneras.

Classifier	Correlation Coefficient	Correlation Coefficient (no ionosphere)
GNB_s UCI	-0.45	-0.65
GNB_a UCI	-0.47	-0.68
AODE UCI	-0.61	-0.77
WAODE UCI	-0.53	-0.59
HNB UCI	-0.52	-0.62
Avg All	-0.54	-0.68
NBC Artificial	0.78	NA

Tabla 5.10: Coeficientes de correlación de Pearson entre el promedio del error absoluto $|\Delta S_{total}|$, promediado sobre todos los vectores de atributos en el conjunto de entrenamiento vs la diferencia relativa en el error del clasificador entre la NBC y GBC para cada clasificador. Todos los coeficientes de correlación son estadísticamente significativos con un nivel de confianza del 95 %.

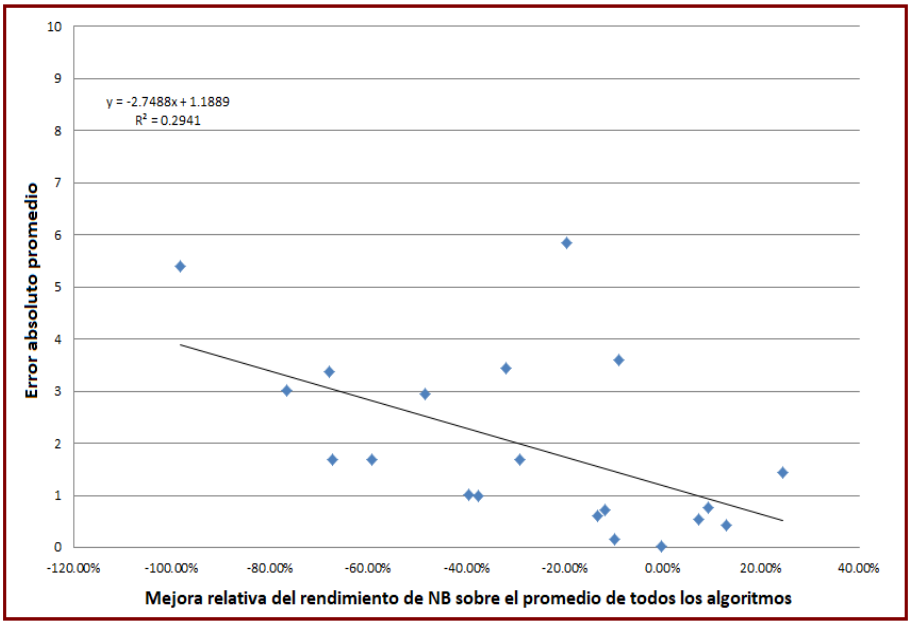


Figura 5.17: Relación entre el porcentaje de diferencia relativa en error entre la NBA y el GBA y el error promedio absoluto para las 20 bases de datos de UCI y promediado sobre los 5 clasificadores GBA.

En la figura 5.17 podemos ver gráficamente las correlaciones entre el error y la mejora del desempeño relativo de los conjunto de datos de la UCI promediados sobre todos los 5 clasificadores mejorados GNB. También podemos notar la presencia de un caso atípico - el conjunto de datos de ionosphere - donde hay un error de correlación muy alta pero solo una pequeña mejora de la GBA sobre el NBA. En efecto, si consideramos los coeficientes

de correlation de Pearson para los conjuntos de datos de UCI en la tabla 5.10 sin el valor atípico de ionosphere, podemos ver un incremento significativo en los coeficientes de correlación. Por supuesto, no deseamos alterar los datos con el fin de mejorar los resultados pero deseamos utilizar este caso para señalar que, si bien es destacable el grado de correlación entre nuestra muy simple medida de error y el desempeño relativo del GBC frente al NBC, sin duda habría que esperar la existencia de otros factores potenciales, potencialmente muchos, que afectan la relación entre estos. Por ejemplo, podemos señalar que el conjunto de datos de ionosphere tiene una fracción sustancialmente menor de casos para valor de atributos, comparado con otros conjuntos de datos. También podemos señalar que la distribución de los errores de correlación contienen potencialmente una gran cantidad de información, la cual podría ser utilizada para intuir el desempeño de un clasificador, mas allá como se ha utilizado aquí como solo una medida general. Finalmente, en la figura 5.18 vemos una gráfica de la distancia promedio por caso, entre las clasificaciones NBC y GBC para el conjunto de vectores de atributos del conjunto de entrenamiento vs la diferencia relativa del error de clasificación entre el GBC y el NBC. Como podemos ver, hay una clara correlación lo cual implica, que el NBC es de mejor desempeño en condiciones en las cuales el rango de distancia promedio por el número total de casos es grande. Dada la enorme heterogeneidad de los problemas de UCI es muy gratificante ver que hay métricas de diagnostico las cuales diferencian entre aquellos problemas donde la NBA, seria adecuada vs aquellos donde un algoritmo mas sofisticado es requerido. De hecho, estos resultados muestran el potencial de desarrollar algoritmos de meta-predicción, los cuales predigan el desempeño de un algoritmo dado para un problema dado.

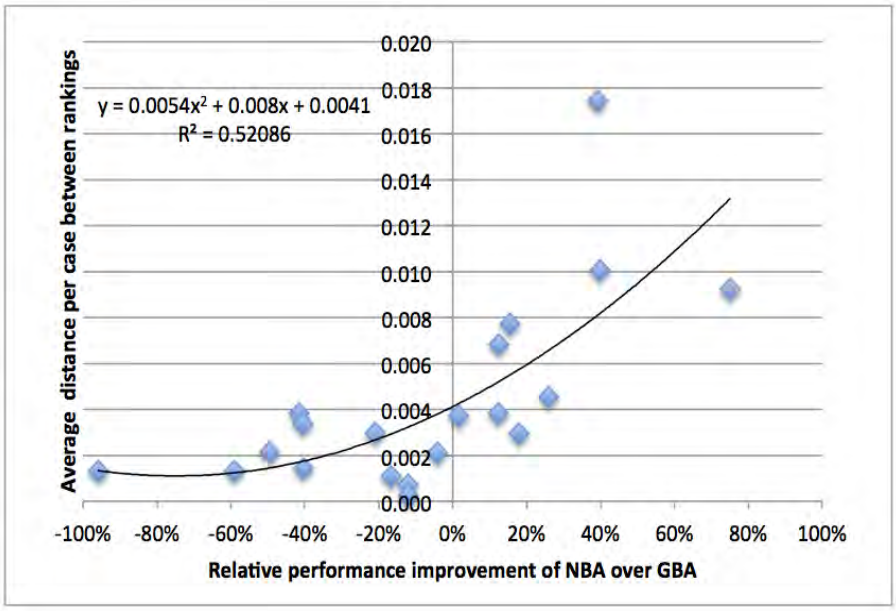


Figura 5.18: Relación entre el porcentaje de la diferencia en error relativo entre los clasificadores NBA y GBA_s y GBA_A y la distancia promedio en el ranqueo por caso.

5.10. Conclusiones Particulares

En cuanto a la comprensión de por que es tan robusta (la NBA), han habido distintas propuestas. Una hipótesis es que es un artefacto de la naturaleza de la medida del desempeño de la mayoría de los problemas donde ha sido aplicado - clasificación binaria. En este

caso, solo el "ranqueo" de un ejemplar con respecto a algún umbral es de importancia, por lo que solo aquellos casos donde el error en la NBA es suficientemente grande para cambiar el ejemplar de un lado del umbral al otro puede causar errores de clasificación. Una segunda linea de pensamiento ha sido la posibilidad de que no sea una condición suficiente la dependencia entre elementos para validar la NBA, mas bien es la distribución de dependencias entre clases y entre los propios elementos, lo que gobierna su validez. Sin embargo, hasta ahora, no ha habido ninguna cuatificación de como exactamente las dependencias pueden cancelar, bajo cuales circunstancias, y como medirlo. Este último punto es fundamental, si es posible cuantificar las desviaciones de la NBA, entonces es posible predecir a priori cuando y bajo que circunstancias, la NBA será adecuada y, por lo tanto, saber cuando hay que implementar un más sofisticado, pero más costoso, algoritmo alternativo. Por otra parte, mediante un mejor entendimiento de la relación entre la estructura del problema y la estructura del algoritmo, tal visión debe permitir el diseño de clasificadores hechos a la medida y que son mas adecuados para un problema dado. Por supuesto, un tercer atributo de la NBA, que ofrece una ventaja relativa, es que esta basado en una factorización máxima de las probabilidades, así cada factor esta asociado con una muestra mas grande y, por lo tanto, de errores de muestreo pequeños, que cualquier otro factor compuesto con más de una variable. Esto se refiere mas a la ventaja relativa de la NBA en términos de la varianza del modelo. En este trabajo, sin embargo, nos interesa por el momento solo en la cuestión del sesgo del modelo de la NBA contra sus generalizaciones, GBA.

Hemos desarrollado un marco que puede ser usado para determinar la presencia de dependencias entre los atributos dentro de una combinación de variables arbitraria (esquema), y más importante, determinar cuando, como y en que medida, afectan a la estimación de las diferentes métricas de desempeño, tales como la estimación de la probabilidad posterior y exactitud en la clasificación utilizando la NBA. Nuestro análisis se basa en el supuesto, utilizado en generalizaciones de la NBA, de que existen mejores factorizaciones, por ejemplo de las funciones de probabilidad $P(\mathbf{X}|C)$ y $P(\mathbf{X}|\bar{C})$, que la factorización completa asociada con la NBA. La pregunta entonces es: ¿Cual es el error asociado con una factorización dada, es decir, una realización determinada de la GBA, con relación a la NBA? En el análisis de este error, no hicimos ningún supuesto de que la factorización óptima de $P(\mathbf{X}|C)$ fue el mismo que para $P(\mathbf{X}|\bar{C})$, presentando evidencia de que hay problemas (distribuciones de probabilidad) donde manifiestamente no eran los mismos. Demostramos que las cancelaciones pueden ocurrir y ocurren en ambos niveles intra e inter - esquema, la primera mostrando que las cancelaciones pueden ocurrir entre las probabilidades de una clase y su complemento, y el segundo muestra que las cancelaciones pueden ocurrir entre las probabilidades para diferentes combinaciones de variables pero la misma clase. De hecho, demostramos que era inevitable que hubieran dichas cancelaciones inter-esquema cuando los signos del error para una combiánción de variables dada para diferentes valores de variable no podran ser todas las mismas.

Con el fin de cuantificar el grado de cancelación de error, o, de hecho, el reforzamiento del error, introdujimos un conjunto de diagnósticos - $\delta(\xi|C)$, $\delta(\xi|\bar{C})$, $\delta(\xi)$, $\delta_s(\xi|C)$, $\delta_s(\xi|\bar{C})$, $\Delta_C(\xi)$, $\Delta_{\bar{C}}(\xi)$, $\Delta(\xi)$, $\Delta_s(\xi|C)$ y $\Delta(\xi)$. $\delta(\xi|C)$, y su análogo para $\bar{C}$, medida del grado de dependencia entre los atributos dentro de un esquema dado en la expresión para la función de probabilidad para la clase C, o su complemento, $\bar{C}$. Demostramos que fuertes dependencias, según lo exhibido por $\delta(\xi|C)$ y $\delta(\xi|\bar{C})$, no fueron suficientes para dar lugar a errores significativos en la estimación de probabilidades posteriores o en la clasificación. Mas bien, mostramos como, para un esquema dado, la distribución de dependencias a través de C y $\bar{C}$ es lo que controla la exactitud de la NBA. Demostramos que los errores se maximizaban cuando los errores en las probabilidades para C y $\bar{C}$ eran de gran magnitud

y, sobre todo, de signo opuesto. Como un diagnóstico correspondiente introducimos $\Delta(\xi)$.

Hemos demostrado que el análisis de errores es más simple en términos de la función de score, donde independientemente de los errores tenían un carácter aditivo y la función de error es una medida de la no-linealidad en el score para el esquema ξ relativo con su homóloga NBA, lo cual corresponde a una suma de scores para cada atributo individual de ξ. También vimos que aunque los errores podrían cancelar/reforzar "localmente", es decir, entre las probabilidades para C y $\bar{C}$ dentro del mismo esquema, debido a la existencia de cancelaciones intra-esquema, un análisis completo para un conjunto de variables dado fue una cuestión "global", donde no era posible decir si el error total para un conjunto dado era grande o no hasta que todas las contribuciones se hubieran calculado. Derivamos formulas explicitas que relacionan errores locales, en subconjuntos de variables potencialmente distintos para cada probabilidad, a errores globales sobre el conjunto completo de variables.

Para relacionar el análisis de errores al desempeño del modelo consideramos distintas métricas: i) exactitud en la clasificación para la NBA o GBA; ii) estimación de las probabilidades posteriores $P(C|\mathbf{X})$ en la NBA o GBA; y, iii) la distancia entre las posiciones relativas de la NBA y GBA. Por supuesto, la distribución actual y el impacto de las correlaciones de atributos en estas métricas de desempeño, depende de las propiedades precisas de la estructura de la correlación subyacente de las distribuciones de probabilidad, que se esta tratando de estimar. Como distribuciones del mundo-real estan asociadas con finitas y muy frecuentemente pequeñas, muestras, los errores de muestreo y, por lo tanto, la varianza del modelo, juegan un papel importante para distinguir entre la NBA y la GBA. Como nuestro interés aquí es la comprensión del papel de las correlaciones de los elementos como la causa del sesgo del modelo sin embargo, optamos por restringir la atención a un conjunto de artificiales, pre-especificadas distribuciones de probabilidad, donde el grado y el tipo de correlación podía ser elegido y ajustado para ilustrar el efecto e impacto de las correlaciones. En concreto, hemos propuesto un conjunto de 12 distribuciones de probabilidad para dos variables binarias que ilustraban diferentes características cuantitativas, que nosotros creemos dan una idea sustancial del funcionamiento interno de la NBA y sus generalizaciones. En esencia, las distribuciones capturan las nociones de correlaciones fuertes vs débiles y correlaciones que refuerzan o se cancelan entre las probabilidades.

Examinamos correlaciones en detalle para las 12 distribuciones de prueba de dos-elementos, mostrando como el error significativo en la NBA dependía fundamentalmente de lo signos relativos de los errores para C y $\bar{C}$; reforzamiento y cancelación de errores asociados con signos opuestos e iguales en $\delta(\xi|C)$ y $\delta(\xi|\bar{C})$ respectivamente. Examinamos el impacto de las diferentes estructuras de correlación sobre el desempeño del modelo y mostramos que nuestros diagnósticos se correlacionan bien con el desempeño del modelo. En otras palabras, cuanto mayor sea el error de acuerdo con nuestros diagnósticos peor el desempeño de la NBA. Esto valida los diagnósticos como predictores potenciales del desempeño de la NBA. Al considerar una clase de generalizaciones de la NBA, donde las probabilidades no están factorizadas al máximo, hemos visto el impacto de la elección de la factorización hecha, en comparación con no, respecto a alguna estructura correlación subyacente en el contexto de un conjunto de distribuciones de probabilidad de 3-variables derivados de nuestro conjunto de 12 distribuciones de dos-variables. Demostramos que una factorización GBA que capturó la estructura de correlación del problema subyacente condujo inevitablemente a un mejor desempeño. También vimos el impacto de las correlaciones que no eran simétricamente distribuidas entre las probabilidades de C y $\bar{C}$, es decir, que involucró a diferentes combinaciones de elementos. Además, demostramos que nuestros diagnósticos de errores se correlacionan bien con la estructura del problema sub-

yacente, lo cual indica que la factorización fue óptima.

Al nivel de dos y tres variables solo errores de cancelación intra-esquema son visibles. Para investigar el rol de las cancelaciones inter-esquema ampliamos nuestro análisis a distribuciones de cuatro, seis y ocho variables, las cuales fueron concatenaciones de nuestras distribuciones originales de dos-variables. Estudiamos un conjunto de 19 distintas, distribuciones concatenadas con diferente número de elementos y estructuras de correlación, mostrando como el error global fue una propiedad emergente, como resultado de un conjunto de cancelaciones y reforzamientos de los errores locales en ambos niveles intra e inter - esquemas. En particular, hemos visto el impacto del hecho de que los errores de cualquier esquema dado tenían diferentes signos a través de los distintos valores de las variables, lo cual garantiza la existencia de cancelaciones de errores inter-esquema. Hemos demostrado que la máxima cancelación de errores ocurrió en distribuciones de probabilidad las cuales exhibían cancelaciones tanto intra como inter - esquemas. También mostramos que el desempeño del modelo estaba altamente correlacionado con nuestras funciones de error global, validando así, una vez mas su valor como diagnóstico predictivo para el desempeño relativo de la NBA. También vimos que diferentes métricas de desempeño fueron más sensibles a las diferentes características de la distribución de error, con error de clasificación siendo particularmente sensible a errores en las probabilidades cerca del umbral del score NB pero insensibles a esas lejos del umbral. Por el contrario, nuestra métrica de distancia como una medida del "ranqueo" global fue igualmente afectada por errores de probabilidad independientemente de sus distancias con el umbral.

Capítulo 6

Minería de Datos Dinámica

Definimos a este tipo o rama de la minería como la habilidad de un modelo (algoritmo) para capturar la dinámica temporal del fenómeno, por tomar en cuenta no solo la presencia o ausencia de un evento sino también el cuándo ocurre dicho evento sobre una ventana temporal previamente definida.

Ahora hablando de minería de datos dinámica o algoritmos dinámicos, existe muy poco o casi nada al respecto de esto en la literatura, la mayoría de las publicaciones que encontramos, las cuales usan definiciones similares o funcionalidades parecidas, buscan capturar y modelar la esencia de fenómenos dinámicos [53][54][55], donde por dinámicos se entiende la naturaleza cambiante de los atributos del fenómeno y que tan rápido cambian estas y no se refieren, como en este caso el algoritmo que proponemos, a la importancia de capturar la ocurrencia o no de un evento con respecto a un intervalo de tiempo, relativo a la ocurrencia del fenómeno. Por otra parte existen trabajos sobre minería de datos temporal [56][57][58] los cuales se acercan a un más a lo propuesto en este proyecto, aunque estos análisis buscan capturar el fenómeno desde el punto de vista de series de tiempo, secuencias y variables continuas, las cuales dependen del tiempo, a diferencia de lo propuesto en este proyecto donde se busca capturar la dinámica de los eventos importantes del fenómeno e incorporarlos con las variables no temporales (estáticas), con el objetivo de crear modelos más robustos y con mejor seguimiento de la naturaleza dinámica del fenómeno y por ende un mejor desempeño del modelo a la hora de clasificar o predecir.

Un elemento novedoso de este algoritmo, es el concepto de no solo capturar el efecto de cuando el evento ocurre o se presenta, sino también cuantificar el efecto del paso del tiempo sin que este evento se presente u ocurra, lo cual resulta ser sorprendentemente cuantificable y contribuye a la buena clasificación del modelo y por lo mismo a su desempeño.

6.1. Metodología

El objetivo del algoritmo dinámico es capturar la dinámica temporal de las variables que pueden clasificarse como eventos o intervenciones, con la cual se puede dotar a los modelos de dinámica con respecto del tiempo, es decir, los modelos sean cambiantes y adaptables en sus predicciones (o clasificaciones), acorde con los eventos que afectan al fenómeno de estudio con el correr del tiempo.

El primer paso es identificar las variables más importantes potencialmente dinámicas, esto es, variables tipo evento, intervención y todas aquellas que puedan tener un efecto distinto en el fenómeno más allá de solo su ocurrencia, si no también por el cuándo ocurren, después codificar estas para que puedan ser interpretadas y codificadas como dinámicas,

la idea propuesta en este trabajo para dotar a un atributo de dinámica, es mediante el uso de esquemas, los cuales son construidos utilizando ventanas de tiempo y la presencia o ausencia del evento o intervención, los cuales cambiaran su valor dependiendo de la ventana de tiempo en la que se encuentren como se muestra en la figura 6.1.

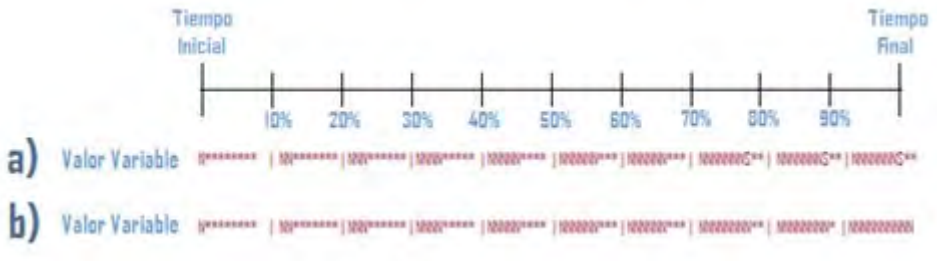


Figura 6.1: Esquemas creados para representar variables dinámicas

Como se puede observar en la figura 6.1, los atributos dinámicas son creado utilizando esquemas, los cuales estarán formados por ventanas, las ventanas son formadas por dividir el periodo de tiempo que hay entre un tiempo inicial y uno final, este periodo de tiempo se divide en partes iguales, las cuales pueden ser variables (en la figura el periodo es dividido en 10 ventanas iguales) en su número, dependiendo de la resolución que se requiera tener para la variable es el número de ventanas a crear, entre más ventas mayor resolución al capturar el efecto dinámico. El periodo de tiempo inicial y final es definido por observación del fenómeno mismo, por ejemplo si es para enfermedades el periodo de tiempo serán los años definidos en la literatura para la evolución de dicha enfermedad o para padecerla después de determinados factores, si estamos hablando de escuelas, el periodo de tiempo sería un año (lo que dura el año escolar) y las ventanas serían los meses del ciclo escolar, si hablamos de un fenómeno descrito con periodicidad mensual las ventanas serian semanas o días, es decir no hay una formula para formar dichas ventanas, estas son variables y deben ser definidas acorde con la resolución que se desea dar a la variable y con el fenómeno a estudiar, por lo cual se deben tener muy claro la periodicidad del fenómeno de estudio.

La variable se formará de la siguiente forma: para cada ventana existirá un valor, si el evento ocurre en el periodo de tiempo correspondiente a la ventana, el valor asignado será S (Si), por el contrario si el evento no ocurre el valor asignado será N (No), las ventanas continuaran sucediendo con el transcurso del tiempo, y para cada ventana la pregunta será la misma ¿el evento ocurrió?, mientras la respuesta sea No se continuara llenando las ventanas con el valor N, una vez que el evento ocurre se completara con * el resto de las ventanas, ya que solo importa que el evento ocurra y será cuando la variable este completa.

Por ejemplo sea una variable dinámica dividida en 5 ventanas y el evento ocurre hasta la 3ra ventana, la codificación de la variable en los 5 tiempos es:

- Tiempo 1: N*****
- Tiempo 2: NN***
- Tiempo 3: NNS**
- Tiempo 4: NNS**
- Tiempo 5: NNS**

Por el contrario, si el evento no ocurre nunca durante ese periodo la codificación es de la siguiente forma:

- Tiempo 1: N*****
- Tiempo 2: NN***
- Tiempo 3: NNN**
- Tiempo 4: NNNN*
- Tiempo 5: NNNNN

Cada posible combinación tendrá un valor de contribución, incluidos los esquemas donde el evento no ha ocurrido aun, esto dota de dinámica el modelo, ya que con el avance del tiempo para cada ventana el modelo cambiara debido a que las contribuciones de las variables cambian, incluso cuando el evento no ha ocurrido aun.

6.2. Funcionamiento del Algoritmo

El algoritmo propuesto en este trabajo para capturar la dinámica temporal de los fenómenos a modelar trabaja de la siguiente forma:

1. Definición de atributos dinámicos: Puede ser tal vez el paso más importante para el buen funcionamiento de este concepto y la obtención de un modelo superior en desempeño (comparado con un modelo tradicional), primero debemos comprender el fenómeno y definir si un determinado evento es relevante por solo su ocurrencia o también por el cuando ocurre, cuando hablamos del último caso entonces este atributo puede ser considerado como dinámico.

2. Agregar el componente de temporalidad a los atributos elegidos: Si el atributo aun no tiene un valor que puede ser ubicado temporalmente debemos asignar o definir la forma de hacerlo, sin esto no es posible codificarlo de forma dinámica.

3. Definir número de ventanas a utilizar: Esto es en cuantos bloques decido codificar la información dinámica, esto muchas veces depende de la periodicidad del fenómeno y que tan rápido puede renovarse este.

4. Definición de archivos de entrenamiento y prueba: Al igual que en cualquier creación de un modelo los datos se dividen en entrenamiento y prueba, la selección es de forma aleatoria, nosotros usamos 70 % y 30 % respectivamente.

5. Entrenamiento del modelo dinámico: Este tipo de modelo se crea y se aplica con dos componentes, cuenta con un componente estático y otro dinámico, el estático puede ser entrenado de forma tradicional, separando los atributos que fueron elegidos como dinámicos y entrenar el resto de forma separada usando el algoritmo tradicional, el entrenamiento dinámico tiene su particularidad para entrenar y se hace de la siguiente forma:

    - Definición de intervalos: Acorde con el número de ventanas y el periodo de estudio del fenómeno definimos los intervalos correspondientes a cada ventana.
    - Codificación de atributos dinámicos: Con la definición de los intervalos se puede ahora proceder con la codificación de los atributos dinámicos, esto se hace ventana por ventana, en la primera ventana se revisa para el primer atributo dinámico, si el evento sucedió o no en ese intervalo de tiempo, si es así se coloca

el valor S en la primera ventana de este atributo, si no, se coloca N, esto se realiza para cada uno de los atributos dinámicos definidos y para cada registro, el resto de las ventanas se llenan con el valor * (no importa).

- Entrenamiento: Una vez codificados los atributos se procede a entrenar el modelo de forma normal o estándar, se aplica el mismo procedimiento que se aplicaría a un modelo estático, con lo cual se obtienen los valores de épsilon y score para la primera ventana de tiempo de cada uno de los atributos dinámicos.
- Iteración de entrenamiento por ventana: Se realiza la misma operación de codificación para la siguiente ventana y el mismo proceso de entrenamiento, así hasta cubrir las N ventanas para tener los valores de todos los atributos en cada una de las ventanas acorde a su ocurrencia o no en una o varias de las ventanas.

6. Prueba del modelo dinámico: Al igual que en el caso de entrenamiento este tipo de modelo se mide por la suma de sus dos componentes, estática y dinámica, la parte estatica es aplicada una sola vez al conjunto de prueba, pero la parte dinámica funciona de la siguiente forma:

   - Codificación de atributos dinámicos: Se codifican las atributos en la primera ventana para el archivo de prueba, al igual que se hace con el proceso de entrenamiento.
   - Aplicación de score: Se aplican los scores correspondientes para la ventana 1 obtenidos en el proceso de entrenamiento, se suman los scores de todos los atributos dinámicos y también los estáticos.
   - Desempeño: Se puede medir en cada ventana el desempeño relativo del modelo, en cada ventana tendremos una medición del actuar de nuestro modelo, lo cual indica la dinámica cambiante y adaptación de nuestro modelo a los eventos temporales que se presenten.

7. Resultados: Después de la medición de los desempeños podemos decir si nuestro modelo es lo suficientemente bueno o debemos re-definir las variables elegidas, al igual podemos saber de manera individual si la parte estática o dinámica son buenas o no, para así ajustar sobre una o ambas.

6.2.1. Pseudocódigo

A continuación se presenta el pseudocódigo del funcionamiento del algoritmo de modelación dinámica:

Función Modelación Dinámica():

{

Parámetros de configuración (entrada)

N = # de ventanas de partición
Arr_Var = Arreglo con los atributos dinámicos
Mat_Datos = Totalidad del conjunto de datos para crear modelo (Train)
M = tamaño intervalo de tiempo
ini = tiempo inicial
Z = Longitud(Arr_Var), (Número de atributos a convertir en dinámicos)
L = Longitud(Mat_Datos), (Número de registros en la muestra)

For i = 1 hasta N, (codificación para cada una de las N ventanas)

```
{
    For  a = 1 hasta L, (codificación atributos dinámicos para cada
        registro
    {
        For  j = 1 hasta Z, (codificación para cada atributo)
        {
            IF(Mat_Datos[a, j] no contiene S) (revisa que el evento no haya
                ocurrido ya para el registro y atributo en turno)
            {
                IF(Evento en Mat_Datos[a, j] ocurrió = verdadero)
                    Mat_Datos[a, j] = S,
                    (Se pone el valor de si ocurrió el evento en la ventana)

                ELSE
                    Mat_Datos[a, j] = N,
                    (Se pone el valor de no ocurrió el evento en la ventana)

                For k = i + 1 hasta N, (completa codificación ventanas
                    no vistas)

                    Mat_Datos[a, j] = *,
                    (Pone el valor * para el resto de las ventanas no recorridas)
            }
        }
    }

    Resultado = Entrenamiento(Mat_Datos, Arr_Var),
        (Calcula los scores de cada atributo dinámico para ventana i)

    Scores_dinamicos[i] = Resultado,
        (Guarda el calculo de los scores dinámicos para ventana i)
}

return Scores_dinamicos[],
    (Devuelve el arreglo de atributos dinámicos y sus scores para
    cada una de las N ventanas)
}
```

Posteriormente se puede probar que tan bueno fue el modelo creado utilizando atributos dinámicos y estáticos, utilizando el siguiente pseudocódigo:

Función Prueba Modelo Dinámico():

```
{
  Parámetros de configuración (entrada)
    N = # de ventanas de partición
    Arr_Var = Arreglo con los atributos dinámicos
    Mat_Datos = Totalidad del conjunto de datos para probar modelo (Test)
    Z = Longitud(Arr_Var), (Número de atributos a convertir en dinámicos)
    L = Longitud(Mat_Datos), (Número de registros en la muestra)
```

```
For  i = 1 hasta N, (codificación para cada una de las N ventanas)
{
    For  a = 1 hasta L, (codificación atributos dinámicos para cada
        registro
    {
        For  j = 1 hasta Z, (codificación para cada atributo)
        {
            IF(Mat_Datos[a, j] no contiene S) (revisa que el evento no haya
                ocurrido ya para el registro y atributo en turno)
            {
                IF(Evento en Mat_Datos[a, j] ocurrió = verdadero)
                    Mat_Datos[a, j] = S,
                    (Se pone el valor de si ocurrió el evento en la ventana)

                ELSE
                    Mat_Datos[a, j] = N,
                    (Se pone el valor de no ocurrió el evento en la ventana)

                For k = i + 1 hasta N, (completa codificación ventanas
                    no vistas)

                    Mat_Datos[a, j] = *,
                    (Pone el valor * para el resto de las ventanas no recorridas)
            }
        }
    }

    Scoring_estatico = Pueba(Mat_Datos, Arr_Var, Scores_estaticos),
        (Aplica el modelo estático para ventana i)

    Scoring_dinámico = Pueba(Mat_Datos, Arr_Var, Scores_dinamicos),
        (Aplica el modelo dinámico para ventana i)

    Resultado[i] = Desempeño(Scoring_estatico, Scoring_dinamcio),
        (Guarda el calculo de desempeño para ventana i)
}

return Resultado[],
    (Devuelve el arreglo de desempeños para
    cada una de las N ventanas)
}
```

6.3. Resultados

En este caso para este tipo de algoritmo no pudo ser aplicado a la base de datos de ENCOPREVENIMSS 2006, por que no encontramos atributos tipo evento, ninguna de las preguntas hechas durante la encuesta tienen el componente temporal y tampoco se pueden ubicar en un intervalo de tiempo especifico. Igualmente resulto casi imposible encontrar datos médicos relacionados a diabetes u obesidad que tuvieran este tipo de atributos

ubicados en intervalos de tiempo, aun así, para el caso de los datos de DxCG 97-99 si pudimos encontrar atributos, los cuales pueden ser transformados a atributos dinámicos y, por lo tanto, aplicar y crear un modelo basado en este algoritmo.

6.3.1. DxCG 97-99

En los datos descritos y analizados en cap. 3.6.2 encontramos 4 atributos relacionados con costos, los cuales pueden ser convertidos en atributos dinámicos para poder aplicar un proceso de modelado dinámico como el descrito en este capítulo, los atributos son: Total (costos totales), Income (costos generados por hospitalización de los pacientes), outcome (costos generados por consultas medicas ambulatorias proporcionadas a los pacientes) y drugs (costos generados por medicinas surtidas a los pacientes), todos estos atributos pueden ser ubicados con periodicidad de 3 meses a lo largo del año, es decir podemos tener 4 ventanas de tiempo a lo largo del año conformadas cada una por 3 meses de longitud (Enero- Marzo, Abril - Junio, Julio - Septiembre, Octubre - Diciembre).

La codificación de los atributos en este caso fue poner no (N) en la ventana para los casos en los cuales el paciente no se encontrara en el top 5 % de gastos para ese trimestre del año y poner si (S) cuando por el contrario si se encontrara.

Los resultados obtenidos muestran una diferencia significativa entre la forma de entrenamiento dinámica y la estática, como se muestra en la figura 6.2, para el atributo dinámico "Total"podemos ver las contribución del score para cada una de las 4 ventanas definidas, la figura 6.2 a) muestra una serie de 4 eventos en los que el paciente no (N) se encuentra en el top 5 % con más gastos, la tendencia de sucesos consecutivos de N (no estar en el top 5 % de gastos totales), muestra descensos progresivos en la contribución al score en el atributo, por el contrario de la forma estática para todo el tiempo (4 ventanas) el valor seria invariante, hay 2 ventajas importantes en este tipo de comportamiento que son visibles al revisar los gráficos entre el algoritmo estático vs dinámico; 1) el progreso temporal de la predicción: mientras que utilizando los atributos de forma estática nuestras predicciones no tendrían ninguna variación a lo largo del año, al utilizar los atributos de forma dinámica podemos ver una tendencia progresiva en las contribuciones al score, por cada una de las ventanas, la secuencia de N (no estar en top 5 % consecutivos) o S (si estar en el top 5 % consecutivos) hacen variar nuestras predicciones en cada una de las ventanas y ser más cercano a la realidad temporal del fenómeno, esto es, mientras un algoritmo estático nos llevaría a una sobre o sub predicción en nuestras clasificaciones, dependiendo de N o S respectivamente, sobre todo en los primeros 3 trimestres del año, un algoritmo dinámico seria más apegado a las evidencias capturadas en cada trimestre del año y nos llevaría a una mejor estimación de las predicciones de nuestro clasificador; 2) mejor descriminación: como se puede observar para el atributo "Total" en la figura 6.2 a) y b) al final, 4 ventana, la consecusión de N o S (no estar o si estar en al top 5 % de gastos) tiene una contribución de score superior al atributo estático, negativamente o positivamente respectivamente figura 6.2, lo cual es lógico, si lo pensamos dado que estamos clasificando al top 5 % de los pacientes más costos una serie de N consecutivos para cada trimestre del año debería llevarnos a afirmar que el paciente no estará en el top 5 % definitivamente, lo cual se representa con una contribución al score muy negativa, esto se logra mejor con el algoritmo dinámico, también con el caso de S consecutivos se logra un score muy positivo para afirmar que el paciente si estará en top 5 %, todo esto ofrece una mejor descriminación al final del año.

La figura 6.2 también muestra el resto de los atributos que fueron considerados como dinámicos (income, outcome, drugs), en todos se observa un comportamiento similar al

ya analizado atributo "Total", con la diferencia entre las contribuciones al score, en "Total"son mas grandes al tratarse de la suma de las otras tres, también podemos observa que el último trimestre parece ser el más definitivo, dado que el cambio en las contribuciones al score es más fuerte en ese trimestre que en los primeros 3, también observamos que el atributo "Outcome" contribuye más a la dinámica del modelo, lo cual es lógico al considerarlo, entre más visitas hay a consulta por parte de los pacientes indica una sintomatología de deterioro, por lo cual a lo largo del tiempo serán más frecuentes y posiblemente de mayor consumo de medicamentos y la inevitable hospitalización.

Aparte de las secuencias de N y S, hay variaciones en las que se combinan N y S, las cuales se muestran en la figura 6.3 para las variable "Total", en estas también se pueden observar las variaciones en las contribuciones al score, en estos gráficos se pueden observar las mismas características ya descritas, sobre todo resalta la parte ya mencionada de que el último trimestre del año parece tener un efecto definitivo, lo cual al pensarlo parece también lógico, como todo lo encontrado en esta modelación, alguien que decae en su salud al final del año es muy probable que para el siguiente año se encontrara entre el top 5 % de los más costosos, dado que muy probablemente este paciente pasara gran parte del año enfermo, además podemos ver en los gráficos de la figura 6.3 que la variación entre los trimestres entre S y N da como resultado contribuciones al score intermedias entre los scores estáticos de S y N, lo cual habla de nuevo de una mejor aproximación y discriminación ventana a ventana de forma dinámica.

El resto de los gráficos para los atributos dinámicos outcome, income y drugs es mostrado en el apéndice B, los cuales en comportamiento son similares el mostrado para el atributo "Total" en la figura 6.3, con las variaciones correspondientes en las magnitudes de las contribuciones de los scores para cada atributo respectivamente. Podemos también observar una variabilidad entre lo positivo y negativo de los scores, al hacerlo de forma dinámica comparado con lo estático, lo cual nos hace comprender de mejor forma por que el dinámico tiene mejor discriminación temporal en cada trimestre del año y al final.

Con respecto a la discriminación del modelo dinámico vs el estático la podemos mostrar utilizando solo los 3 atributos dinámicos como son: Income, Outcome y Drugs aplicando el modelo ventana a ventana podemos ver la evolución de la probabilidad en cada una de las mismas, esto se muestra en la figura 6.4, en la cual se puede ver como ventana a ventana el progreso de la discriminación en probabilidad va en aumento y al final comparando con la discriminación obtenida por el modelo estático podemos ver que esta es superior.

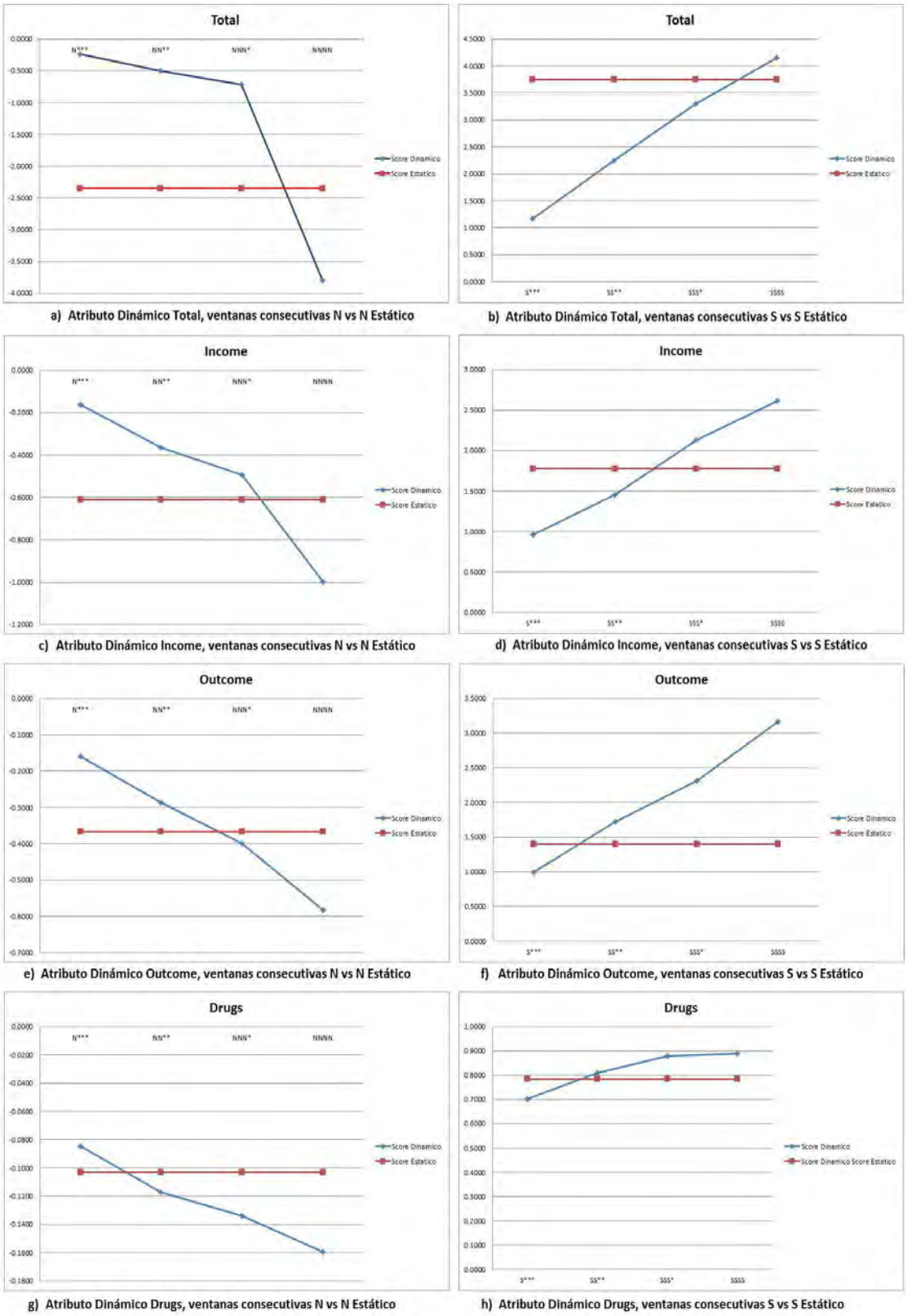


Figura 6.2: Atributos dinámicos tipo evento codificados a través de los 4 trimestres con S y N si ocurre o no ocurre el evento y la comparación de los scores obtenidos vs el score del mismo atributo codificado de forma estática.

Figura 6.3: Atributo dinámico Total tipo evento codificado a través de los 4 trimestres con combinaciones S y N, comparación de los scores obtenidos vs el score del mismo atributo codificado de forma estática.

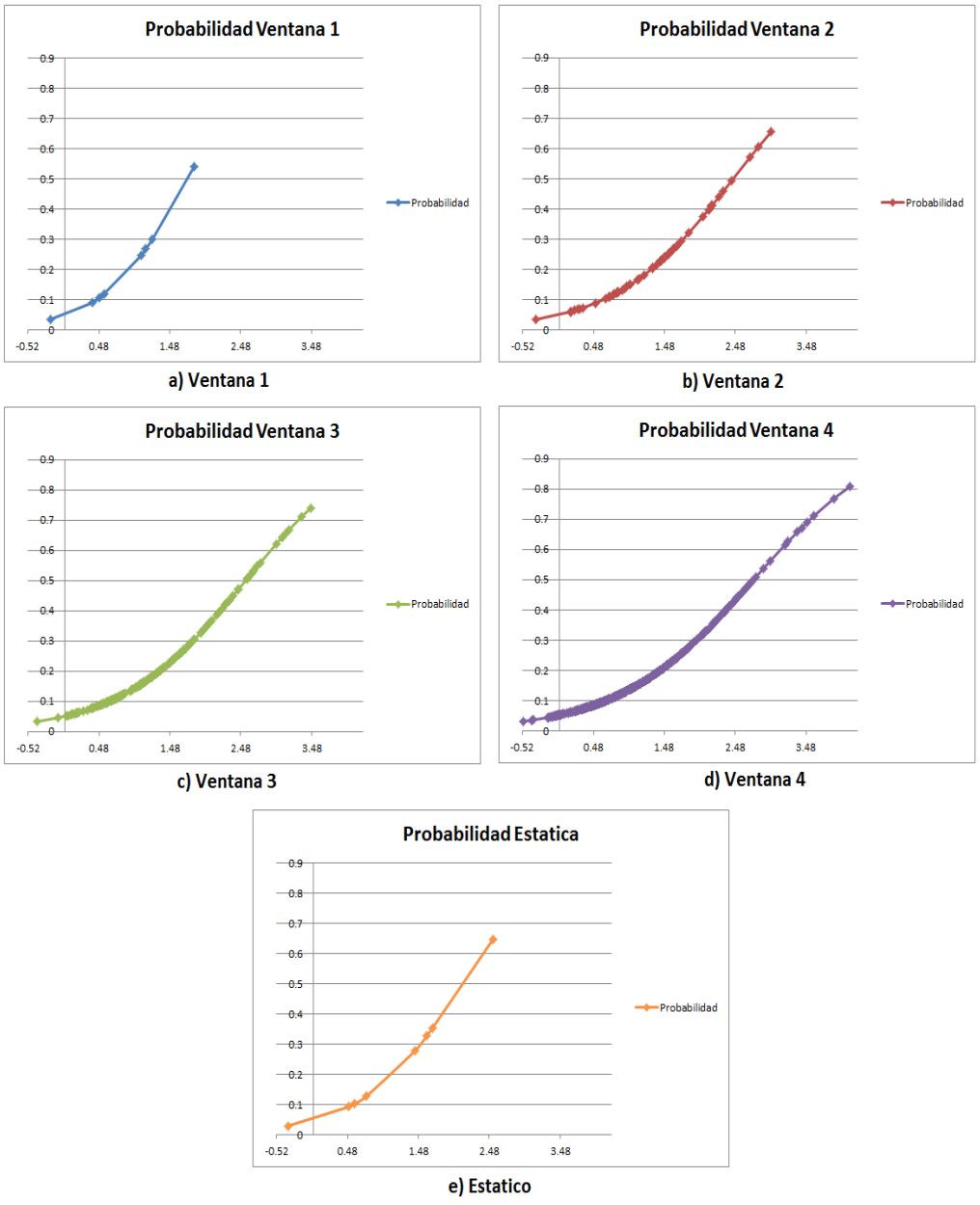


Figura 6.4: Evolución de la discriminación en probabilidad ventana a ventana del modelo dinámico vs estático.

Capítulo 7

Minería de Datos Adaptativa

Con respecto del algoritmo adaptativo, se puede interpretar en la literatura como actualización dinmica del modelo o actualización en tiempo real (online) del modelo o simplemente actualización del modelo, no existen muchos trabajos respecto a esto, puede considerarse algo simple desde el punto de vista de la investigación, bastaría con re-entrenar o volver a construir el modelo cada que se requiera, pero desde el punto de vista computacional poder lograr esto de forma automática y sin intervenir el software para hacer la actualización es de suma importancia, dado que permite crear software para el usuario final más duradero, confiable y atractivo. Desde el punto de vista de minera de datos adaptativa existen algunos trabajos[59][60][61], los cuales se refieren a lidiar en tiempo real con grandes volúmenes de datos de entrada y poder hacer predicciones que den al usuario la impresión de ser en tiempo real, además de adaptarse a la velocidad y demanda de esos datos. En este proyecto se hace algo similar pero en diferente concepto, adaptarse a nueva información, pero no porque esta entre de forma continua y en grandes volúmenes al sistema, sino por la importancia y relevancia que tenga para producir un cambio en el fenómeno, es decir, si factores los cuales no eran importantes después de un tiempo se convierten en importantes, se debe cuantificar e incorporar al modelo esta aportación y, por el contrario, factores que eran importantes ahora ya no lo son entonces deben de igual forma ser cuantificados y actualizar su aportación actual al modelo.

El elemento innovador es capturar este efecto para incluir nuevas características, las cuales se vuelvan importantes, como pueden ser intervenciones y quitar aquellas cuya aportación al modelo deja de ser importante, estas pueden ser características las cuales han sido combatidas efectivamente y después ya no son un factor de riesgo para la enfermedad por ejemplo, además de no acumular historia si no al contrario tener una ventana de control, donde la información más antigua sale y se inserta la más actual, para dar al modelo y a las nuevas variables la oportunidad de reflejar esos cambios en el ambiente de forma casi inmediata y no tener que esperar tanto para ver el efecto de nuevas características, esto por el simple hecho de no contar con una muestra suficientemente grande para equiparar el efecto de la histórica, lo cual logramos superar por eliminar la información antigua al mismo ritmo que agregamos nueva. Esto es: si se incierta un año de nueva información al modelo, se elimina un año de la historia más antigua, con lo cual se puede eliminar el efecto del muestreo, lo que permitirá ver reflejados los cambios actuales en el fenómeno, más temprano que tarde en sus predicciones.

7.1. Metodología

El algoritmo adaptativo dotara al modelo de la habilidad para adaptarse a nuevas circunstancias a un cambio en el comportamiento del fenómeno y al surgimiento de nuevos factores que lo afectan. Al igual que en el algoritmo del capítulo anterior esto favorece al modelo para ser dinámico, una vez que los datos de 2010 ingresan a la BD el modelo se actualizara nuevamente utilizando la información de 2000 a 2010 descartando los datos de 1999 y 1998.

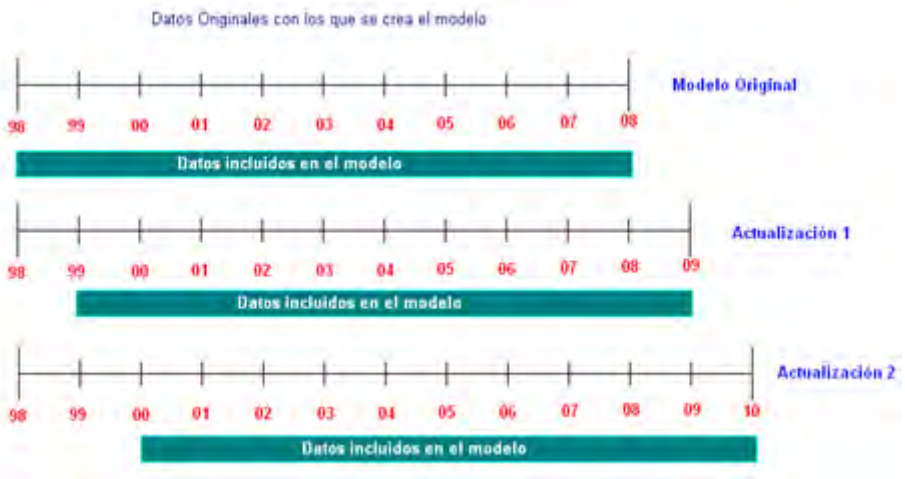


Figura 7.1: Algoritmo Adaptativo, ejemplo a 10 años.

La figura 7.1 describe la forma de implementar el algoritmo adaptativo, esto es, a un periodo de tiempo para el cual se vence la valides del modelo, al igual que en el caso del algoritmo dinámico, la definición de este periodo depende del fenómeno, el algoritmo es muy simple por cada nuevo conjunto de datos este actualizará las contribuciones al score de cada variable sin tomar en cuenta el último conjunto de datos históricos, por ejemplo en la figura 7.1 tenemos un periodo de vencimiento del modelo de 10 años, esto es de 1998 a 2008, una vez que nuevos datos arriban y están completos para 2009, el modelo se actualizará utilizando la información de 1999 a 2009 descartando la información histórica de 1998, y as sucesivamente una vez que los datos de 2010 ingresan a la BD el modelo se actualizará nuevamente utilizando la información de 2000 a 2010 descartando los datos de 1999 y 1998.

Esta actualización es sencilla ya que el clasificador NB está basado únicamente en conteos, por lo cual la actualización consiste en re-escribir estos conteos y re- calcular las contribuciones para cada variable, el cálculo de esta contribución es:

$$\mathbf{Score} = \ln\left(\frac{\frac{\mathbf{N_{xc}}}{\mathbf{N_c}}}{\frac{\mathbf{N_{x\sim c}}}{\mathbf{N_{\sim c}}}}\right) \tag{7.1}$$

Donde:

- Nxc es el número de elementos con la variable x y que pertenecen a la clase.
- Nc es el número de elementos en la clase.

- Nx$\sim$ c es el número de elementos con la variable x y que no pertenecen a la clase.
- N$\sim$ c es el número de elementos que no pertenecen a la clase.

7.2. Funcionamiento del Algoritmo

Este algoritmo propuesto para adaptarse a la circunstancias cambiantes de los fenómenos con el pasar del tiempo o la evolución de los mismos trabaja de la siguiente forma:

1. Definición de periodicidad del fenómeno: Es un paso muy importante el definir que tan periódicamente este fenómeno debe ser actualizado, lo cual depende de si en ese periodo puede sufrir un cambio significativo en su comportamiento, también dependemos de la periodicidad de la información que vamos a recibir, combinando ambos conocimientos podemos definir un periodo adecuado de actualización (adaptabilidad) para nuestro fenómeno.
2. Recorrido de adaptación: Con el periodo definido podemos ahora enfrentar la llegada de nuevos datos, acorde al periodo de actualización descartamos el periodo más viejo de datos disponibles e integramos el periodo actual más reciente recibido por el algoritmo.
3. Actualización de conteos: con los nuevos datos se procede a recontar cada uno de los números requeridos en la ecuación 7.1, descontando aquellas contribuciones del viejo periodo y agregando la contribución del nuevo periodo.
4. Actualización de scores: Con los nuevos conteos se puede recalcular el score utilizando la ecuación 7.1 y a su vez con las nuevas aportaciones individuales de score se puede actualizar el score total de cada uno de los registros del modelo.

7.2.1. Pseudocódigo

A continuación se presenta el pseudocódigo del funcionamiento del algoritmo adaptativo:

Función Modelo Adaptativo():

{
Parámetros de configuración (entrada)
N = Periodo
Arr_Var = Arreglo con atributos modelo
Mat_Datos = Nueva información
Datos_totales = Datos totales actuales del modelo
Z = Longitud(Arr_Var), (Número de atributos a convertir en dinámicos)

For i = 1 hasta Z, (conteos para cada uno de los atributos del modelo)
{
Conteos[i] = Restar(Datos_totales,Periodo_eliminar),
(calcula la contribución del periodo a eliminar)

New_Conteos[i] = Contar(Mat_Datos),
(calcula la contribución del periodo a agregar)

```
            Nxc[i] = Nxc[i] + New_Conteos[i].Nxc - Conteos[i].Nxc,
            (actualiza los conteos para la variable Nxc)

            Nx∼c[i] = Nx∼c[i]. + New_Conteos[i].Nx∼c - Conteos[i].Nx∼c,
            (actualiza los conteos para la variable Nx∼c)

            Nc[i] = Nc[i] + New_Conteos[i].Nc - Conteos[i].Nc,
            (actualiza los conteos para la variable Nc)

            N∼c[i] = N∼c[i] + New_Conteos[i].N∼c - Conteos[i].N∼c,
            (actualiza los conteos para la variable N∼c)

            actualizar(Nxc,Nx∼c,Nc,N∼c,i),
            (actualiza las contribuciones al score para el atributo i-esimo)
        }
    }
```

7.3. Resultados

Para aplicar este algoritmo y mostrar los resultados no fue posible utilizar los datos que se han venido utilizando desde el principio de este proyecto (ENCOPREVENIMSS 2006, DxCG 97-99, Datos de Ervine Universidad de California), debido a que no son temporales, no hay forma de conseguir la misma información para distintos periodos de tiempo o continua a través de los años de tal forma que se pudiera aplicar el algoritmo descrito en este capítulo, sin embargo fue posible conseguir información relacionada con salud. esto fue gracias a las encuestas aplicadas por el Sistema Nacional de Encuestas de Salud las cuales son llamadas ENSANUT (Encuesta Nacional de Salud y Nutrición), en las cuales se pudieron obtener los datos recopilados en las encuestas de 2006 y 2012, los cual no da suficiente información para probar el algoritmo presentado en este capítulo.

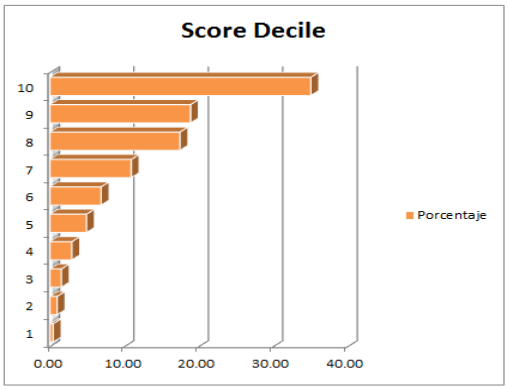
7.3.1. ENSANUT 2006 y 2012

La encuestas nacionales de salud y nutrición (ENSANUT) son aplicadas con el propósito de recaudar información en lo general sobre la salud, cobertura de programas y las condiciones en lo general de salud y nutrición de la población mexicana, con el objetivo de diagnosticar y planear estrategias futuras para la atención de las necesidades en esta materia por parte de la población. Las encuestas son muy parecidas en información a las hechas por ENCOPREVENIMSS 2006, por lo cual es posible generar un modelo similar al mostrado en cap. 3.6.1 y utilizando el mismo objetivo la predicción de diabetes en la población entrevistada, utilizando los datos recavados por estas para crear el modelo descrito.

Se creo un modelo utilizando los datos de la ENSANUT 2006, para predecir diabetes, utilizando las respuestas obtenidas en cada una de las preguntas hechas en la encuesta como atributos o predictores para el modelo. Se obtuvieron los resultados en desempeo mostrados en la figura 7.2, podemos observar que existe una buena discriminación en el modelo creado al igual que en los datos de ENCOPREVENIMSS 2006, lo cual además de demostrar que es posible modelar la diabetes en los datos de ensanut también nos permite concluir que los perfiles encontrados en la modelación para ENCOPREVENIMSS 2006,

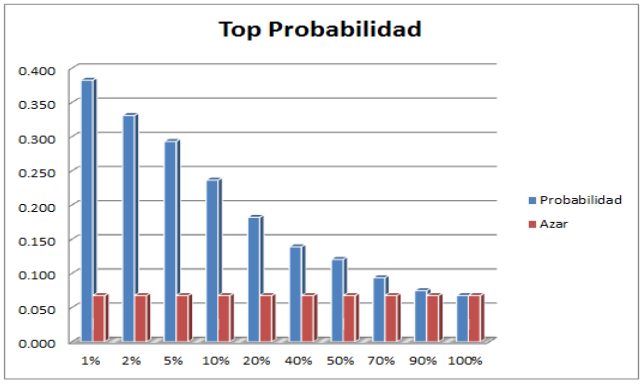
lo cuales son una muestra de los derecho habientes de esa institución en la Ciudad de México, son muy similares a los encontrados en la población general de todas las entidades del país. El modelo para ENSANUT fue creado para todos los adultos de 20 a 99 sin diferencias de género o adultos mayores (de 60 años en adelante) como en el caso del modelo ENCOPREVENIMSS.

Decile	Aciertos	Porcentaje
10	321	35.04
9	173	18.89
8	160	17.47
7	100	10.92
6	63	6.88
5	45	4.91
4	27	2.95
3	14	1.53
2	9	0.98
1	4	0.44



a) Desempeño Score Decile

Top %	Aciertos	Probabilidad	Azar
1%	52	0.382	0.068
2%	90	0.331	0.068
5%	199	0.293	0.068
10%	321	0.236	0.068
20%	494	0.182	0.068
40%	754	0.139	0.068
50%	817	0.120	0.068
70%	889	0.094	0.068
90%	912	0.075	0.068
100%	916	0.068	0.068



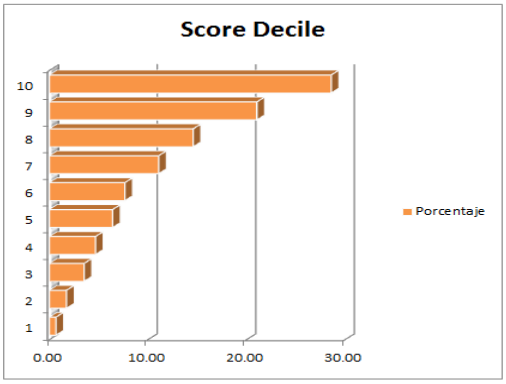
b) Desempeño Top Probabilidad

Figura 7.2: Desempeño del modelo ENSANUT 2006 para clasificación de diabetes en adultos de 20 a 99 años

Posteriormente se aplico el modelo creado para los datos de ENSANUT 2006 a los datos recavados en la ENSANUT 2012, como lo platicamos en este capítulo una de las primeras cosas a las que nos enfrentamos es el cambio de las preguntas hechas entre la encuesta de 2006 y la de 2012, esto debido a que de acuerdo a los análisis y resultados obtenidos en la ENSANUT 2006, el Sistema Nacional de Encuestas de Salud decide realizar ajustes y preguntar algunas cosas distintas en la ENSANUT 2012, esto resulta en que el modelo original de ENSANUT 2006 fue creado utilizado 99 atributos (preguntas de la encuesta) como predictores del modelo, pero en ENSANUT 2012 solo se encontraron 50 atributos (preguntas de la encuesta) iguales a los utilizados en 2006 aquí podemos observar una perdida de atributos de casi el 50 %, lo cual es una perdida importante de información para el modelo y nos habla de la importancia de mantener actualizados los modelos, en este caso en un periodo de 6 años las preguntas e información recabada por las encuestas cambio en un 50 % por considerarse ya inapropiada u obsoleta, dado que todos los fenómenos tienden a cambiar u evolucionar con el paso del tiempo como lo mencionamos en la introducción de este capítulo y ahora lo podemos observa en datos reales y sobre todo en un fenómeno de salud publica.

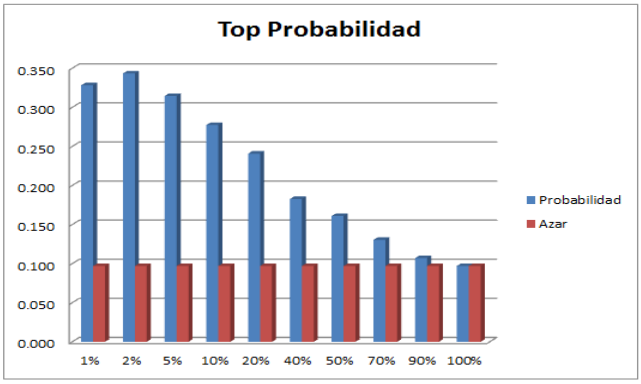
El resultado en desempeño de aplicar el modelo creado para datos 2006 en información obtenida en 2012 se muestra en la figura 7.3, en esta podemos observar que se mantiene un buen desempeño, es decir, mantenemos buena predicción para el modelo en los primeros deciles a pesar de la perdida del 50 % de los atributos, pero aun así vemos una disminución en el desempeño comparado con el modelo original, por ejemplo en score decile para el decile 10 existe una disminución del 18 % en el porcentaje de aciertos obtenidos en el modelo original y tomando los 3 primeros deciles se pasa del 71.4 % a 64.2 % de aciertos, pero la parte importante que podemos observar es el crecimiento de la tasa de diabetes (probabilidad de azar), en 2006 era del 6.8 % de la población encuestada y para 2012 la tasa crece a 9.7 % de la población encuestada, esto es un crecimiento del 42 % en la tasa de incidencia de diabetes en la población mexicana en tan solo 6 años, esto nos da una muestra mas de la importancia de la actualización de los modelos, observando un fenómeno real como el de la salud publica en México, con datos reales recavados por encuestas, podemos ver como los fenómenos son cambiantes con paso del tiempo y deben ser ajustados para mantener su buena predictibilidad, así como su adaptación a un entorno cambiante como son los fenómenos de las enfermedades emergentes.

Decile	Aciertos	Porcentaje
10	1284	28.60
9	945	21.05
8	655	14.59
7	498	11.09
6	344	7.66
5	288	6.41
4	210	4.68
3	158	3.52
2	78	1.74
1	30	0.67



a) Desempeño Score Decile

Top %	Aciertos	Probabilidad	Azar
1%	152	0.328	0.097
2%	318	0.343	0.097
5%	728	0.315	0.097
10%	1284	0.277	0.097
20%	2229	0.241	0.097
40%	3382	0.183	0.097
50%	3726	0.161	0.097
70%	4224	0.130	0.097
90%	4460	0.107	0.097
100%	4490	0.097	0.097



b) Desempeño Top Probabilidad

Figura 7.3: Desempeño del modelo ENSANUT 2006 para clasificación de diabetes en adultos de 20 a 99 años aplicado a los datos obtenidos en ENSANUT 2012

Después haciendo un análisis con épsilon de las preguntas comunes de las encuestas de 2006 vs 2012 para detectar posibles cambios en los predictores, observamos en algunos cambios muy fuertes, en otros cambios pequeños y otros factores se mantuvieron casi idénticos en ambos periodos, a continuación analizamos algunos de estos factores sobre los

que notamos estos efectos entre los distintos periodos y se muestran los resultados en la tabla 7.1, por ejemplo la primera pregunta de análisis es, si el entrevistado tiene seguro medico y en que institución, ahí podemos ver un cambio significativo en el aumento como factor de riesgo para la diabetes de la instituciones como el IMSS y el ISSSTE, IMSS paso de ser un factor no relevante a elevar su valor de épsilon para 2012 y colocarse como un factor de riesgo importante e ISSSTE paso de ser un factor protector con una épsilon negativa a convertirse en un factor de riesgo importante también, podríamos pensar que estas instituciones empeoraron en la atención a sus derecho-habientes y fallaron a la hora de prevenir el padecimiento entre su población, pero más bien podemos considerar que es por el contrario debido a las campañas recientes de salud y detección temprana de la diabetes que estos han aumentado en número y sobre todo en estas dos instituciones, las cuales lo llevan con mayor régimen, ambas han podido aumentar la detección de casos que tal vez en 2006 no eran detectados hasta muy tarde ya cuando el paciente sufría una accidente bascular o algún otro debido a la falta de detección del padecimiento, el resto de las instituciones se mantiene casi igual a excepción del Seguro Popular, lo cual podemos considerar como sospechoso que ahora es un factor protector contra la diabetes al contrario que lo era en 2006, pero observando la muestra esto nos habla mas bien de un crecimiento en la población, que ahora cuenta con el seguro popular lo que produce un efecto de disminución, en 2006 posiblemente solo la gente con un padecimiento grave como la diabetes solicitaba el servicio o era ingresada al sistema, ahora para 2012 un mayor número de personas tienen acceso a este sistema y ya no necesariamente por estar graves o padeciendo una enfermedad como la diabetes, en la encuesta de 2006 la mayoría de las personas entrevistadas estaban afiliadas a IMSS y en 2012 esto cambia a una mayoría de afiliados al Seguro Popular, otra de las evidencias de los cambios que sufren los ambientes en los que se desarrollan los fenómenos con el transcurso del tiempo.

En cuanto a la pregunta del último grado de estudios cursado o aprobado, podemos ver un ligero aumento en el riesgo para persona que no estudiaron más allá de la primaria y al contrario una disminución para educación superior a la primaria pero en conclusión la tendencia se mantiene tal cual en 2006 como para 2012. De igual forma con la pregunta de estado conyugal podemos ver una aumento significativo como factor de riego en 2012 para las personas casadas o viudas, en comparación con 2006 y disminución igualmente significativa para personas en unión libre o solteros, esto también puede intrínsecamente contener un factor de edad, sabemos que la mayoría de las personas mayores son mas propicios a tener diabetes. Un cambio drástico lo podemos ver en la pregunta acudió a detección de hipertensión, en 2006 era un factor de riesgo para quien contestaba que “Si” a esta pregunta, pero para 2012 se convierte en un factor protector, esto puede significar que en el 2006 la gente solo acudía al doctor cuando ya estaban enfermas y para 2012 con las campañas de prevención, tal vez más personas acuden con mas regularidad a chequeos médicos con el objetivo de cuidar su salud. Por último el sexo del entrevistado es una muestra de como las condiciones para un atributo pueden mantenerse, en este caso lo observamos en la tabla 7.1 los épsilon de 2006 y 2012 son muy perecidos por no decir casi iguales, esto es muestra que también hay atributos que se mantienen como factores de riesgo o como protectores sin importar el paso del tiempo y que son característicos de los fenómenos.

Aplicando el algoritmo adaptativo a los datos de ENSANUT con una periodicidad de 6 años para ajustarlo al mismo ritmo con el que son llevadas acabo las encuestas por el Sistema Nacional de Encuestas de Salud se obtiene una actualización del modelo y de los atributos utilizados como predictores para el modelo, dando como resultado las medidas de desempeño mostradas en la figura 7.4, en el cual podemos observar una mejora con

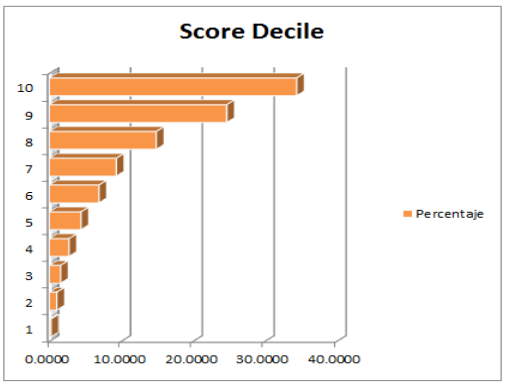
Atributo	Valor	Epsilon 06	Epsilon 12
¿ Tiene seguro medico por... ?	IMSS	-0.32	6.86
	Seguro Popular	9.54	-3.27
	ISSSTE Estatal	8.46	3.74
	ISSSTE	-3.29	9.62
	Marina o Defensa	3.82	2.06
	PEMEX	10.80	4.06
	Particular	0.41	-0.50
¿ Cuál es el último grado de educación que pasó(aprobó) en la escuela?	Ninguno	11.29	13.57
	Preescolar	1.31	2.15
	Primaria	11.59	17.16
	Secundaria	-11.43	-14.32
	Secundaria Técnica	-1.60	1.68
	Carrera Técnica	-2.96	-0.90
	Normal Básica	0.38	2.35
	Preparatoria	-8.41	-13.37
	Carrera Técnica Com.	-2.22	-2.17
	Normal Superior	-1.96	-0.15
	Licenciatura	-6.86	-8.99
	Maestría	-0.86	-2.37
	Doctorado	-0.10	-0.39
Estado Conyugal	Unión Libre	-5.26	-10.93
	Casado(a)	2.06	5.51
	Separado(a)	3.05	1.95
	Divorciado(a)	0.60	-0.18
	Viudo(a)	17.26	20.61
	Soltero(a)	-12.88	-14.23
¿ Acudió a detección de hipertensión?	Si	12.96	-2.26
	No	-7.15	-26.42
Sexo	Hombre	-4.83	-4.72
	Mujer	4.27	4.08

Tabla 7.1: Comparación de atributos recabados en la encuesta de 2006, comparados con los atributos recabados en la encuesta de 2012, utilizando épsilon como métrica

respecto al modelo original en cuanto al número de aciertos en los primeros 3 deciles (los más altos en ranqueo de score), para 2006 se obtuvo el 71.4 % de los aciertos en estos 3 primeros deciles y después de aplicar el algoritmo adaptativo incluyendo los datos de 2012 se obtuvo el 73.9 % de aciertos en los 3 primeros deciles, optando por el lado de los top porcentajes vemos que el modelo creado utilizando los datos de 2012, es superior en todos y cada uno de los top porcientos en cuanto a la probabilidad en cada sección. Y también esta de mas mencionar que la actualización del modelo utilizando datos de 2012 es superior en desempeño a utilizar el modelo de 2006 sobre los datos de 2012.

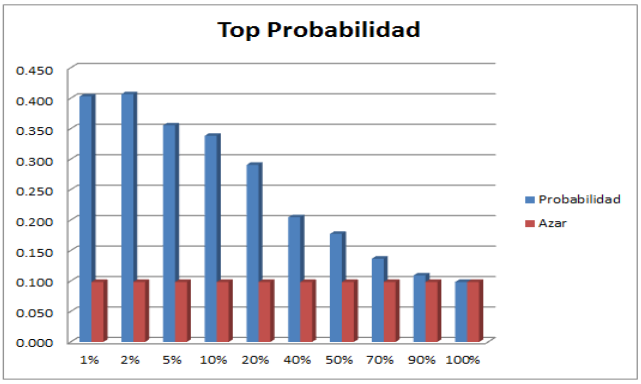
Ahora si aplicamos el algoritmo adaptativo con una periodicidad de 12 años seria a un ritmo del doble de años del modelo anterior obtenemos los desempeños mostrados en la figura7.5, podemos observar de estos que el modelo conseguido es ligeramente peor en desempeño compara con el anterior de periodicidad de 6 años, pero seria mas robusto al contar con información histórica de 12 años, por lo tanto, es necesaria una prueba mas,

Decile	Aciertos	Percentaje
10	470	34.4100
9	337	24.6700
8	203	14.8600
7	127	9.3000
6	94	6.8800
5	60	4.3900
4	37	2.7100
3	21	1.5400
2	14	1.0200
1	3	0.2200



a) Desempeño Score Decile

Top %	Aciertos	Probabilidad	Azar
1%	56	0.403	0.098
2%	113	0.407	0.098
5%	247	0.355	0.098
10%	470	0.338	0.098
20%	807	0.291	0.098
40%	1137	0.205	0.098
50%	1231	0.177	0.098
70%	1328	0.137	0.098
90%	1363	0.109	0.098
100%	1366	0.098	0.098

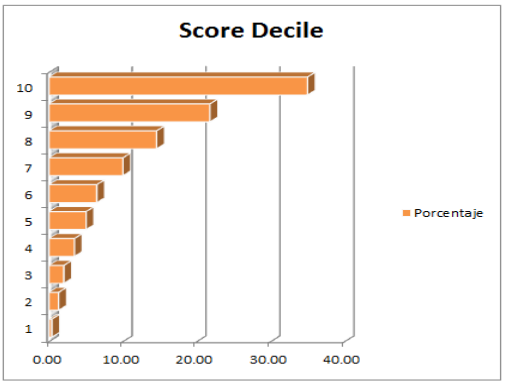


b) Desempeño Top Probabilidad

Figura 7.4: Desempeño del modelo ENSANUT aplicando algoritmo adaptativo con datos de 2006 y 2012, para clasificación diabetes en adultos de 20 a 99 años con periodicidad de 6 años

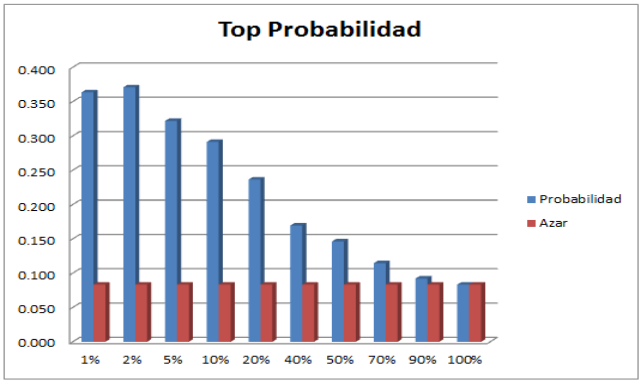
utilizando la ENSANUT de 2018, una vez finalizada, para decidir si la periodicidad del modelo debe ser de 6 o 12 años.

Decile	Aciertos	Porcentaje
10	800	34.98
9	498	21.78
8	333	14.56
7	229	10.01
6	148	6.47
5	115	5.03
4	79	3.45
3	46	2.01
2	30	1.31
1	9	0.39



a) Desempeño Score Decile

Top %	Aciertos	Probabilidad	Azar
1%	100	0.364	0.083
2%	204	0.371	0.083
5%	442	0.322	0.083
10%	800	0.291	0.083
20%	1298	0.236	0.083
40%	1860	0.169	0.083
50%	2008	0.146	0.083
70%	2202	0.115	0.083
90%	2278	0.092	0.083
100%	2287	0.083	0.083



b) Desempeño Top Probabilidad

Figura 7.5: Desempeño del modelo ENSANUT aplicando algoritmo adaptativo con datos de 2006 y 2012, para clasificación diabetes en adultos de 20 a 99 años con periodicidad de 12 años

Capítulo 8

Conclusiones

En este trabajo presentamos un breve contexto general sobre lo que es la minería de datos y los beneficios de utilizarla aplicada a las enfermedades emergentes, como apoyo a las herramientas actuales y tradicionales para el estudio de la epidemiología, además de sugerir algunos algoritmos que podrían ayudar a mejorar los desempeños obtenidos al aplicar la minería de datos al fenómeno de las enfermedades emergentes y sobre todo para la identificación de los atributos predictores de estos modelos, para poder diseñar intervenciones que apoyen y ayuden a disminuir el impacto de estos fenómenos, además de un análisis profundo de la medición de dependencia e independencia de los atributos y la perdida que esto significa al utilizar el Naive Bayes Aproximación (NBA) versus utilizar un algoritmo más complejo como el Naive Bayes Generalizado, los errores que acarrea cada asumpción de independencia que hace el algoritmo tradicional de NBA en cuanto a clasificación y ranqueo, así como la conveniencia de utilizar un algoritmo más complejo dependiendo del problema sobre el que se aplicara.

Empezamos demostrando la validez y utilidad de un algoritmo clasificador simple sencillo de utilizar y de muy buen desempeño al clasificar todo tipo de fenómenos, además de su capacidad explicativa para el diseño de intervenciones y el buen entendimiento de los factores que contribuyen como factores de riesgo o protectores para las enfermedades, por lo cual podemos concluir que el NBA es lo suficientemente bueno y robusto para crear modelos epidemiológicos básicos, simples y capaces de superar muchas de las herramientas tradicionales de la epidemiología, también demostramos la utilidad de utilizar conceptos como la correción de Laplace en su forma más simple como el llamado “ADD ONE” y la mejoría que se obtiene al utilizar una versión derivada aplicada directamente al NBA, en problemas donde el muestreo o datos obtenidos contienen caso sin ninguna evidencia para atributo/valor en particular para una de las clases de análisis o su complemento.

Presentamos una versión propia del algoritmo de selección de características ajustado al clasificador NBA y buscando la optimización del número de atributos utilizados para la creación de los modelos y mejorar el desempeño de los mismos, podemos concluir que un algoritmo de este tipo no solo ayuda computacionalmente a reducir el problema, al disminuir el número de atributos o dimensionalidad del modelo al elegir automáticamente solo aquellos que contribuyen a un buena clasificación del modelo (mejor desempeño), si no que también contribuye a obtener modelos más robustos de mejor desempeño y mayor estabilidad en pruebas de validación cruzada en donde la desviación estándar para distintas medidas de desempeño es menor. Asi pues el uso de un algoritmo de selección de características no solo favorece en la disminución de atributos que participan en el modelo, si no que también, como beneficio secundario, ayuda a crear modelos más estables y de mejor desempeño.

La NBA, y el asociado NBC, es ampliamente utilizado en multitud de diferentes contextos. Ha demostrado ser muy robusto y eficaz a través de problemas de multiples dominios a pesar de la fuerte suposición de que todos los atributos son independientes. Hay casos donde no funciona muy bien sin embargo, y por esto ha sido una área activa de investigación tanto para entender por que funciona y también para mejoras en el diseño.

En cuanto al algoritmo Naive Bayes Generalizado (NBG) presentado en el capítulo 5 podemos resumir que hemos propuesto y probado un conjunto de diagnóstico de error para detectar y cuantificar el efecto de las correlaciones en los elementos en ambos niveles local (subconjuntos de elementos - esquemas) y global (el conjunto completo de elementos). Por interpolación entre los niveles locales y globales permiten un entendimiento completo de como los errores se cancelan a través de las diferentes combinaciones de elementos. Por lo tanto, uno puede no solo predecir el potencial desempeño de la NBA, si no también determinar cuales subconjunto de elementos deberían combinarse en una generalización de la NBA. La factorización óptima para un ejemplar de la GBA debe ser la cual respeta mejor la estructura de la correlación subyacente del problema en cuestión. Nuestros diagnósticos son una ayuda en la búsqueda para esta estructura de correlación. Obviamente en un problema del mundo real esta estructura debe ser inferida de muestras finitas y por lo tanto está sujeta a errores de muestreo. Nuestro énfasis aquí fue sobre el sesgo del modelo asociado con la NBA frente a la GBA o la estructura de correlación exacta.

Uno de los conceptos más interesantes que aporta este trabajo a nuestra consideración, junto con el descrito para NBG, es el de considerar atributos dinámicos, como lo mostramos en el capítulo 6, la importancia de tomar los atributos tipo evento y codificarlos de forma dinámica con un valor en cada una de las ventanas, esto dota al modelo y sobre todo a las predicciones de flexibilidad temporal, esto es el modelo puede ir progresando ventana a ventana (con el tiempo) para ajustarse mejor a un fenómeno que depende de la temporalidad o eventos ocurridos en los atributos que lo describen y sobre todo eso se puede ver en los fenómenos relacionados a las enfermedades emergentes, por que difícilmente una enfermedad podrá comportarse de forma lineal y estática a lo largo del tiempo o evolución del fenómeno, por ejemplo padecimientos como la diabetes u obesidad no son condiciones que aparezcan de la noche a la mañana en una persona, ni de un año a otro o en dos, etc., son padecimientos que aparecen tras años de una serie de eventos a lo largo de la vida de un persona como puede ser el estilo de vida, ejercicio e inclusive la obesidad puede ser un atributo tipo evento para la diabetes, no solo es importante si padece o no obesidad la persona si no desde cuando y cuantos años lleva en esa condición, de ahí la importancia de crear modelos dinámicos que mejor representen la evolucion temporal de un fenómeno como el de las enfermedades, de ahí la importancia de un algoritmo dinámico como el presentado en este trabajo como un primer acercamiento o intento de solución para la modelación de fenómenos dinámicos o minería de datos dinámica.

También analizamos la importancia de la actualización periódica de los modelos, sobre todo aquellos fenómenos donde su medio ambiente es cambiante con el paso del tiempo y por ende los atributos que describen e influencian al fenómeno como tal, si hay fenómenos que por su naturaleza son cambiantes con el paso del tiempo esos son los de las enfermedades emergentes, las intervenciones y distintas estrategias utilizadas por el sector salud, alteran el medio ambiente en el que se desenvuelven estos fenómenos a tal grado de que atributos los cuales en un primer modelo podrían mostrarse como factores de riesgo o protectores con el paso del tiempo pierdan su influencia sobre el fenómeno y pierdan ese estatus de factores de riesgo o protectores, desde el punto de vista del computo no seria

funcional desarrollar un modelo y varios años después regresar con el usuario del mismo para re-hacer o actualizar el modelo, de ahí la propuesta del algoritmo adaptativo que hicimos en este trabajo, el cual es un simple algoritmo de computo, el cual permite a un modelo creado con NBA actualizarse de forma automática sin la intervención humana para dar así a lo modelos durabilidad y sobre todo vigencia con el paso del tiempo y cambios en el fenómeno sufridos por el medio ambiente donde se desenvuelve el mismo.

Este trabajo presenta varias propuestas de modificaciones al algoritmo tradicional de clasificación NBC con el objetivo de mejorar el desempeño y la precisión en el ranqueo de los modelos creados al utilizar NBA, aun cuando los algoritmos fueron diseñados para funcionar bien y adaptarse al NBA, las ideas y conceptos propuestos pueden ser utilizados por diferentes algoritmo de minería de datos, no necesariamente los algoritmos e ideas propuestas son exclusivamente funcionales con el NBA. La mayoría de las pruebas se elaboraron sobre datos relacionados con la diabetes, pero al igual que mencionamos sobre los algoritmos, esto no necesariamente nos dice que los algoritmos propuestos en el trabajo solo funcionan sobre fenómenos de enfermedades emergentes, los conceptos y algoritmos presentados son perfectamente funcionales con otros tipos de fenómenos y datos. El hecho de haber elegido a las enfermedades emergentes (en especifico a la diabetes), como datos de prueba y fenómeno de modelación para probar los algoritmos, es debido a que este tipo de datos son lo suficientemente dinámicos, cambiantes y complejos como para poder mostrar la necesidad y pertinencia de utilizar más allá de la minería de datos tradicional algo de mayor sofisticación que capture la esencia de fenómenos cambiantes y complejos como es el padecimiento de la diabetes.

Apéndice A

Doce Distribuciones de Probabilidad

A continuación, en la tabla A, se muestran las doce distribuciones de probabilidad de dos-variables para las probabilidades $P(X_1X_2|C)$ y $P(X_1X_2|\bar{C})$ que utilizamos como bloques fundamentales para la construcción de nuestro análisis. La primera columna, Dis., denota la distribución, $1-12$, la segunda, Conf., la clase y la configuración de las variables (CX_1X_2 con $C=1$, $\bar{C}=0$ y $X_i=0,\ 1$), la tercera, Lik, la probabilidad correspondiente para la clase/combinación de variables, la cuarta y quinta , δ_C y $\delta_{\bar{C}}$, las funciones de error correspondientes para las probabilidades individuales, ecuaciones 5.37) y (5.38); la sexta columna es la función de error Δ (5.41); la séptima columna, Post.,es la probabilidad posterior exacta y la octava la probabilidad posterior de la NBA, con la columna 9 siendo el porcentaje de diferencia entre ellos; la columna 10, SNB, es el score en la NBA y la columna 11, SGNB, el score en la GBA (el score exacto en este caso), y por último, la columna 12 es el porcentaje de diferencia entre ellos. También se muestra al final de cada distribución el error absoluto promedio de las estimaciones de la NBA para la probabilidad posterior y la función de score.

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
1	111	0.01	0.043	1.636	-3.641	0.03	0.56	-94.24 %	0.24	-3.40	2196.87 %
1	110	0.456	1.957	0.222	2.175	0.93	0.61	53.18 %	0.44	2.62	6226.83 %
1	101	0.49	1.835	0.363	1.618	0.88	0.59	48.46 %	0.38	1.99	-5483.14 %
1	100	0.044	0.164	1.777	-2.378	0.14	0.64	-77.88 %	0.58	-1.80	-1389.69 %
1	011	0.45	1.636	0.042	3.640	0.97	0.44	119.77 %	-0.24	3.40	2196.87 %
1	010	0.05	0.222	1.957	-2.175	0.07	0.39	-82.61 %	-0.44	-2.62	6226.83 %
1	001	0.1	0.363	1.835	-1.618	0.12	0.41	-70.58 %	-0.38	-1.99	-5483.14 %
1	000	0.4	1.777	0.164	2.378	0.86	0.36	138.63 %	-0.58	1.80	-1389.69 %
						Error Abs Prom		85.67 %	Error Abs Prom		3824.13 %
2	111	0.01	0.042	0.363	-2.136	0.13	0.56	-76.69 %	0.24	-1.90	1289.34 %
2	110	0.456	1.957	1.777	0.096	0.63	0.61	3.72 %	0.44	0.54	275.03 %
2	101	0.49	1.835	1.636	0.114	0.62	0.59	4.61 %	0.38	0.49	-388.45 %
2	100	0.044	0.164	0.222	-0.298	0.57	0.64	-11.14 %	0.58	0.28	-174.69 %
2	011	0.1	0.363	0.042	2.136	0.87	0.44	97.47 %	-0.24	1.90	1289.34 %
2	010	0.4	1.777	1.957	-0.096	0.37	0.39	-5.78 %	-0.44	-0.54	275.03 %
2	001	0.45	1.636	1.835	-0.114	0.38	0.41	-6.72 %	-0.38	-0.49	-388.45 %

Continua en la siguiente página

Dis	**Conf**	**Lik**	δ_C	$\delta_{\bar{C}}$	Δ	**Post**	**PostNB**	**% diff**	**SNB**	**SGNB**	**% diff**
2	000	0.05	0.222	0.164	0.298	0.43	0.36	19.83 %	-0.58	-0.28	-174.69 %
						Error Abs Prom		28.25 %	Error Abs Prom		531.88 %
3	111	0.01	0.0429	1.777	-3.723	0.04	0.61	-94.06 %	0.44	-3.28	-10658.34 %
3	110	0.456	1.957	0.363	1.683	0.87	0.56	55.89 %	0.24	1.92	-1015.52 %
3	101	0.49	1.835	0.222	2.111	0.94	0.64	46.23 %	0.58	2.69	1233.57 %
3	100	0.044	0.164	1.636	-2.295	0.13	0.59	-78.43 %	0.38	-1.92	7775.56 %
3	011	0.4	1.777	0.042	3.723	0.96	0.39	146.10 %	-0.44	3.28	-10658.34 %
3	010	0.1	0.363	1.957	-1.683	0.13	0.44	-71.03 %	-0.24	-1.92	-1015.52 %
3	001	0.05	0.222	1.835	-2.111	0.06	0.36	-82.29 %	-0.58	-2.69	1233.57 %
3	000	0.45	1.636	0.164	2.295	0.87	0.41	114.22 %	-0.38	1.92	7775.56 %
						Error Abs Prom		86.03 %	Error Abs Prom		5170.75 %
4	111	0.2	1.045	0.826	0.234	0.91	0.89	2.41 %	2.07	2.30	14.13 %
4	110	0.015	0.634	1.021	-0.476	0.07	0.11	-35.26 %	-2.11	-2.59	18.92 %
4	101	0.69	0.987	1.085	-0.094	0.93	0.93	-0.65 %	2.65	2.56	-4.20 %
4	100	0.095	1.100	0.989	0.106	0.19	0.18	9.01 %	-1.53	-1.42	-5.49 %
4	011	0.03	0.826	1.045	-0.234	0.09	0.11	-19.03 %	-2.07	-2.30	14.13 %
4	010	0.3	1.021	0.634	0.476	0.93	0.89	4.26 %	2.11	2.59	18.92 %
4	001	0.08	1.085	0.987	0.094	0.07	0.07	9.20 %	-2.65	-2.56	-4.20 %
4	000	0.59	0.989	1.100	-0.106	0.81	0.82	-1.96 %	1.53	1.42	-5.49 %
						Error Abs Prom		10.22 %	Error Abs Prom		10.68 %
5	111	0.53	1.092	0.595	0.607	0.89	0.81	9.34 %	1.47	2.07	57.27 %
5	110	0.24	0.842	1.127	-0.291	0.38	0.45	-15.82 %	-0.22	-0.51	46.72 %
5	101	0.1	0.690	1.944	-1.035	0.52	0.75	-31.14 %	1.10	0.07	-148.11 %
5	100	0.13	1.527	0.701	0.777	0.55	0.36	53.04 %	-0.58	0.20	-78.93 %
5	011	0.1	0.595	1.092	-0.607	0.11	0.19	-40.43 %	-1.47	-2.07	57.27 %
5	010	0.6	1.127	0.842	0.291	0.63	0.55	12.71 %	0.22	0.51	46.72 %
5	001	0.14	1.944	0.690	1.035	0.48	0.25	94.01 %	-1.10	-0.07	-148.11 %
5	000	0.16	0.701	1.527	-0.777	0.45	0.64	-29.70 %	0.58	-0.20	-78.93 %
						Error Abs Prom		35.77 %	Error Abs Prom		82.76 %
6	111	0.15	0.659	1.183	-0.584	0.31	0.45	-30.54 %	-0.21	-0.80	94.49 %
6	110	0.5	1.183	0.659	0.584	0.83	0.74	13.25 %	1.02	1.61	94.49 %
6	101	0.2	1.632	0.659	0.906	0.67	0.45	49.21 %	-0.21	0.69	-146.47 %
6	100	0.15	0.659	1.632	-0.906	0.53	0.74	-28.05 %	1.02	0.12	-146.47 %
6	011	0.5	1.183	0.659	0.584	0.69	0.55	24.67 %	0.21	0.80	94.49 %
6	010	0.15	0.659	1.183	-0.584	0.17	0.26	-36.90 %	-1.02	-1.61	94.49 %
6	001	0.15	0.659	1.632	-0.906	0.33	0.55	-39.74 %	0.21	-0.69	-146.47 %
6	000	0.2	1.632	0.659	0.906	0.47	0.26	78.15 %	-1.02	-0.12	-146.47 %
						Error Abs Prom		37.56 %	Error Abs Prom		120.48 %
7	111	0.08	0.314	0.363	-0.146	0.55	0.58	-6.18 %	0.33	0.18	189.51 %
7	110	0.3	2.392	1.777	0.296	0.53	0.46	16.27 %	-0.18	0.12	-50.79 %
7	101	0.59	1.420	1.636	-0.141	0.66	0.69	-4.45 %	0.82	0.68	-34.33 %

Continua en la siguiente página

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
7	100	0.03	0.146	0.222	-0.415	0.47	0.58	-17.90 %	0.31	-0.11	437.46 %
7	011	0.1	0.363	0.314	0.146	0.45	0.42	8.58 %	-0.33	-0.18	189.51 %
7	010	0.4	1.777	2.392	-0.296	0.47	0.54	-13.60 %	0.18	-0.12	-50.79 %
7	001	0.45	1.636	1.420	0.141	0.34	0.31	10.08 %	-0.82	-0.68	-34.33 %
7	000	0.05	0.222	0.146	0.415	0.53	0.42	24.42 %	-0.31	0.11	437.46 %
						Error Abs Prom		12.69 %	Error Abs Prom		178.02 %
8	111	0.1	0.363	1.069	-1.078	0.20	0.43	-52.61 %	-0.29	-1.37	154.87 %
8	110	0.4	1.777	0.887	0.695	0.67	0.50	33.47 %	0.00	0.69	-170.59 %
8	101	0.45	1.636	0.439	1.313	0.96	0.86	11.57 %	1.80	3.11	94.22 %
8	100	0.05	0.222	1.913	-2.153	0.48	0.89	-45.62 %	2.09	-0.06	-127.92 %
8	011	0.59	1.069	0.363	1.078	0.80	0.57	39.33 %	0.29	1.37	154.87 %
8	010	0.3	0.887	1.777	-0.695	0.33	0.50	-33.40 %	0.00	-0.69	-170.59 %
8	001	0.03	0.439	1.636	-1.313	0.04	0.14	-70.01 %	-1.80	-3.11	94.22 %
8	000	0.08	1.913	0.222	2.153	0.52	0.11	368.34 %	-2.09	0.06	-127.92 %
						Error Abs Prom		81.80 %	Error Abs Prom		136.90 %
9	111	0.1	0.363	0.146	0.908	0.83	0.67	24.67 %	0.70	1.61	307.14 %
9	110	0.4	1.777	1.420	0.224	0.50	0.45	12.48 %	-0.21	0.02	-36.61 %
9	101	0.45	1.636	2.392	-0.379	0.69	0.77	-9.72 %	1.19	0.81	-48.37 %
9	100	0.05	0.222	0.314	-0.346	0.48	0.57	-15.11 %	0.28	-0.06	280.28 %
9	011	0.03	0.146	0.363	-0.908	0.17	0.33	-49.73 %	-0.70	-1.61	307.14 %
9	010	0.59	1.420	1.777	-0.224	0.50	0.55	-10.14 %	0.21	-0.02	-36.61 %
9	001	0.3	2.392	1.636	0.379	0.31	0.23	31.98 %	-1.19	-0.81	-48.37 %
9	000	0.08	0.314	0.222	0.346	0.52	0.43	20.03 %	-0.28	0.06	280.28 %
						Error Abs Prom		21.73 %	Error Abs Prom		168.10 %
10	111	0.1	0.363	2.392	-1.883	0.33	0.77	-56.53 %	1.19	-0.69	-239.90 %
10	110	0.4	1.777	0.146	2.495	0.95	0.62	52.97 %	0.50	3.00	2625.35 %
10	101	0.45	1.636	0.314	1.650	0.89	0.62	44.59 %	0.48	2.13	2140.89 %
10	100	0.05	0.222	1.420	-1.854	0.11	0.45	-74.84 %	-0.21	-2.06	302.53 %
10	011	0.3	2.392	0.363	1.883	0.67	0.23	185.96 %	-1.19	0.69	-239.90 %
10	010	0.03	0.146	1.777	-2.495	0.05	0.38	-87.38 %	-0.50	-3.00	2625.35 %
10	001	0.08	0.314	1.636	-1.650	0.11	0.38	-72.24 %	-0.48	-2.13	2140.89 %
10	000	0.59	1.420	0.222	1.854	0.89	0.55	60.81 %	0.21	2.06	302.53 %
						Error Abs Prom		79.42 %	Error Abs Prom		1327.17 %
11	111	0.02	0.078	1.816	-3.142	0.06	0.58	-90.37 %	0.31	-2.83	3260.35 %
11	110	0.48	1.959	0.115	2.828	0.96	0.59	63.71 %	0.35	3.18	-5021.02 %
11	101	0.49	1.921	0.0418	3.827	0.99	0.62	60.35 %	0.47	4.30	5984.44 %
11	100	0.01	0.040	2.038	-3.910	0.03	0.62	-94.84 %	0.51	-3.40	-3760.22 %
11	011	0.51	1.816	0.078	3.142	0.94	0.42	123.09 %	-0.31	2.83	3260.35 %
11	010	0.03	0.115	1.959	-2.828	0.04	0.41	-90.33 %	-0.35	-3.18	-5021.02 %
11	001	0.01	0.041	1.921	-3.827	0.01	0.38	-96.51 %	-0.47	-4.30	5984.44 %

Continua en la siguiente página

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
11	000	0.45	2.038	0.040	3.910	0.97	0.38	157.85 %	-0.51	3.40	-3760.22 %
						Error Abs Prom		79.42 %	Error Abs Prom		4506.51 %
12	111	0.5	1.814	1.957	-0.075	0.62	0.64	-2.76 %	0.57	0.50	-45.14 %
12	110	0.02	0.081	0.164	-0.700	0.41	0.58	-29.93 %	0.32	-0.38	791.49 %
12	101	0.03	0.117	0.0429	1.010	0.82	0.62	31.78 %	0.49	1.50	1150.28 %
12	100	0.45	1.994	1.835	0.083	0.58	0.56	3.65 %	0.24	0.32	-49.46 %
12	011	0.456	1.957	1.814	0.075	0.38	0.36	4.90 %	-0.57	-0.50	-45.14 %
12	010	0.044	0.164	0.081	0.700	0.59	0.42	41.10 %	-0.32	0.38	791.49 %
12	001	0.01	0.042	0.117	-1.010	0.18	0.38	-52.04 %	-0.49	-1.50	1150.28 %
12	000	0.49	1.835	1.994	-0.083	0.42	0.44	-4.63 %	-0.24	-0.32	-49.46 %
						Error Abs Prom		21.35 %	Error Abs Prom		509.09 %

Tabla A.1: Características de las doce distribuciones de probabilidad

A continuación, en la tabla A, se muestra el desempeño de la NBA y NBC para cada una de las doce distribuciones de probabilidad de dos-variables para las probabilidades vistas anteriormente en la tabla A. Las primeras tres columnas, Dis, Conf y Lik son las mismas como en la tabla A. La cuarta columna es la clase real para la configuración, la quinta columna es la clase predicha usando la NBA, la sexta columna la clase predicha usando la GBA (la cual es el caso exacto). Del mismo modo, las columnas siete y ocho el rango relativo de score de la configuración en la NBA y GBA respectivamente. Finalmente, debajo de cada conjunto de configuración para una distribución dada esta la exactitud de la clasificación global - el número de elementos correctamente identificados en la NBA y GBA y la distancia entre las aproximaciones de la NBA y GBA.

Dis	Conf	Lik	Class	Class NB	Class GNB	Rank NB	Rank GNB
1	111	0.01	0	1	0	1	1
1	110	0.456	1	1	1	3	4
1	101	0.49	1	1	1	2	3
1	100	0.044	0	1	0	4	2
1	011	0.45	1	0	1		
1	010	0.05	0	0	0		
1	001	0.1	0	0	0		
1	000	0.4	1	0	1		
			Desempeño	4	8	Distancia	2.45
2	111	0.01	0	1	0	1	1
2	110	0.456	1	1	1	3	4
2	101	0.49	1	1	1	2	3
2	100	0.044	1	1	1	4	2
2	011	0.1	1	0	1		
2	010	0.4	0	0	0		
2	001	0.45	0	0	0		
2	000	0.05	0	0	0		
			Desempeño	6	8	Distancia	2.45

Continua en la siguiente página

Dis	Conf	Lik	Class	Class NB	Class GNB	Rank NB	Rank GNB
3	111	0.01	0	1	0	3	1
3	110	0.456	1	1	1	1	3
3	101	0.49	1	1	1	4	4
3	100	0.044	0	1	0	2	2
3	011	0.4	1	1	1		
3	010	0.1	0	1	0		
3	001	0.05	0	1	0		
3	000	0.45	1	1	1		
			Desempeño	4	8	Distancia	2.83
4	111	0.2	1	1	1	3	3
4	110	0.015	0	0	0	1	1
4	101	0.69	1	1	1	4	4
4	100	0.095	0	0	0	2	2
4	011	0.03	0	0	0		
4	010	0.3	1	1	1		
4	001	0.08	0	0	0		
4	000	0.59	1	1	1		
			Desempeño	8	8	Distancia	0.00
5	111	0.53	1	1	1	4	4
5	110	0.24	0	0	0	2	1
5	101	0.1	1	1	1	3	2
5	100	0.13	1	0	1	1	3
5	011	0.1	0	0	0		
5	010	0.6	1	1	1		
5	001	0.14	0	0	0		
5	000	0.16	0	1	0		
			Desempeño	6	8	Distancia	2.45
6	111	0.15	0	0	0	1	1
6	110	0.5	1	1	1	3	4
6	101	0.2	1	0	1	2	3
6	100	0.15	1	1	1	4	2
6	011	0.5	1	1	1		
6	010	0.15	0	0	0		
6	001	0.15	0	1	0		
6	000	0.2	0	0	0		
			Desempeño	6	8	Distancia	2.45
7	111	0.08	1	1	1	3	3
7	110	0.3	1	0	1	1	2
7	101	0.59	1	1	1	4	4
7	100	0.03	0	1	0	2	1
7	011	0.1	0	0	0		
7	010	0.4	0	1	0		
7	001	0.45	0	0	0		
7	000	0.05	1	0	1		
			Desempeño	4	8	Distancia	1.41
8	111	0.1	0	0	0	1	1
8	110	0.4	1	1	1	2	3
8	101	0.45	1	1	1	3	4

Continua en la siguiente página

Dis	Conf	Lik	Class	Class NB	Class GNB	Rank NB	Rank GNB
8	100	0.05	0	1	0	4	2
8	011	0.59	1	1	1		
8	010	0.3	0	1	0		
8	001	0.03	0	0	0		
8	000	0.08	1	0	1		
			Desempeño	5	8	Distancia	2.45
9	111	0.1	1	1	1	3	4
9	110	0.4	1	0	1	1	2
9	101	0.45	1	1	1	4	3
9	100	0.05	0	1	0	2	1
9	011	0.03	0	1	0		
9	010	0.59	0	1	0		
9	001	0.3	0	0	0		
9	000	0.08	1	1	1		
			Desempeño	4	8	Distancia	2.00
10	111	0.1	0	1	0	4	2
10	110	0.4	1	1	1	3	4
10	101	0.45	1	1	1	2	3
10	100	0.05	0	0	0	1	1
10	011	0.3	1	0	1		
10	010	0.03	0	1	0		
10	001	0.08	0	1	0		
10	000	0.59	1	1	1		
			Desempeño	4	8	Distancia	2.45
11	111	0.02	0	1	0	4	3
11	110	0.48	1	1	1	3	2
11	101	0.49	1	1	1	2	1
11	100	0.01	0	1	0	1	4
11	011	0.51	1	1	1		
11	010	0.03	0	1	0		
11	001	0.01	0	1	0		
11	000	0.45	1	1	1		
			Desempeño	4	8	Distancia	3.46
12	111	0.5	1	1	1	4	3
12	110	0.02	0	1	0	2	1
12	101	0.03	1	1	1	3	4
12	100	0.45	1	1	1	1	2
12	011	0.456	0	0	0		
12	010	0.044	1	0	1		
12	001	0.01	0	0	0		
12	000	0.49	0	0	0		
			Desempeño	6	8	Distancia	2.00

Tabla A.2: Comparación de desempeño para las 12 distribuciones de probabilidad

Apéndice B

Atributos Dinámicos

En este apendice se muestra los graficos de varias combinaciones de los atributos dinmicos del modelo creado en 6.3.1.

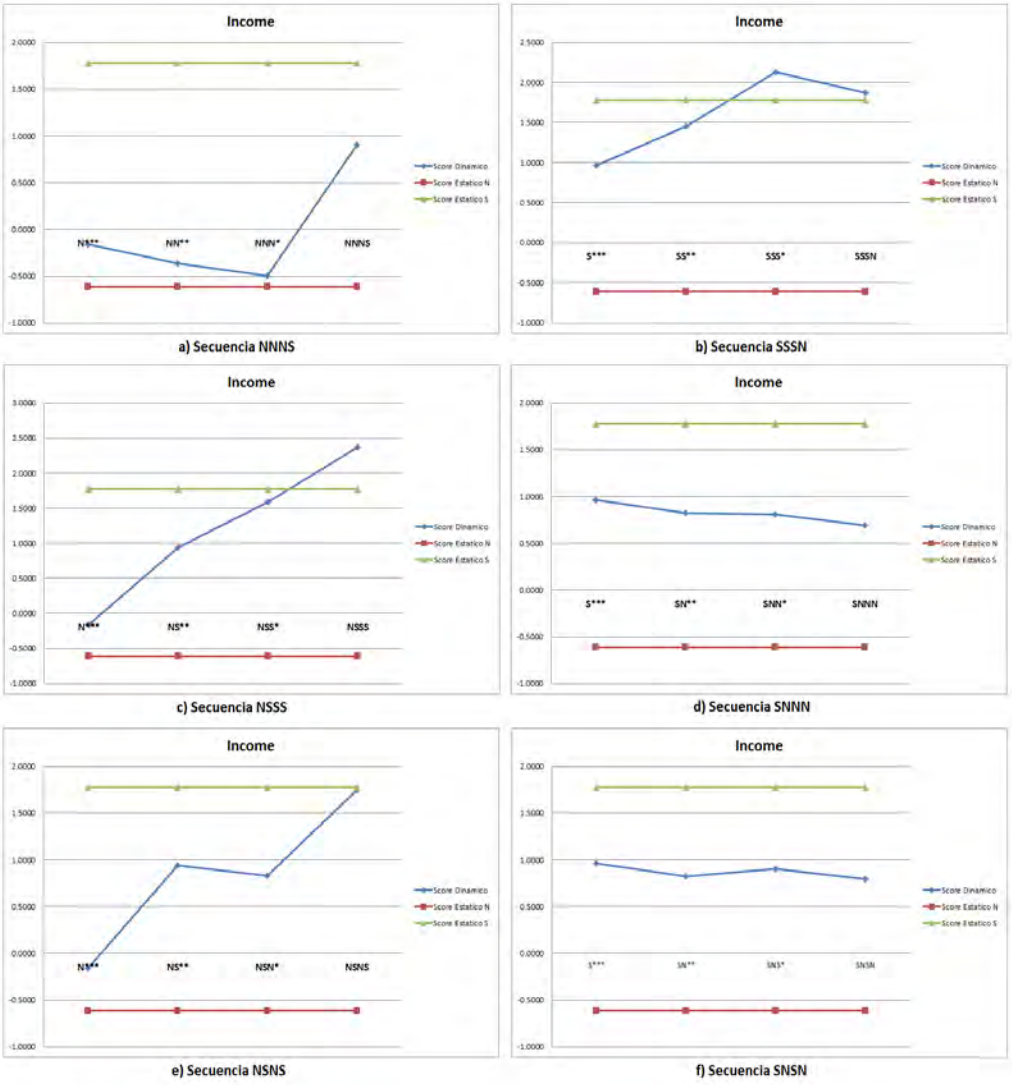


Figura B.1: Atributo dinámico Income tipo evento codificado a través de los 4 trimestres con combinaciones S y N, comparación de los scores obtenidos vs el score del mismo atributo codificado de forma estática.

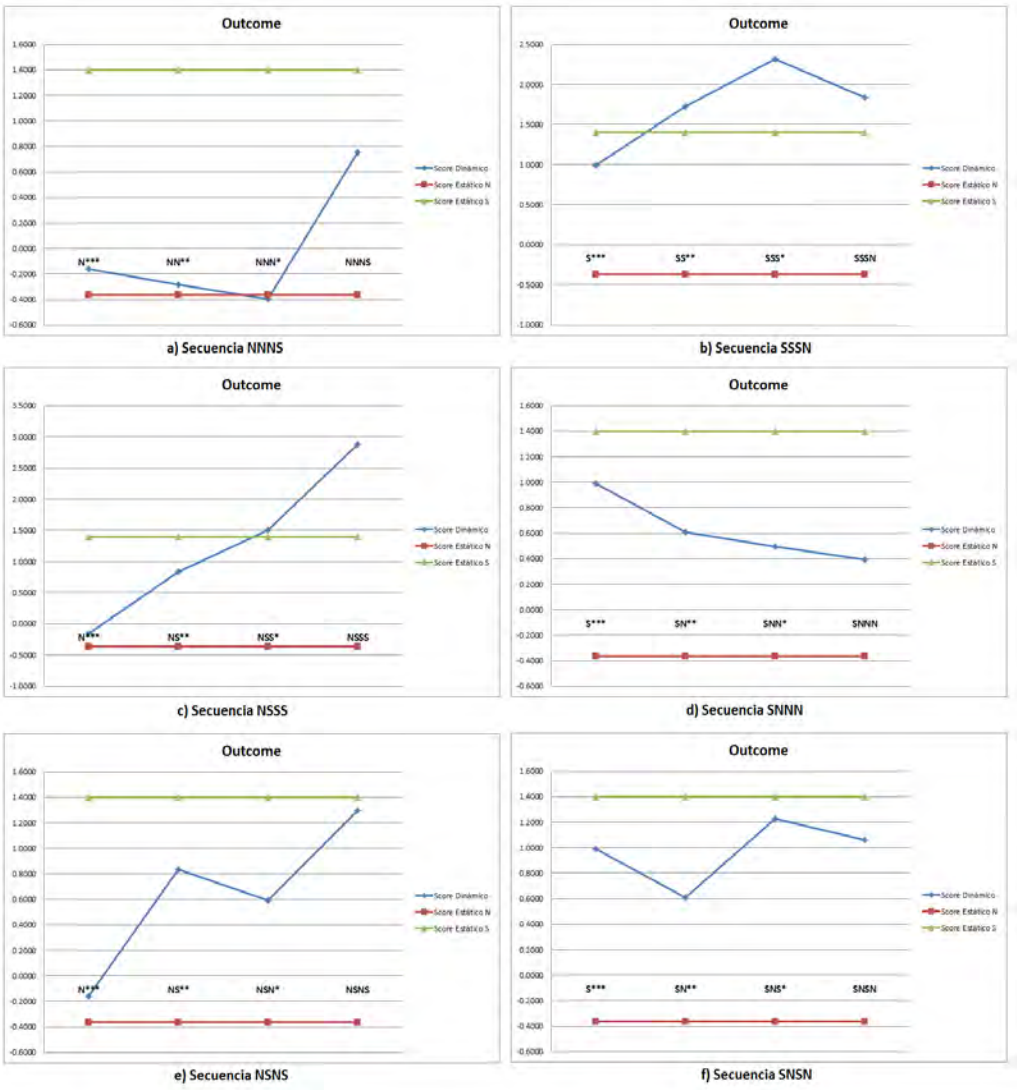


Figura B.2: Atributo dinámico Outcome tipo evento codificado a través de los 4 trimestres con combinaciones S y N, comparación de los scores obtenidos vs el score del mismo atributo codificado de forma estática.

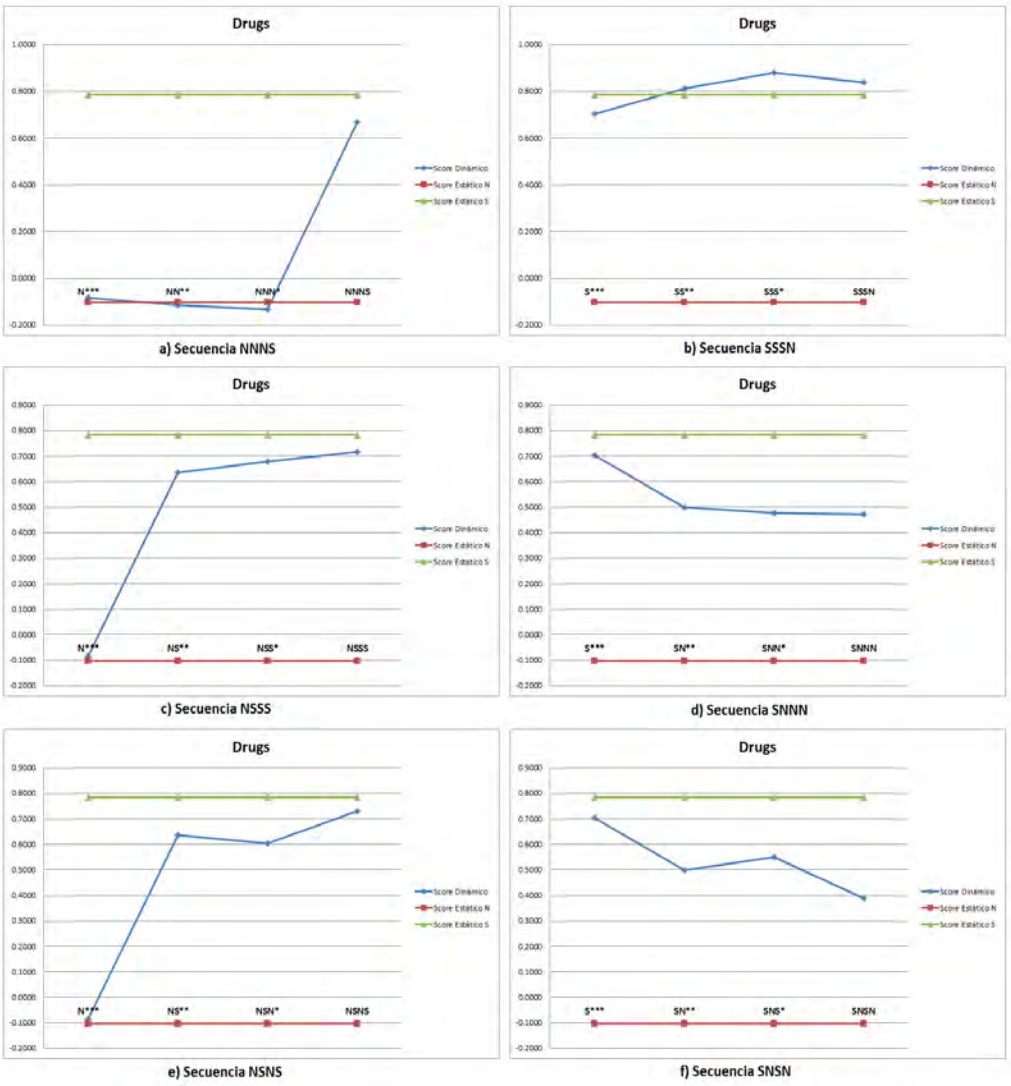


Figura B.3: Atributo dinámico Drugs tipo evento codificado a través de los 4 trimestres con combinaciones S y N, comparación de los scores obtenidos vs el score del mismo atributo codificado de forma estática.

Bibliografía

[1] http://www.webopedia.com/TERM/B/Business_Intelligence.html

[2] http://www.who.int/topics/epidemiology/es/

[3] http://es.wikipedia.org/

[4] R.Bonita, R. Beaglehole, T. Kjellstrom (2008). "Epidemiología Básica", Editado por Organización Panamericana de la Salud, pp. 268, ISBN 978-92-41-547079, Washington D.C.

[5] http://editorialmedsalud.blogspot.mx/2009/11/sistemas-adaptativos-complejos.html

[6] John Holland, Sistemas Adaptativos Complejos, Universidad de Michigan, U.S.A http://ruc.udc.es/bitstream/2183/9449/1/CC_019_art_10.pdf

[7] http://manuelgross.bligoo.com/20120418-big-data-historia-del-dato-de-la-informacion-al-conocimiento

[8] Hernndez Jos (2004). "Introducción a la Minera de Datos", editado por Pearson Prentice Hall, pp. 656, ISBN: 84-205-4091-9, Espaa.

[9] Jasmina Novakovic, The Impact of Feature Selection on the Accuracy of Nave Bayes Classifier, 18th Telecommunications forum TELFOR, 2010. http://2010.telfor.rs/files/radovi/TELFOR2010_10_13.pdf

[10] Chotirat Ratanamahatana, Dimitriosgunopulos, Feature Selection For The Naive Bayesian Classifier Using Decision Trees, Applied Artificial Intelligence, 2003, pp. 475 487.

[11] Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qu, Feature selection for text classification with Nave Bayes, Expert Systems with Applications, Volume 36, Issue 3 Part 1, April 2009, Pages 54325435.

[12] Karl-Michael Schneider, A new feature selection score for multinomial naive Bayes text classification based on KL-divergence, 04 Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004. http://dl.acm.org/citation.cfm?id=1219068

[13] Chanju Kim, Kyu-Baek Hwang, Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking, In Proc. Europ. Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2008. http://www.kde.cs.uni-kassel.de/ws/rsdc08/pdf/4.pdf

[14] M. Morita, R. Sabourin, F. Bortolozzi, C. Y. Suen, Unsupervised Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Word

Recognition, Seventh International Conference on Document Analysis and Recognition, 2003 http://neuro.bstu.by/ai/To-dom/My_research/Papers-0/For-courses/MOGA/moga-hand-w-char.pdf

[15] D. Ruiz, J. García, Algoritmos de Minado de Datos como un servicio en la nube. Posgrado en Ciencia e Ingeniería de la Computación, UNAM-IIMAS, Mayo 2015.

[16] Christos Emmanouilidis, Evolutionary Multi-Objective Feature Selection And Roc Analysis With Application To Industrial Machinery Fault Diagnosis, Evolutionary Methods For Design, Optimization And Control, 2002. http://www.ipet.gr/~chrisem/Files/CEM_ENPGA_ROC.pdf

[17] L. S. Oliveira, R. Sabourin, F. Bortolozzi, C. Y. Suen, Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Digit Recognition, ICPR '02 Proceedings of the 16 th International Conference on Pattern Recognition , 2002.

[18] Christos EMMANOUILIDIS, Andrew HUNTER, John MACINTYRE, Chris COX, A Multi-Objective Genetic Algorithm Approach to Feature Selection in Neural and Fuzzy Modeling, Evolutionary Optimization An International Journal on the Internet, Volumen 3, Numero 1, 2001, pp.1-26 http://www.ceti.gr/~chrisem/Files/eoj_v3n1_001.pdf

[19] Venkatadri.M , Srinivasa Rao.K, A Multiobjective Genetic Algorithm for Feature Selection in Data Mining, International Journal of Computer Science and Information Technologies, Vol. 1 (5) , 2010,pp. 443-448.

[20] Juan Arturo Herrera-Ortiz, Katya Rodriguez-Vazquez, A RankMOEA to Approximate the Pareto Front of a Dynamic Principal-Agent Model, 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011.

[21] Breiman L., Friedman J.H., Olshen R.A., Classification and Regression Trees, Wadsworth and Books/Cole, Monterrey, 1984.

[22] Quinlan J.R., Induction of Decision Trees, Machine Learning. Morgan Kaufmann, 1987.

[23] Quinlan J.R., C4.5 Programs for Machine Learning, San Francisco, Morgan Kaufmann, 1993.

[24] http://pubs.rsc.org/en/content/articlehtml/2012/an/c2an16122b

[25] C.R. Stephens y R. Sukumar. An introduction to Data Mining. The Handbook of Marketing Research. SAGE Publications. 2006.

[26] http://www.veriskhealth.com/answers/population-answers/dxcg-risk-analytics

[27] http://archive.ics.uci.edu/ml/

[28] Stephens C., Flores H., Ruiz A. (2018). When is the Naive Bayes approximation not so naive?, Machine Lerning Vol. 107, pp 397 - 441.

[29] Domingos, P., and Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In Proceedings of the Thirteenth International Conference on Machine Learning (pp. 105-112). Morgan Kaufmann.

[30] Frank, E.; Trigg, L.; Holmes, G.; and Witten, I. H. (2000) Naive Bayes for regression, Machine Learning 41(1):5-15.

[31] Ling, C.X.; Huang, J.; and Zhang, H. (2003) AUC: a statistically consistent and more discriminating measure than accuracy, Proceedings of the 18th international joint conference on Artificial intelligence, 519-524.

[32] Stephens, C. R., Waelbroeck, H., and Talley, S. (2005, June). Predicting healthcare costs using GAs. In Proceedings of the 2005 workshops on Genetic and evolutionary computation (pp. 159-163). ACM.

[33] Bennett, P. N. 2000. Assessing the calibration of Naive Bayes'posterior estimates. In Technical Report No. CMU-CS00-155.

[34] Monti, S., and Cooper, G. F. (1999) A Bayesian network classifier that combines a finite mixture model and a Naive Bayes model. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann. 447-456.

[35] Lowd, D., Domingos, P. (2005) Naive Bayes models for probability estimation, ICML '05 Proceedings of the 22nd International Conference on Machine learning, 529-536 ACM New York, NY, USA.

[36] Zhang, H. and Ling, C.X. (2003) AI 2003, LNAI 2671, pp. 591595, Y. Xiang and B. Chaibdraa (Eds.): Springer-Verlag Berlin Heidelberg.

[37] Zhang, H. (2004) The optimality of naive Bayes, In Proceedings of the FLAIRS Conference, vol. 1, no. 2, pp. 3-9.

[38] Rish, I. (2001) An empirical study of the naive Bayes classifier IJCAI 2001 workshop on empirical methods in artificial intelligence, Volume 3, 22, 41-46.

[39] Friedman, J. (1997a). On bias, variance, 0/1 - loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery, 1, 5577.

[40] Liangxiao, J., Zhang, H., and Cai, Z. (2009) A Novel Bayes Model: Hidden Naive Bayes IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 10, p. 1361.

[41] Zheng, Z., and Webb, G. I. Lazy learning of Bayesian Rules. Machine Learning, 41(1), 53-84.

[42] Keogh, E., and Pazzani, M. (1999). Learning augmented Bayesian classifiers: A comparison of distribution based and classification based approaches. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, pp. 225-230.

[43] Wolpert, David (1996), The Lack of A Priori Distinctions between Learning Algorithms, Neural Computation, pp. 1341-1390.

[44] Wolpert, D.H., Macready, W.G. (1997), No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation 1, 67.

[45] Friedman, N., Geiger, D., and Goldszmidt, M. (1997) Bayesian network classifiers, Machine Learning, 29(2), 131-163.

[46] Kononenko, I. (1991). Semi-naive Bayesian classifier. In Proceedings of the Sixth European Working Session on Learning, pp. 206-219 Berlin. Springer-Verlag.

[47] Holland, J.H., (1975) Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor.

[48] Poli, R., and Stephens, C. R. (2014). Taming the Complexity of Natural and Artificial Evolutionary Dynamics, In Evolution, Complexity and Artificial Life (pp. 19-39). Springer Berlin Heidelberg.

[49] Domingos, P., and Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In Proceedings of the Thirteenth International Conference on Machine Learning (pp. 105-112). Morgan Kaufmann.

[50] J. Demsar (2006) Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research 7, 1-30.

[51] Harry Zhang and Jiang Su, (2004) Naive Bayesian Classifiers for Ranking, J.-F. Boulicaut et al. (Eds.): ECML 2004, LNAI 3201, pp. 501-512. Springer-Verlag Berlin Heidelberg 2004

[52] Harry Zhang and Jiang Su (2008) Naive Bayes for optimal ranking, Journal of Experimental & Theoretical Artificial Intelligence, 20:2, 79-93.

[53] Vijay Raghavan, Alaaeldin Hafez, Dynamic Data Mining, Journal Of The American Society For Information Science, 2000, pp. 220-229.

[54] Sergey Brin, Lawrence Page, Dynamic Data Mining: Exploring Large Rule Spaces by Sampling, Technical Report. Stanford InfoLab, 2008. `http://ilpubs.stanford.edu:8090/424/1/1999-68.pdf`

[55] Ching Lien Huang, Tsung-Shin Hsu , Chih-Ming Liu, The Neural Network Algorithm for Data-Mining in Dynamic Environments, Eighth International Conference on Intelligent Systems Design and Applications, 2008, pp. 622-625. `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4696278&tag=1`

[56] Mohd. Shahnawaz, Ashish Ranjan, Mohd Danish, Temporal Data Mining: An Overview, International Journal of Engineering and Advanced Technology (IJEAT), Volume-1, Issue-1, October 2011, pp. 20-24.

[57] C. M. Antunes, A. L. Oliveira, Temporal data mining: An overview, KDD 2001 Workshop, 2001. `http://www.citeulike.org/user/bluestan/article/2807775`

[58] Luca Chittaro, Carlo Combi, Giampaolo Trapasso, Data Mining on Temporal Data: a Visual Approach and its Clinical Application to Hemodialysis, Journal of Visual Languages and Computing, vol. 14, no. 6, December 2003, pp. 591-620. `http://hcilab.uniud.it/publications/2003\$03/DataMining_JournalVisualLanguages03`

[59] Mohamed Medhat Gaber, Shonali Krishnaswamy, Arkady Zaslavsky, Adaptive Mining Techniques for Data Streams using Algorithm Output Granularity, 2003 Congress on Evolutionary Computation (CEC 2003),2003. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.3862`

[60] Alberto Ochoa-Zezzatti, Fernando Montes, Jns Snchez, Improve Decision Support using Adaptive Data Mining, 2009 International Conference on Electrical, Communications, and Computers, 2009. `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5163890`

[61] Conny Franke, Adaptivity in Data Stream Mining, DISSERTATION for the degree DOCTOR OF PHILOSOPHY in computer science, UNIVERSITY OF CALIFORNIA, 2009. `http://www.cs.ucdavis.edu/research/tech-reports/2010/CSE-2010-10.pdf`

[62] Kohavi, R. (1996). Scaling up the accuracy of naive Bayes classifiers: A decision-tree hybrid. In Proceedings of the Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 202-207 Portland, Or.

[63] Langley, P. (1993). Induction of recursive Bayesian classifiers. In Proceedings of the 1993 European Conference on Machine Learning, pp. 153-164 Berlin. Springer-Verlag.

[64] Langley, P., and Sage, S. (1994). Induction of selective Bayesian classifiers. In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 399-406. Morgan Kaufmann.

[65] Pazzani, M. J. (1996). Constructive induction of Cartesian product attributes. In ISIS: Information, Statistics and Induction in Science, pp. 66-77 Singapore. World Scientific.

[66] Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 334-338 Menlo Park, CA. AAAI Press.

[67] Geoffrey Webb, Janice Boughton, Zhihai Wang Singh, M., and Provan, G. M. (1996). Efficient learning of selective Bayesian network classifiers. In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 453-461 San Francisco. Morgan Kaufmann.

[68] Webb, G. I., and Pazzani, M. J. (1998). Adjusted probability naive Bayesian induction. In Proceedings of the Eleventh Australian Joint Conference on Artificial Intelligence, pp. 285-295 Berlin. Springer.

[69] Webb, G. I., Candidate elimination criteria for Lazy Bayesian Rules. In Proceedings of the Fourteenth Australian Joint Conference on Artificial Intelligence, pp. 545-556 Berlin. Springer.

[70] Xie, Z., Hsu, W., Liu, Z., and Lee, M. L. (2002). SNNB: A selective neighborhood based naive Bayes for lazy learning. In Chen, M.- S., Yu, P. S., and Liu, B. (Eds.), Advances in Knowledge Discovery and Data Mining, Proceedings PAKDD 2002, pp. 104-114 Berlin. Springer.

[71] Zheng, Z., Webb, G. I., and Ting, K. M. (1999). Lazy Bayesian Rules: A lazy semi-naive Bayesian learning technique competitive to boosting decision trees. In Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99), pp. 493-502. Morgan Kaufmann.

[72] Zheng, Z., and Webb, G. I. Lazy learning of Bayesian Rules. Machine Learning, 41(1), 53-84.

[73] `http://en.wikipedia.org/wiki/Receiver_operating_characteristic`