UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"DETERMINACIÓN DE LA MUESTRA ÓPTIMA DE BASES DE DATOS PARA LA MINERÍA DE DATOS"

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN INGENIERÍA (COMPUTACIÓN)

P R E S E N T A:

ALEXIS RENÉ JAVIER LOZANO CÁRDENAS

DIRECTOR DE TESIS:

DR. ÁNGEL FERNANDO KURI MORALES

México, D.F. 2011.

A mi familia, mi papá, mamá, mis hermanos Rulo, Paty, Lolita, a ustedes les debo toda mi formación, este trabajo es también de ustedes.

A mis amigos de la coordinación de Administración Esther, Emilia, Pablo, Anaíd, Israel por su apoyo y paciencia todos estos años.

我非常感谢我的汉语班的朋友们，特别感谢 Jorge, Lucy, 金娅曦 和 Sandra. 你们教会了我很多生活重要的地方，无论我们再不在一起，永远会是我的朋友.

A Alicia por su paciencia y orientación durante tantos años de apoyo.

A mis compañeros de ciencias de la computación, especialmente a Laura Leonides por su apoyo.

A mi tutor Angel Kuri por su paciencia, por su orientación tanto en lo académico como en lo personal, le agradezco mucho.

Índice general

Capítulo 1
Introducción

1.1 Descripción general

En el presente trabajo se expone el análisis de una metodología para la optimización de la búsqueda de grupos de datos similares, aplicable en bases de datos de grandes dimensiones por medio de la selección de una muestra representativa. Esta muestra es fundamentada y validada a través de conceptos de teoría de la información, estadística, aprendizaje de máquinas, y teoría de la aproximación manteniendo un bajo costo computacional en cada etapa, sin hacer para ello suposiciones en la distribución de los datos.

La organización de este trabajo es la siguiente: el capítulo 1 presenta una introducción al tema, descripción de los objetivos y alcances de la investigación así como una descripción en el estado del arte de los elementos en el proceso de agrupamiento de datos principalmente enfocado a grandes volúmenes de datos. El segundo capítulo presenta los elementos teóricos manejados en la metodología aquí propuesta para una clara presentación de las técnicas usadas y de las razones por las que son seleccionadas. El capítulo 3 presenta la metodología propuesta llamada FDM [LOZANO10], del inglés *Fast Data Mining*, presentando las etapas necesarias para la obtención y manejo de una muestra de bases de datos para la búsqueda de grupos de datos. El capítulo 4 describe los resultados que se pueden obtener a través del uso de los elementos aquí descritos y la comparación con alternativas estadísticas en algunas etapas del proceso. Finalmente el capítulo 5 contiene las conclusiones, la descripción de los resultados generales obtenidos en esta investigación, así como puntos que pueden abrir paso a nuevas investigaciones en el tema a partir de los procesos propuestos e implementados.

1.2 Objetivos

La investigación realizada tiene por objetivo los siguientes puntos

- Un estudio y análisis de las técnicas más usadas para el preprocesamiento y agrupamiento de datos que presenten menos limitaciones por la suposición de distribución en los datos o requerir el uso de equipo de supercómputo o paralelismo en la etapa de procesamiento de la información, sin descartar la posibilidad de su uso.
- Desarrollo de una metodología para la obtención de una muestra representativa para la minería de datos, específicamente para algoritmos de agrupamiento.
- Presentación de formas de validación de la muestra obtenida que sean aplicables a grandes volúmenes de datos sin las restricciones comúnmente impuestas por las pruebas estadísticas

Los conceptos aquí manejados son provenientes de distintas áreas, por lo que pueden usarse con distinta nomenclatura dependiendo del área desde la cual se analice. Para no dar lugar a dudas en la nomenclatura básica de algunos términos en este trabajo se consideran sinónimos tuplas, instancias, registros y elementos de una base de datos, así como atributo, columna o variable al estar analizando los datos distribuidos en tablas de bases de datos. Otros términos similares entre sí se aclararán a lo largo del trabajo.

1.3 Antecedentes

El almacenamiento, transferencia e interpretación de la información es un aspecto fundamental en la sociedad actual, ya que esta transferencia espacial y temporal del conocimiento es una base del desarrollo humano. Claros ejemplos de su importancia se presentan desde el uso del lenguaje hasta el desarrollo de tecnologías de información.

Estas tecnologías de la información han aumentado la disponibilidad y abundancia de la información almacenada digitalmente creando una gran necesidad del análisis de minería de datos. La minería de datos es un conjunto de técnicas y algoritmos que exploran grandes volúmenes de datos extrayendo modelos y descubriendo patrones ocultos con fines descriptivos o predictivos.

La minería de datos es la parte central del descubrimiento de conocimiento en bases de datos (*KDD*[1]) que se encarga de realizar un análisis automático exploratorio y descriptivo para identificar patrones validos, novedosos y útiles en grandes y complejos conjuntos de datos. Estas técnicas combinan técnicas de campos como la estadística, probabilidad y ciencias e ingeniería de la computación, especialmente de inteligencia artificial.

El proceso de descubrimiento de conocimiento en bases de datos incluye la definición de objetivos, análisis y preprocesamiento de la información, selección de herramientas de minería de datos, interpretación y retroalimentación de resultados.

Las técnicas de minería de datos utilizan conceptos y algoritmos de inteligencia artificial, probabilidad y estadística para alcanzar resultados significativos y utilizables; sin embargo, no existe un algoritmo o una metodología que sea superior a las demás universalmente, ya que al usarse en distintos conjuntos de datos o con distintos objetivos presentarán ventajas y desventajas en exactitud, eficiencia o precisión de resultados. Es por ello que en la minería de datos y todo el proceso de descubrimiento de conocimiento no existe un

[1] *KDD=Knowledge Discovery in Databases*

algoritmo que sea universalmente aplicable y que pueda asegurar el mejor resultado en todos los aspectos.

A pesar de lo anterior, al conocer las capacidades y resultados alcanzables con algún algoritmo se puede hacer una elección informada de qué algoritmo usar o buscar optimizaciones orientadas a la solución del problema en cuestión.

Las bases de datos de grandes dimensiones son particularmente difíciles de manejar y analizar, ya que pueden eliminar la posibilidad de usar algunas técnicas más comunes de minería de datos. Las principales razones se presentan a continuación.

1. La complejidad de muchos algoritmos de minería de datos es exponencial con el número de dimensiones y es común el manejo de grandes cantidades de variables.
2. El concepto de distancia y similitud entre elementos no es claro en un espacio de tan alta dimensionalidad, ya que una distancia euclidiana llega a ser muy poco sensible a variaciones en estas condiciones, provocando que la distancia entre los datos más lejanos y los más cercanos no sea demasiado diferente.
3. Existe la posibilidad de alta relevancia local de atributos, es decir que una variable o atributo puede ser irrelevante para el proceso en un gran número de instancias pero es relevante en otro conjunto de instancias menor, por lo que algunos métodos de selección de características globales no producen los resultados esperados.
4. El número de instancias es tan alto que un algoritmo se vuelve ineficiente al procesar una gran cantidad de datos redundante, ya que es necesario que se hagan múltiples accesos a memoria secundaria.
5. El almacenamiento y organización de datos se vuelve más complejo y requiere de bases de datos con diferentes estructuras y herramientas.

Los primeros 4 puntos corresponden a problemas que tienen los algoritmos de minería de datos y el último punto se relaciona con la organización y almacenamiento de la base de datos, lo cual no corresponde a la minería de datos, sino a sistemas gestores de bases de datos y los llamados almacenes de datos (*data warehouses*) desarrollados para solucionar problemas de acceso, generación de reportes y análisis de información.

El resto de los problemas son consecuencia de la dimensionalidad de la base de datos y la heterogeneidad que es consecuencia de ésta, ya que afectan fuertemente el desempeño de las técnicas de minería de datos.

Estas bases de datos son comunes en aplicaciones comerciales y científicas. Por ello se ha realizado una gran cantidad de investigación con el objetivo de encontrar alternativas a los

algoritmos tradicionales para superar estas dificultades. Estas investigaciones han sido enfocadas en los siguientes puntos:

- Optimización de los algoritmos, reduciendo su complejidad.
- Hibridación de algoritmos para alcanzar mejores resultados
- Preprocesamiento de la información para eliminar variables y reducir el número de instancias
- Desarrollo de algoritmos de procesamiento secuencial o en paralelo en las que los datos son procesados por partes.

A pesar de todas las investigaciones realizadas en los puntos anteriores, el resultado es un conjunto de técnicas que deben ser aplicadas dependiendo de los objetivos, características y alcances deseados en la búsqueda y no se tiene una guía para elegir la herramienta correcta de manera definitiva.

Las optimizaciones más generales que se pueden encontrar son aquellas que son, en la medida de lo posible, independientes de la técnica de búsqueda utilizada. Dado lo anterior, la etapa de preprocesamiento resulta interesante, ya que se realiza previamente a la aplicación del algoritmo de minería, trabajando directamente sobre las características de los datos, pudiendo ser manejado de manera independiente al análisis de minería.

Este preprocesamiento de la información es fundamental y debe incluir procesos de normalización, limpieza de los datos y reducción dimensional.

La reducción de dimensiones se puede hacer en el número de variables manejadas, llamada reducción vertical o selección de características; o se puede aplicar en la cantidad de instancias, llamada reducción horizontal o selección de instancias. La reducción del número de variables se realiza por diferentes métodos basados en la cantidad de información que pueda aportar la variable a un algoritmo, por lo que se han desarrollado diferentes métodos basados en proyecciones y diferentes representaciones espaciales con importantes resultados. La reducción del número de instancias se logra a través de una selección de los datos, por lo general basándose en un criterio estadístico.

Esta reducción horizontal no ha sido ampliamente estudiada con respecto a la aplicación de técnicas de inteligencia artificial, ya que en la práctica se toman criterios tradicionales estadísticos basados en la suposición de distribuciones comunes de una o dos dimensiones [LENTH01] o en heurísticas sin fundamentos claros.

Un muestreo efectivo de una población obtiene una muestra representativa que pueda generar los mismos o mejores resultados, que el conjunto de datos completo de la base de datos.

Entre las técnicas de minería de datos, el conjunto de algoritmos cuyo objetivo es el agrupamiento de datos similares identificados a través de sus características, llamados grupos de datos, es uno de los procesos fundamentales para obtener una descripción general del conjunto de datos ya que permite hacer predicciones, caracterización de elementos, o aplicar procesos de minería de manera selectiva en los datos pertenecientes a grupos con características de interés [HAM01].

El resultado obtenido de la aplicación de algoritmos de agrupamiento es altamente útil y difícil de identificar por medio de otras técnicas, ya que utiliza la información contenida en los datos para hacer una clasificación sin necesidad de una guía en el proceso ya que se trata de algoritmos de aprendizaje no supervisado.

1.4 Metodología**:** justificación y relevancia

Esta investigación se enfocó en el uso de técnicas que sean utilizables en casos reales y aplicables en conjuntos de datos y problemas de distintos tipos, con el fin de hacer agrupamiento de datos que es un proceso muy sensible a la distribución de datos ya que estos guían totalmente el aprendizaje.

Independientemente del algoritmo que sea utilizado para el análisis, la distribución de los datos determinará el modelo resultante con base a los patrones presentes en ellos. En esta investigación se busca encontrar la forma apropiada de seleccionar una muestra de una población sin tener pérdidas importantes en estos patrones identificables, para permitir aplicar cualquier otro método de agrupamiento.

El muestreo estudiado desde el punto de vista estadístico, por lo general, requiere de suposiciones respecto a la distribución de la información, ya que en ocasiones no se tiene acceso a priori a los datos. Así sucede en las ciencias sociales o biológicas; sin embargo en el caso de las bases de datos se tiene acceso a la información, con la única limitación del tiempo de acceso a los datos, permitiendo la selección de un criterio ad hoc a la información contenida sin hacer suposiciones arriesgadas acerca de una distribución que no se ajuste a los datos. En la minería de datos, a diferencia de los conjuntos usados en los análisis estadísticos básicos, es muy difícil encontrar una distribución bien definida en una población tan grande.

La presencia de patrones en los datos que permite la identificación de grupos de datos, es analizada en la metodología aquí presentada, a través de los conceptos de teoría de la información, tomando la entropía de la población como estadística guía en el proceso de muestreo y siendo validada a través de pruebas dependientes de las relaciones no lineales

entre variables sin hacer suposiciones con respecto a la distribución de los datos ni transformaciones que ocasionen pérdidas de información como puede suceder con las pruebas estadísticas no paramétricas tradicionales.

1.5 Estado del arte

1.5.1 Minería de datos

Grandes avances se han dado en el aspecto físico en las tecnologías de la información. Actualmente se cuentan con equipos personales capaces de almacenar cantidades enormes de datos, aunque pueden parecer insignificantes frente a la capacidad de los equipos de supercómputo. La capacidad de recolectar y almacenar información ha superado la capacidad de interpretación de esta, dando lugar una gran evolución en algoritmos de minería de datos en todo tipo de aplicaciones, desde su aplicación en la creación de taxonomías biológicas [FLAKE70], análisis de datos médicos[JENSEN01], hasta análisis de clientes [CABANES09], recuperación de información en textos[FANG06], análisis financiero[POWELL08], etc.

1.5.2 Algoritmos de agrupamiento

Los algoritmos de agrupación son básicos en la minería de datos al mostrar los grupos implícitos en la población de datos, proporcionando una descripción global de las características de la distribución. Existe una gran cantidad de algoritmos de búsqueda de grupos de datos. Los algoritmos se pueden clasificar en los siguientes grupos [HAM01]

- Algoritmos jerárquicos.
- Algoritmos particionales.
- Algoritmos basados en densidad de los datos.
- Algoritmos basados en la generación de modelos

1.5.2.1 Algoritmos jerárquicos y algoritmos basados en teoría de gráficas

Estos algoritmos se basan en la construcción de grupos clasificando las instancias en un proceso de aglomeración o división secuencial. Los algoritmos aglomerativos inician considerando que cada objeto es un grupo por sí mismo e, iterativamente, se va uniendo con los elementos más cercanos hasta que la estructura del grupo es obtenida. Los algoritmos divisivos consideran que inicialmente todos los datos pertenecen a un mismo

grupo y sucesivamente se va dividiendo en subgrupos seleccionando los elementos con diferencias más significativas.
La estrategia usada para medir la distancia entre grupos define características que tendrá el algoritmo en su ejecución. Si la distancia es usada como una medida de similitud entre grupos, esta puede ser tomada de tres formas:

- Enlace simple: en el que la distancia entre dos grupos de datos es medida como la mínima distancia entre cualquiera de los individuos de ambos grupos.
- Enlace completo: la distancia entre dos grupos es igual a la máxima distancia entre dos elementos de ambos grupos.
- Enlace promedio: la distancia entre dos grupos es dada por la distancia entre los centroides de los grupos dados por el promedio de los elementos que lo conforman.

El enlace simple tiende a unir grupos por la presencia de unos pocos elementos en la frontera del grupo pero es más versátil que el enlace completo al poder identificar formas más complejas. El enlace promedio suele romper o unir grupos con formas alargadas [GUHA98]. Estos métodos tienen la gran desventaja de no poder ser escalados ya que son de orden $O(m^2)$, donde m es el número de instancias.

También existen algoritmos basados en teoría de grafos donde las instancias son representadas como nodos y por medio de la aplicación del algoritmo de árbol de expansión mínima [ZAHN71] genera resultados similares a los de los algoritmos jerárquicos tradicionales.

1.5.2.2 Métodos particionales

Estos métodos iterativamente asignan las instancias a diferentes grupos de datos en búsqueda de la combinación que genere un grupo de mayor calidad minimizando la distancia entre los elementos de un mismo grupo. Estos algoritmos requieren conocer el número de grupos que se buscan a priori.

Operan minimizando una medida de error, que por lo general es la suma de errores al cuadrado con respecto al centro del grupo al que es asignado. El algoritmo más utilizado de de este tipo es el de *k-medias* [ROKACH05], que a partir de una selección aleatoria de un conjunto de centros de grupos, iterativamente se asignan los datos más cercanos al grupo correspondiente a ese centro, para posteriormente calcular de nuevo el centro del grupo con base en la media de los elementos pertenecientes al grupo encontrado. Este algoritmo presenta fuertes deficiencias siendo altamente sensible a la elección inicial de centros. Sin embargo es ampliamente usado por su complejidad $O(T\ K\ n\ V)$ donde T es el número de iteraciones, K el número de grupos, n el número de instancias y V el número de variables y por la facilidad de interpretación e implementación [DILLON01].

Otros algoritmos particionales que intentan minimizar los errores de clasificación, como el algoritmo de *PAM* (*Partition Around Medians*) o *k-medianas* en el que el centro del grupo es representado por el objeto central en el grupo y no por su media. Es un algoritmo menos sensible a ruido pero de procesamiento más costoso.

Algunas alternativas han sido propuestas para solucionar los problemas de k medias, una es la selección de diferentes métricas de error para aumentar su capacidad de detección de grupos. Otra opción que ha sido exitosamente utilizada para disminuir la sensibilidad de la selección inicial de centros es el uso de la lógica difusa para la asignación a grupos no disjuntos, a través de un valor de membresía a cada uno de los grupos, es decir que cada elemento pertenece en un cierto grado a cada uno de los grupos existentes [HOPPNER00]. El algoritmo más popular es *c-medias difusas* que presenta una importante mejora a *k-medias* en la sensibilidad a la selección inicial de centros, aunque también puede caer a mínimos locales. En este tipo de algoritmos el problema más importante es el diseño de la función de pertenencia a cada uno de los grupos basada en una medida de similitud.

1.5.2.3 Métodos basados en densidad

Estos métodos suponen que los datos en cada grupo, son obtenidos de una distribución de probabilidad bien definida [BANFIELD93]. Están diseñados para la búsqueda de grupos con formas arbitrarias y no necesariamente convexas.

Forman cúmulos de datos al hacer crecer un grupo mientras la densidad en una vecindad exceda un valor umbral. Una vez que se tienen estos grupos se analiza para obtener su distribución, que por lo general se supone como una distribución Gaussiana en el caso de datos numéricos, o una distribución multinomial en el caso de variables nominales o categóricas. El algoritmo de esperanza-maximización [DEMPSTER77] basado en la búsqueda de la distribución de máxima verosimilitud a través de la selección de parámetros, es aplicado para la optimización de los *clusters* encontrados.

El algoritmo DBSCAN [ESTER96] (*Density Based Spatial Clustering of Aplications with Noise*) sigue un enfoque similar buscando en la vecindad de cada objeto verificando si esta vecindad contiene un mínimo número de objetos.

Otros algoritmos usados ampliamente para búsqueda de *clusters* basados en densidad incluyen AUTOCLASS [CHEESEMAN96], SNOB [WALLACE94] y MCLUST [FRALEY98], todos orientados a la búsqueda de grupos con base a distribuciones probabilísticas.

1.5.2.4 Métodos basados en modelos

Estos métodos generan modelos que se ajustan a los datos, además de identificar los grupos de datos. Los algoritmos más usados son los árboles de decisión y las redes neuronales.

En los árboles de decisión cada hoja se refiere a un concepto y contiene una descripción probabilística de esta. El algoritmo COBWEB [FISHER87] es el más comúnmente usado enfocado a datos nominales o categóricos, asumiendo independencia de los atributos. CLASSIT es una extensión para datos continuos, pero ninguno es capaz de manejar grandes conjuntos de datos.

Cuando se usan redes neuronales para la identificación de grupos de datos, cada grupo puede ser representado por una neurona o prototipo. Cada conexión neuronal posee un peso ajustado durante el proceso de aprendizaje. El tipo de red neuronal más popular para este tipo de problemas son los mapas auto-organizados propuestos por Kohonen [KOHONEN90] (SOM), que realizan un mapeo del espacio de alta dimensionalidad a un espacio de menor dimensionalidad definido por el número de neuronas a través de los pesos en las mismas. Este algoritmo tiene la desventaja de ser sensible a las condiciones iníciales y requerir del ajuste de parámetros como la tasa de aprendizaje o el radio de vecindad que afecta cada neurona durante el proceso de aprendizaje.

Otro enfoque interesante es el análisis multidimensional a través hipervoxels [KURI10] consistente en la búsqueda de un agrupamiento eficiente sin usar una métrica de distancia tradicional, sino la búsqueda de la combinación más eficiente de los llamados voxels minimizando una medida de calidad de los *cluster*, como la entropía de la teoría estadística de la comunicación, o cualquier modelo que se quiera aplicar por medio de algoritmos evolutivos.

1.5.3 Análisis de grupos en grandes bases de datos

Una buena cantidad de investigación se ha hecho para adaptar algoritmos a grandes bases de datos, teniendo en cuenta las dificultades de análisis y de costos que pueden ocasionar. Este análisis de grandes volúmenes de información es el análisis para el que las técnicas de minería de datos fueron diseñadas, ya que otros tipos de análisis resultan insuficientes e inaplicables.

Como se mencionó en la sección 1.3 el tratamiento en grandes bases de datos de grandes dimensiones prohíbe el uso de algoritmos tradicionales con resultados eficientes. Es por

ello que se han desarrollado métodos con siguientes objetivos que se presentan a continuación.

1.5.3.1 Reducción de dimensionalidad en el conjunto de datos.

Selección de variables

La selección de variables o características se realiza con la finalidad de eliminar variables redundantes o irrelevantes para reducir la dimensionalidad del conjunto de datos que puede llegar a tener decenas, centenas o miles de variables. Esta selección se hace de acuerdo con una ordenación de las variables basada en un criterio de importancia, seleccionando los elementos más relevantes y eliminando los que aportan menos a un análisis dado [GUYON03]. Los métodos de selección de variables más utilizados pueden ser clasificados en 2 tipos [TANG05] .

- De filtro: son independientes del algoritmo de aprendizaje que vaya a ser aplicado
- *Wrappers*: Dependen del objetivo del proceso de minería de datos que se vaya a utilizar.

La selección de variables de tipo de filtro es la herramienta más útil al ser aplicada en un mayor número de casos. Esta reducción puede enfocarse a la búsqueda de correlación lineal entre variables, criterios de información mutua o transformaciones [GUYON03], de este modo se asume que dos variables que están fuertemente correlacionadas no aporten gran cantidad de información en la inferencia. El análisis de componentes principales (*PCA*) busca la dirección de los ejes que contienen mayor cantidad de variaciones para limitar la redundancia presente en el conjunto. Otros métodos realizan proyecciones de los datos en búsqueda de aquella proyección en la cual la redundancia o relevancia sea más clara a través del uso de una medida, que por lo general es una estadística basada en la distribución de los datos proyectados.

Selección de instancias

La selección de instancias se centra en la reducción del número de elementos que se usarán en el proceso de minería, seleccionando un conjunto de datos que represente el comportamiento de la población.

El muestreo estadístico es la selección de instancias más básica, al tratar de obtener un conjunto de objetos que sea representativo de la población eligiendo elementos aleatoriamente. En este caso la representatividad implica que los patrones encontrados en la población y en la muestra serán prácticamente equivalentes. Existen investigaciones

enfocadas al análisis del muestreo en si mismo [OLKEN94] o del cálculo del tamaño de una muestra aleatoria unidimensionalmente a través de las fórmulas de muestreo de Cochran [COCHRAN77] suponiendo distribuciones conocidas [BARLETT01].

Opciones más complejas que el simple muestreo, enfocadas a la reducción de instancias en la minería de datos han sido propuestas aunque presentan deficiencias al depender de otros métodos de aprendizaje o tratar de comprimir la información complicando su uso posterior. Una propuesta para la generación del conjunto reducido es la selección de prototipos a través de los cuales se pueda reproducir la distribución de la población, un ejemplo de esta selección es la propuesta por Bradley, et al. [BRADLEY98] que usa una iteración de *k medias* para identificar prototipos y generar los datos necesarios a través de una distribución de probabilidad alrededor de estos prototipos, siendo una solución poco eficiente al presentar las deficiencias de *k medias* combinada con la dificultad de definir una distribución de los datos alrededor del prototipo. Otro enfoque basado en el agrupamiento jerárquico propone identificar elementos que correspondan a jerarquías superiores en el agrupamiento para utilizarlos en las muestras, aunque alcanzando baja eficiencia.

La selección de instancias por compresión requiere de la declaración de un modelo estadístico para recuperar la información compactada, pero permite la regeneración de los datos sin grandes pérdidas si se logra identificar la distribución adecuada.

1.5.3.2 Optimización de algoritmos

La optimización de algoritmos para su aplicación ha sido ampliamente investigada para solucionar el problema de búsqueda de grupos de datos en grandes bases de datos, aunque no se ha llegado a un algoritmo superior en todos los aspectos [KRIEGEL09]. Se pueden observar diferentes corrientes;

- Búsquedas de grupos en subespacios en vez de analizar todas las dimensiones.
- Optimizaciones en el tiempo de ejecución.
- Adaptaciones para la aplicación en equipos de cómputo de mayor capacidad, permitiendo paralelización.

Búsqueda en subespacios y proyecciones

La selección de variables en bases de datos de grandes dimensiones por métodos como el análisis de componentes principales puede generar problemas en la búsqueda de grupos al guiarse por la relevancia de las variables de manera global, mientras que se puede dar el problema de relevancia local de una variable, de modo que a pesar de que globalmente un

atributo no es suficientemente relevante, lo es en un subconjunto de instancias y dimensiones. Para solucionar este problema se han propuesto distintos algoritmos que pueden agruparse de manera general en 3 grupos [KRIEGEL09]:

- Algoritmos de proyección de *clusters*: asignan cada elemento a un sólo *cluster* buscando en diferentes proyecciones de los datos en vez de usar el espacio original de datos de alta dimensionalidad. Ejemplos de este tipo de algoritmos son:
    - PROCLUS [AGRAWAL99] basado en k-medianas que realiza búsquedas de grupos en muestras de datos. Una variación a este algoritmo es SSPC [YIP2005] que permite la introducción de conocimientos del dominio.
    - PreDeCon [BOHM04] aplica DBSCAN utilizando una medida de distancia que explora la existencia de grupos en subespacios.
    - CLTree [LIU00] basado en la construcción de arboles de decisión en forma similar a la que se usa en un aprendizaje supervisado, agregando datos uniformemente dispersos y clasificando los nodos en zonas densas (*clusters*) o vacías (dispersas), permitiendo explorar subespacios dentro de las zonas densas. El criterio de separación de nodos puede ser costoso.

- Algoritmos de proyección "suave": buscan la asignación óptima de los elementos a *k-clusters* asignados por medio de un valor de pertenencia a cada uno de los *clusters* en diferentes proyecciones. Algunos algoritmos de este tipo son :
    - LAC (*Locally adaptive Clustering*) [DOMENICONI07], aproxima el agrupamiento por medio de k gaussianas de parámetros adaptables
    - COSA [FRIEDMAN04] genera una matriz de similitud que puede ser utilizada para aplicar agrupamientos posteriormente, indicando el subespacio al que corresponde un elemento de manera similar a PreDeCon a través de medidas de distancia
- Algoritmos de agrupamiento en subespacios: Analizan los grupos en subespacios de menor dimensionalidad, pudiendo generar una guía para la búsqueda en subespacios de mayor dimensionalidad. Ejemplos de estos son
    - CLIQUE [AGRAWAL98] con una búsqueda de grupos orientada a mallas, analiza el conjunto de elementos en busca de celdas con un número de elementos superior a un mínimo dado como parámetro, para considerar que está densamente poblada. Se analiza en subespacios de menor dimensionalidad y se va aumentando gradualmente la dimensionalidad para identificar el subespacio inmediato superior en el que la celda dejo de ser densa. Otras variaciones de CLIQUE se han planteado, como ENCLUS [CHENG99] que busca la existencia de más de un grupo en cada celda, o MAFIA [NAGESH01] propone celdas de tamaño adaptable.

    - SUBCLU [KAILING04] utiliza el modelo DBSCAN en diferentes subespacios obteniendo un número mayor de grupos con formas arbitrarias aprovechando el conocimiento de la densidad en subespacios superiores, aunque aumentando el tiempo de ejecución. Mejoras para evitar un sesgo en ciertas dimensiones son propuestas en el algoritmo DUSC [ASSENT07], aunque causan pérdida de monotonicidad en el proceso.
- Algoritmos híbridos: Algoritmos que buscan en ciertos subespacios de interés basados en el uso combinado de algoritmo de búsqueda y optimización combinatoria. Algunos algoritmos de este tipo son [KRIEGEL09]:
    - DOC y MINECLUS buscan a través de hipercubos en diferentes dimensiones o regiones densas.
    - DiSH usa el mismo enfoque que PreDeCon para hacer agrupamiento jerárquico,
    - FIRES busca *clusters* en cada dimensión para después buscar *clusters* de grupos. P3C analiza intervalos densos en cada dimensión para unir grupos en dimensiones superiores.

Optimización del tiempo de procesamiento

Para reducir el tiempo de procesamiento se propone la descomposición del conjunto de datos en subconjuntos o el análisis incremental de la población, tratando de conservar una complejidad lineal sin hacer uso de gran cantidad de memoria.

CURE [GUHA98] es un algoritmo que intenta por una parte encontrar grupos de formas arbitrarias y por otra solucionar problemas de volumen de información a través de un muestreo aleatorio que crea particiones en el conjunto de datos asumiendo el número de grupos y una aproximación del número de elementos en ellos para después unir los grupos encontrados en las diferentes muestras.

CLARANS (*Clustering Large Applications based on RANdom Search*)[NG94] usa muestreo aleatorio para identificar centroides de los datos y posteriormente unir los centroides encontrados en diferentes muestras .

BIRCH (*Balanced Iterative Reducing and Clustering*) [ZHANG96], hace uso de estructuras de datos como árboles dinámicos para almacenar la información de grupos candidatos permitiendo reducir el tiempo de procesamiento en la búsqueda de jerarquías.

Optimizaciones aplicando la técnica de divide-y-vencerás permiten también optimizar los accesos a memoria secundaria [JAGADISH01] o para aplicar algoritmos jerárquicos por secciones de la población [CHENG06].

A pesar del desarrollo de estos algoritmos adaptados para el manejo de cantidades mayores de datos, es común el uso de algoritmos particionales tradicionales como *k-medias*, *c-medias difusas* o mapas auto-organizados [ROKACH05], aprovechando que tienen una complejidad lineal en el número de elementos e iteraciones, a pesar de que no permita obtener grupos con formas irregulares como otros algoritmos basados en densidad.

Capítulo 2. Marco teórico

En este trabajo se manejan un conjunto de conceptos y técnicas correspondientes a diferentes áreas de estudio. En este capítulo se presentarán los conceptos básicos usados en el desarrollo de la metodología para mayor claridad en su presentación en los siguientes capítulos.

2.1 Bases de datos y el descubrimiento de conocimiento (*KDD*).

Una base de datos es una colección de elementos, almacenados de manera organizada y estructurada para su posterior uso. Las bases de datos almacenadas digitalmente han permitido el almacenamiento masivo de la información con fines comerciales, científicos e incluso personales.

Debido a que en la época actual acumular datos es cada vez más sencillo, es cada vez más común encontrar grandes repositorios de datos. Desafortunadamente al aumentar el volumen almacenado de información, la capacidad de extraer el significado no aumenta al mismo ritmo [FAYAD96], generando la necesidad del proceso de descubrimiento de conocimiento en bases de datos y las técnicas de minería de datos.

El proceso de KDD debe ser iterativo y retroalimentado y no existe una metodología bien establecida a seguir para extraer patrones interesantes de la información ya que es un problema de alta complejidad y dependiente de los objetivos del análisis y los datos disponibles, entre otros factores. En la figura 2.1 se muestran las diferentes etapas de KDD.

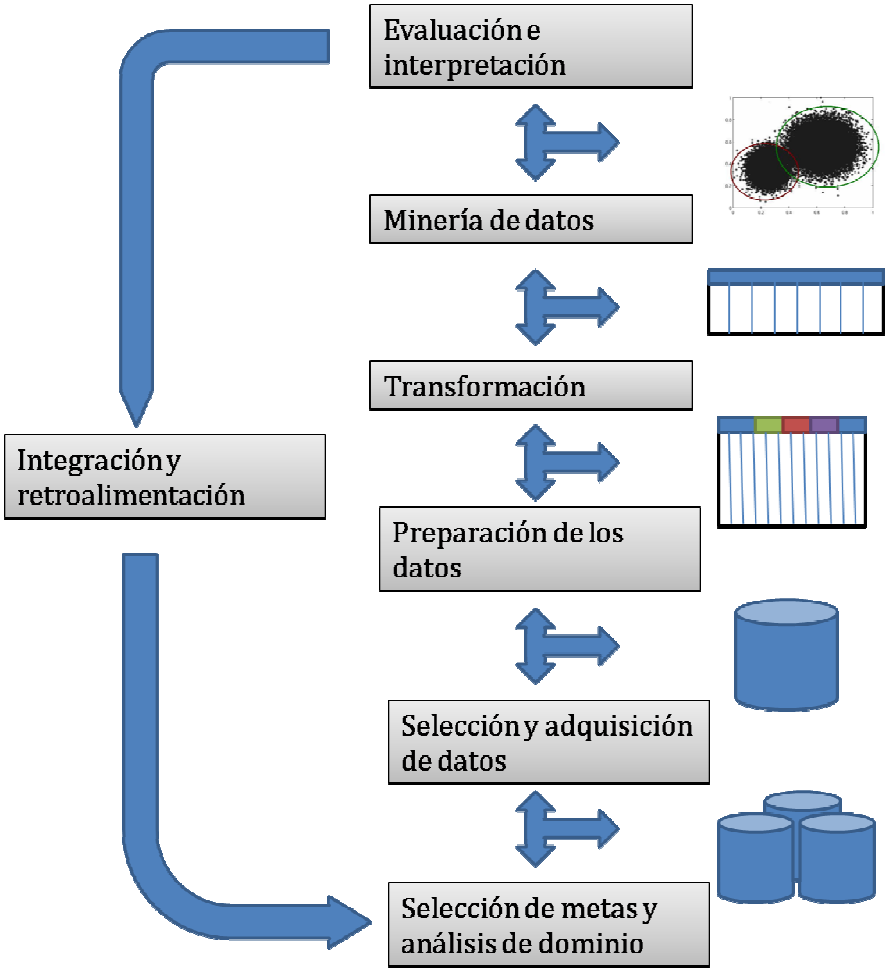

Figura 2.1. Diagrama del proceso de descubrimiento de conocimiento en bases de datos

Este proceso consiste en:

1- Análisis del dominio y establecimiento de metas.
   Consiste en el análisis de la información disponible, selección de algoritmos y representación de los objetivos por alcanzar.
2- Selección y adquisición de datos.
   Selección de la información utilizable o adquisición de ella, tratando de seleccionar sólo la información más relevante para no complicar el proceso
3- Preparación de los datos.
   Etapa de preparación de la información para depurar posibles errores presentes en el conjunto de datos, corrigiendo detalles que pueden ser deducidos, como la detección de anomalías, normalización de los datos, deducción o eliminación de elementos con valores faltantes, entre otros.
4- Transformación
   Consiste en las optimizaciones para tener un comportamiento deseable en el análisis.

Puede consistir en la aplicación de funciones en los datos para hacer cambios de tipo de variable, discretización y normalización. Esta etapa incluye también reducciones de dimensionalidad por selección de variables o por selección de instancias.

5- Minería de datos
Esta etapa consiste en la selección de tipo de algoritmos que se va a aplicar, del algoritmo específico y de los parámetros que se vayan a utilizar en él, para finalmente aplicar los algoritmos de minería. En algunos casos cada una de estas acciones son tomadas como pasos independientes en el proceso de KDD

6- Evaluación e interpretación de resultados
Su objetivo es la evaluación de los resultados analizando su exactitud y utilidad para posteriormente analizarlos en búsqueda de una interpretación. Esta interpretación es el verdadero resultado del proceso de minería en el que se puede obtener información valiosa y aprovechable en la práctica, pudiendo ser el soporte para la toma de decisiones, detección de errores, cualidades, etc.

7- Integración y retroalimentación del conocimiento obtenido.
Los resultados obtenidos pueden ser de utilidad para futuros análisis o indicar las correcciones que se deban hacer en el ciclo del descubrimiento de conocimiento. Estos resultados pueden integrarse en un proceso de mayores dimensiones y alcances, o bien permitir la depuración del proceso para refinar y aumentar los alcances del análisis. Esta es una etapa que en ocasiones no es incluida en el proceso, pero es importante al permitir la integración del análisis a sistemas de mayores dimensiones e importancia. Esta integración directa o automática entre los resultados y sistemas de toma de decisión no es sencilla, por las dificultades que representan el tiempo de análisis y la complejidad de la interpretación de resultados.

2.3 Minería de datos

Los métodos de minería de datos pueden ser utilizados con diferentes objetivos, pero fundamentalmente se pueden definir dos tipos principales:

- Métodos orientados a la verificación de hipótesis
- Métodos orientados al descubrimiento de patrones y reglas de manera autónoma.

Los métodos orientados al descubrimiento de patrones son los más interesantes en general y pueden requerir de un análisis más cuidadoso. Estos a su vez pueden dividirse en métodos descriptivos o métodos predictivos.

Los métodos predictivos tienen la función de crear un modelo que permita prever valores de ciertas variables a partir de elementos que no se conocieron previamente. Los métodos

descriptivos tienen una función de ayudar a la interpretación, visualización o simplificación de las relaciones presentes en las variables.

Estos métodos están basados en teoría estadística o en aprendizaje de máquinas. El aprendizaje de máquinas es el proceso de extraer patrones de la información para generar modelos. El aprendizaje supervisado consiste en la búsqueda de modelos que representen la relación entre un conjunto de variables de entrada y un atributo, como los algoritmos de clasificación y regresión. El aprendizaje no supervisado agrupa los elementos y muestra relaciones entre variables en conjuntos de datos, en los cuales no se especifica una variable dependiente, como en el caso de los algoritmos de agrupamiento.

La minería de datos y en general el proceso descubrimiento de conocimiento se integran con las tecnologías de la información para formar un sistema de soporte de decisiones ya que los resultados obtenidos pueden respaldar hipótesis o crear importantes sugerencias en algún servicio o proceso. Es por ello que comúnmente se mezcla a la minería de datos con los servicios de almacenes de datos, como OLAP (On Line Analytical Processing) que incluye análisis de nivel básico, como reportes e informes de la información contenida en las tablas de la base de datos, y de nivel intermedio como consultas que requieren de un análisis en varios niveles de las bases de datos, sin llegar a necesitar de la minería de datos, sin embargo los alcances y la complejidad del análisis de minería de datos son significativamente más altos, aunque los resultados obtenidos directamente a través de OLAP pueden ser más útiles en casos comunes de necesidad de información. Es por ello que es importante que el usuario de la información posea la capacidad de definir el problema que desea solucionar para usar una herramienta eficiente y adecuada para obtener buenos resultados.

2.4 Reducción de dimensionalidad

La reducción de dimensionalidad es el proceso que permite expresar las características del conjunto de datos a través de un conjunto de elementos de menor tamaño. Esto se puede lograr eliminando la información redundante presente en los datos. Por ello es considerado un proceso de compresión, ya que logra expresar prácticamente el mismo sentido a través de un conjunto de elementos de menor tamaño sin tener pérdidas significativas. Si las características de interés están bien definidas entonces se puede realizar la compresión procurando que no se tengan pérdidas con respecto a esta característica.

La reducción de dimensionalidad se puede hacer reduciendo el número de variables, llamada selección de características o reducción de número de instancias, como se presentó en la sección 1.5.3.

La reducción por selección de características consiste en lo siguiente: si en el conjunto de datos compuesto por un conjunto de elementos vectoriales de *m* dimensiones, se puede encontrar un subconjunto con *c* dimensiones que mantiene la información presente en el conjunto original, entonces el análisis se puede hacer sólo en las *c* dimensiones. En ocasiones el conjunto de datos se le pueden aplicar transformaciones previas para una identificación más sencilla de las variables o características redundantes. Este análisis de redundancia o importancia de variables ha sido ampliamente estudiado y las condiciones y consecuencias de su uso son bien conocidas [YU04] y por lo general permiten buenos resultados. Un procedimiento bien probado y aplicable a cualquier conjunto numérico es el análisis correlacional. El análisis de correlaciones es directamente aplicable en el conjunto de datos y conserva los datos en el espacio original sin aplicar transformaciones que puedan complicar la interpretación de los resultados en el proceso y alcanzando reducciones significativas en el conjunto de variables necesarias para análisis posteriores. Esta selección se basa en la eliminación de aquellas variables que están altamente correlacionadas entre sí. Esta correlación lineal entre dos variables *x* y *y* puede ser estimada a través del coeficiente de correlación lineal de Pearson definido como

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Donde

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

el valor de r_{xy} varía entre -1 y 1 expresando el grado de dependencia lineal entre ambas variables, de modo que una correlación cercana a -1 indica una fuerte correlación negativa, es decir que una variación en X está relacionada con una variación de la misma magnitud en Y pero de signo contrario; mientras que $r_{xy} = 1$ indica que una variación en X está relacionada con una variación de la misma magnitud en Y. Si un conjunto de variables presenta una fuerte correlación lineal entre sí, entonces puede seleccionarse sólo una de ellas para el análisis de los datos sin grandes pérdidas, ya que la información aportada por el resto de las variables ya es utilizada por la variable seleccionada.

2.5 Algoritmos de agrupamiento

Los algoritmos de agrupamiento son fundamentales en la minería de datos. Su principal finalidad es descriptiva descubriendo grupos de interés que pueden llegar a dar lugar a

predicciones de características de elementos, la aplicación eficiente de otros algoritmos predictivos o estadísticos como muestreo por conglomerados, además de ayudar a la visualización del conjunto de datos.

Los grupos de datos son subconjuntos de instancias en los que los elementos dentro de los grupos tienen características similares y mantienen diferencias notables con los elementos de otros grupos. Un algoritmo de búsqueda de grupos se define como sigue.

Sea X un conjunto de n datos x_i, tal que $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ está definido en un espacio m-dimensional $\mathcal{L}$ y c_j uno de k grupos de datos, con $2 \leq k \leq n$. Se pueden definir las funciones

$$f_j(x_i) = u_j$$

que es una función de membresía del elemento x_i al grupo c_j que maximiza la similitud entre los elementos del mismo grupo y las diferencias de los elementos de otros grupos.

Existen una gran cantidad de algoritmos de búsqueda de grupos (véase sección 1.5.2) ya que aun suponiendo que se conozca el número de grupos en la población es un problema *NP-completo* [SLAGLE75] ya que requiere de la exploración de todas las combinaciones de elementos en todos los posibles subespacios para la definición de una función de membresía óptima. Los algoritmos más comúnmente usados son aquellos que poseen una complejidad que no alcance a ser cuadrática sobre el número de datos por procesar, como *k-medias*, *c-medias difusas* o mapas auto-organizados de Kohonen, que además son interpretables e implementables de manera relativamente sencilla. Las pruebas realizadas en este trabajo son hechas usando estos algoritmos por esa razón.

Estos métodos particionales que buscan puntos representativos en el espacio $\mathcal{L}$ son considerados de cuantización vectorial, es decir que buscan el describir una población a través de ciertos objetos centrales o prototipos que pueden describir al resto de los individuos, siendo una forma de compresión de información. Se considera una compresión en este contexto, a una forma reducida de expresar un conjunto de datos y está relacionado con la forma en la que los datos son codificados. En este caso los individuos llamados prototipos, en conjunto con una distribución de probabilidad supuesta alrededor de ellos permiten expresar con sólo un dato y algunos parámetros, un conjunto de elementos de mucho mayores dimensiones aunque con cierta pérdida. Estos métodos suponen que se pueden encontrar estos objetos centrales y estos representarán de forma eficiente a un grupo y a los datos que contiene. Desafortunadamente esta idea por lo general limita a la búsqueda de grupos con formas regulares, por lo general híper-esféricos.

Los algoritmos de agrupamiento buscan la similitud de los objetos a través de una medida de distancia que permita asociarla la distribución de los datos en un espacio n-dimensional, asumiendo que si las variables elegidas para el proceso son de utilidad, esta cercanía representa una relación implícita entre los objetos ya que comparten ciertas características.

Diferentes medidas de distancia pueden ser utilizadas, dependiendo de los resultados esperados en el análisis, las medidas de distancia generan grupos con formas esféricas en su propia métrica, por lo que no siempre son óptimos. Algunas medidas de distancia pueden aumentar la sensibilidad a la separación entre los datos al basarse en medidas estadísticas, como por ejemplo la distancia de Mahalanobis [MAHALANOBIS36]. Las medidas de distancia más utilizadas entre dos instancias son las medidas de Minkowski [HAM01] definidas de la siguiente manera para dos vectores $x = \{x_1 x_2, \dots, x_n\}$ y $y = \{y_1 y_2, \dots, y_n\}$:

$$d(x, y) = (|x_1 - y_1|^g + |x_2 - y_2|^g + \cdots + |x_n - y_n|^g)^{1/g}$$

La distancia euclidiana es obtenida cuando g=2. Cuando g=1 llamada la medida de Manhatan se obtiene la suma de las diferencias absolutas entre dos elementos y cuando $g = \infty$ se obtiene la máxima diferencia entre todas las dimensiones.

Otra medidas de similitud comunes son el coseno del ángulo entre dos vectores y la correlación de Pearson antes presentada. Estas distancias permiten el análisis de la calidad interna de los grupos obtenidos por los algoritmos de agrupamiento.

2.5.1 *C-medias difusas*

El algoritmo de c-medias difusas es un algoritmo particional de agrupamiento de datos basado en *k-medias* que utiliza lógica difusa para asignar los elementos a diferentes grupos en cierta medida, a diferencia de la asignación tradicional en la que cada instancia pertenece a un solo grupo en todo el proceso de agrupamiento [DUNN73]. El resultado es un algoritmo menos sensible a los valores iniciales que *k medias* y con mayor flexibilidad al definir una medida de pertenencia a cada grupo, que es tomado como un conjunto difuso en este contexto,

La teoría de conjuntos difusos fue introducida por Lofti Zadeh [ZADEH65] tratando de presentar una generalización a la teoría clásica de conjuntos en la que un objeto simplemente pertenece o no pertenece a un conjunto. Usando teorías de conjuntos difusos un objeto x puede pertenecer a un conjunto en menor o mayor medida dependiendo del valor del grado de membresía $u_k(x)$ al conjunto k. La membresía en este caso puede estar relacionada con una medida de similitud.

En el algoritmo de c-medias difusas, se define una matriz de membresías U de $n \times c$ donde n es el número de vectores x_i correspondientes a las instancias en la base de datos y c el número de grupos buscados. Cada elemento de la matriz U, $u_{i,j}$ corresponde al valor de la membresía del vector i del conjunto al grupo j. El valor de esta matriz se va alterando durante el algoritmo para finalmente resultar en un conjunto de membresías que permite analizar la pertenencia de cada elemento a los grupos encontrados. Este algoritmo además requiere de establecer un parámetro $m \geq 1$ que indicará el grado de borrosidad de los conjuntos formados, donde $m = 1$ corresponde a un agrupamiento nítido como el obtenido al usar *k-medias*, y $m \to \infty$ indica que el grado de membresía de cada elemento será $\frac{1}{c}$, es decir que pertenecerá en igual grado a todos los grupos, siendo totalmente borroso.

El algoritmo consiste en las siguientes etapas
1.-Se inicializa U asignando aleatoriamente la membresía de cada elemento a todos los *clusters*, de manera que

$$\sum_{j=1}^{k} u_{ij} = 1 \ \forall i \ y \ U_{kj} \in [0,1] \ \forall k, l$$

2.-Se calculan los valores de cada centro

$$c_j = \frac{\sum_{i=1}^{n} {u_{ij}}^m \cdot x_j}{\sum_{i=1}^{n} {u_{ij}}^m}, \forall j$$

3.-Se recalcula el grado de membresía a los grupos a través de la siguiente función

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d(x_i, c_j)}{d(x_i, c_k)})^{\frac{2}{m-1}}}$$

4.-Repetir el paso 2 en caso de que los *grupos* asignados sean suficientemente diferentes a los asignados en la iteración anterior, es decir

$$\|U - U_{anterior}\| < \epsilon$$

Donde ϵ es un valor pequeño asignado por el usuario como criterio de paro.

2.5.2 Mapas auto-organizados

Los mapas auto-organizados son una clase especial de redes neuronales, basados en el aprendizaje competitivo para producir un mapeo de información de alta dimensionalidad a un espacio discreto de salida de las neuronas [HAYKIN99] y ha sido exitosamente utilizado como herramienta de agrupamiento en diferentes áreas. Está compuesto por dos capas de neuronas, una capa de entrada y una de salida. La capa de entrada posee una neurona por cada variable de entrada y cada una de ellas está conectada a todas las neuronas de la capa

de salida. La capa de salida es organizada como un mapa generalmente bidimiensional o tridimensional fácilmente visualizable.

La capa de entrada está compuesta por m neuronas, que reciben los vectores de entrada $x_1, x_2, \dots x_m$ donde m es el número de variables de entrada. En la capa de salida cada neurona está conectada a todas las neuronas de entrada por lo que los pesos de sus conexiones W generan coordenadas en el espacio de los vectores de entrada. La capa de salida consiste en l neuronas comúnmente distribuidas en dos o tres dimensiones. El caso más común es el bidimensional con $a \times b$ neuronas donde cada neurona tiene asociada una coordenada de m dimensiones W_{ij}. Estas coordenadas o pesos de conexiones se ajustarán durante el entrenamiento de la red.

2.5.2.1 Entrenamiento de la red

Las redes neuronales requieren de una etapa de entrenamiento en la que ajustan sus parámetros para después ser usadas como modelo. Este entrenamiento consiste en la presentación de datos a la red con la finalidad de que ajuste los pesos de las conexiones conforme se va presentando la información nueva por medio de los datos. Este entrenamiento se puede repetir varias veces para asegurar la convergencia a un conjunto de pesos útiles en la red. Cada vez que se presentan todos los datos de entrenamiento a la red para su aprendizaje se denomina una época.

Existen distintos algoritmos de aprendizaje en redes neuronales pero el más utilizado en el caso de los mapas auto-organizados es el aprendizaje competitivo. El algoritmo es presentado a continuación:

1. Se inicializan todos los pesos de las conexiones a las l neuronas de la capa de salida con valores aleatorios.
2. Se toma un elemento x del espacio de entradas, que se presentará como un patrón de activación a las neuronas.
3. Se busca la neurona con coordenadas más cercanas al vector presentado a través de una medida de distancia.
$$i(x) = \min\big(d(x, w_j)\big), j = 1,2, \dots, l$$
4. Los pesos de las w_j conexiones son actualizados de acuerdo a lo siguiente
$$w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(x(n) - w_j(n))$$
donde $\eta(n)$ es la tasa de aprendizaje en la iteración n, que establecerá la velocidad del aprendizaje y es comúnmente inferior a 0.01 para evitar inestabilidad en el entrenamiento.
$h_{j,i(x)}(n)$ es la función de vecindad que define la influencia que tendrá la neurona ganadora $i(x)$ sobre la neurona j. Es una función centrada en la

neurona ganadora generalmente definida por:

$$h_{j,i(x)}(n) = e^{-\frac{d(i(x),j)^2}{\sigma^2}}$$

donde σ es el radio de aprendizaje.

5. Regresar al paso 2 hasta que se terminen los datos de entrada.
6. Se reduce el valor de la tasa de aprendizaje y el radio de aprendizaje a través de un factor de aprendizaje f_η y factor radial f_σ, ambos de valor muy cercano a 1.

$$\eta(n+1) = f_\eta \eta(n)$$
$$\sigma(n+1) = f_\sigma \sigma(n)$$

7. Se repite el algoritmo procesando de nuevo la población hasta que se alcance el número de épocas establecido.

2.5.2.2 Propiedades de los mapas auto-organizados

El mapa de características generado en la capa de salida con sus respectivas conexiones tiene las siguientes características importantes [HAYKIN99]:

El mapeo producido por los pesos de las neuronas produce una relación que tiene como dominio el espacio de las neuronas y mapea de manera eficiente el espacio de entrada. Esto indica que el mapeo apunta a lugares representativos en el espacio de entada pudiendo ser estos los centros de un grupo de datos

- El mapa formado en la capa de salida se encuentra topológicamente ordenado en el sentido que la localización de la neurona en la capa de salida está relacionada con un dominio o característica en el espacio de entrada.
- El mapa de características refleja las variaciones estadísticas en la distribución de la población de entrada por lo que regiones densamente pobladas en el espacio de entradas, serán representadas por dominios más amplios en la capa de salida.
- Dados los datos distribuidos de manera no lineal en el espacio de entrada, el mapa realiza una selección de las características para aproximar la distribución.

La segunda propiedad es la que le da el nombre de mapa auto organizado. La auto-organización implica la generación de un efecto que aumenta la complejidad interna de proceso sin ser guiado por factores externos, es decir que por la organización interna del sistema se generan efectos independientes de las funciones que originalmente tenía cada individuo. En este caso las neuronas cercanas mapean dominios cercanos en el espacio de entrada sin ser esta una función explícita en el entrenamiento.

2.6 Análisis Estadístico y la minería de datos

La estadística es un campo que comparte ciertas funciones con la minería de datos, pero está enfocada a conjuntos de datos de tamaño menor en los que se pueden hacer suposiciones, o inferencias con bases teóricas que permiten acotar los posibles resultados. Aunque la estadística permite el manejo de errores e incertidumbre, estos están fuertemente acotados por límites teóricos, mientras que en la minería de datos en muchas ocasiones se realizan búsquedas sin una guía que permita acotar a priori el error al que se pueda llegar; a pesar de estas diferencias, importantes avances en el área de aprendizaje de máquinas son derivados del análisis estadístico, permitiendo analizar a detalle la información disponible y los resultados alcanzables en un proceso de aprendizaje. El manejo de alta incertidumbre en la minería de datos no es un gran defecto ya que permite obtener modelos con buena capacidad de generalización a pesar de los problemas impuestos por la alta dimensionalidad o la heterogeneidad en los datos.

Conceptos de estadística como la teoría del muestreo o las pruebas estadísticas de hipótesis pueden ser aplicadas en el proceso de minería para un adecuado preprocesamiento o validación de procesos. La teoría del muestreo consiste en los métodos para asegurar la representatividad de una muestra a través del uso de una estadística.

Una estadística es una función de la muestra y es considerada una forma de compresión de la información [CASSELA02] ya que su valor indica de manera compacta características distribuidas en los valores de la población. La selección de una estadística para el análisis estadístico es la base sobre la cual se generarán inferencias sobre parámetros, distribuciones, hipótesis, etc.
El muestreo más utilizado y sencillo es el muestreo aleatorio simple ya que no requiere tener conocimientos sobre distribución de los datos y el obtener muestras sólo requiere de obtener un número dado de instancias aleatoriamente sin otro tipo de procesamiento. Muestreos que van adaptando la probabilidad de selección de algunos elementos han mostrado alta eficiencia en el muestreo, al modificar esa probabilidad con base a la distribución de los elementos que se van muestreando iterativamente. Un ejemplo de este muestreo es el algoritmo de Metropolis basado en métodos de Monte Carlo que puede ser aplicado en grandes conjuntos unidimensionales [GUO04]. Desafortunadamente estos métodos no se adaptan fácilmente para el muestreo de poblaciones con un gran número de variables al tener que manejar distribuciones multidimensionales.

En el caso de volúmenes mayores de información en los que no se tiene noción de la distribución de los datos, es necesario hacer pruebas que no dependan de una distribución conocida. Este enfoque es basado en las pruebas estadísticas no paramétricas y el muestreo

implementado a partir de las pruebas es llamado muestreo orientado a potencia de una prueba no paramétrica.

Un análisis no paramétrico se logra haciendo modificaciones en la forma en que son analizados los datos que por lo general son sustituidos por valores de rangos o por su distribución empírica. La distribución empírica de un conjunto de n datos $x_1, x_2, \dots, x_n$ se define como:

$$\widehat{F_n}(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

donde

$$I(X_i < x) = \begin{cases} 1 \; si \; X_i \leq x \\ 0 \; si \; X_i > x \end{cases}$$

Los rangos son la posición que le corresponde a un dato después de ser ordenados. El manejar rangos en vez de los datos originales genera una pérdida importante de información pero a cambio permite saber con certeza, características de la distribución de los datos ya que estos rangos deben estar distribuidos de manera uniforme al tratarse de números consecutivos.

2.6.1 Pruebas estadísticas

Una prueba estadística es una función de decisión que permite rechazar una hipótesis si se encuentra suficiente evidencia de que esta hipótesis es falsa. La hipótesis que se supone inicialmente cierta H_0 es llamada hipótesis nula.

Existen dos tipos de errores en las pruebas estadísticas. El error tipo I es aquel que se comete en la prueba de hipótesis cuando la hipótesis nula es falsa pero no se rechaza; el error de tipo II se da al rechazar una hipótesis nula que es verdadera.

El valor p es otro valor de alta importancia al usar las pruebas estadísticas, ya que permite no solo aceptar o rechazar una hipótesis sino conocer la probabilidad con la que se está aceptando o rechazando en vista de la evidencia. El valor p es definido como la probabilidad de observar un valor de la estadística de prueba utilizada si la hipótesis nula fuera verdadera y si el valor p obtenido es menor que un cierto umbral la prueba rechaza la hipótesis nula. Un valor de 0.05 es generalmente tomado como el valor p mínimo para no rechazar H_0.
A continuación se presentan algunas de las más utilizadas pruebas estadísticas no paramétricas

2.6.1.1 Kruskal-Wallis

La prueba de Kruskal-Wallis [KRUSKAL52] es una prueba para verificar la equivalencia en un conjunto de muestras a través de las variaciones en las medias de cada una de ellas. Esta es una variación de la prueba paramétrica del análisis de varianzas para la verificación de distribuciones normales (ANOVA[2])

Como otras pruebas no paramétricas, la prueba de Kruskal-Wallis requiere de la ordenación de los datos para procesar los rangos en vez de los datos en sí. La prueba consiste en el cálculo de la estadística *H* definida por:

$$H = \frac{N-1}{N}\sum_{i=1}^{k} n_i \frac{\left[\bar{R}_\iota - {}^1\!/_2\,(N-1)\right]^2}{\frac{N^2-1}{12}}$$

donde N es el número total de datos, k es el número de muestras, n_i es el número de datos en la muestra *i* y $\bar{R}_\iota$ es el promedio de los valores de los rangos en la muestra i. Esta estadística tiene una distribución χ^2_{k-1}, chi cuadrada con k-*1* grados de libertad, lo cual permite obtener una probabilidad de que las muestras provengan de la misma población con base a las variaciones de las medianas.

La suposición de la distribución de *H* es válida dado que no se manejan los datos directamente, sino sus rangos, que tienen una distribución uniforme ya que se trata de una secuencia de valores correspondientes a la posición del dato en el conjunto ordenado.

2.6.1.2 Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov es una prueba estadística de bondad de ajuste, es decir que evalúa la calidad de la aproximación de la distribución de los datos a través de un modelo dado. Esto se evalúa a través del estadístico:

$$D = \sup_{1 \le i \le N} |\hat{F}\,(x_i) - F_0(x_i)1|$$

donde x_i es el i-*ésimo* valor observado en la muestra de datos ordenados, $\hat{F}(x_i)$ es la función de distribución empírica y $F_0(x_i)$ es la función de distribución acumulada. Esta estadística es una medida de la máxima diferencia entre dos funciones de distribución acumulada.

La probabilidad de que exista una diferencia significativa entre la distribución empírica y la teórica se limita por medio de desigualdad de Dvoretzky–Kiefer–Wolfowitz aplicable a distribuciones continuas:

[2] ANOVA: Analysis of Variance

$$P(D > \epsilon) \leq 2e^{-2n\,\epsilon^2} \quad \forall\, \epsilon > 0$$

Esta desigualdad permite además calcular el tamaño de la muestra que asegura que la diferencia entre la distribución empírica de la muestra y la población no sea mayor a ϵ con respecto a la distribución de la población.

2.6.1.3 Prueba de bondad de ajuste χ^2 y prueba de χ^2 para dos muestras.

La prueba de bondad de ajuste de χ^2 analiza la diferencia entre una distribución de los datos y un modelo que se ajuste a dicha distribución. Para esta prueba es necesario dividir el espacio de posibles valores de la variable analizada x_i en m intervalos, de modo que permita calcular la frecuencia de datos en cada uno de estos intervalos tanto en los datos manejados como en los que se obtendrían a través del modelo. La estadística de prueba se define como

$$X^2 = \sum_{i=1}^{m} (O_i - E_i)^2$$

Donde X^2 es la estadística de prueba cuya distribución se aproxima asintóticamente a la distribución χ^2 con m grados de libertad, E_i y O_i son la frecuencia esperada y la frecuencia observada respectivamente en el intervalo *i*.

La aproximación de la distribución de esta estadística de prueba es eficiente sólo si la diferencia en las frecuencias en cada intervalo tiene una distribución normal por lo que es necesario que el número de elementos en cada intervalo contenga al menos 5 elementos. La longitud de los intervalos no tiene que ser la misma para todos los intervalos

De acuerdo a esta prueba si el modelo representa un buen ajuste a la distribución de los datos entonces la distribución de las diferencias en intervalos debe ser normal con media cero y varianza 1.

Existe una variación de esta prueba para comparar la distribución de dos muestras en vez de compararla con un modelo dado en la que la estadística de prueba es:

$$X^2 = \sum_{i=1}^{m} \frac{(KR_i - S_i/K)}{R_i + S_i}$$

En la que

$$K = \sqrt{\frac{\sum_{i=1}^{m} S_i}{R_i}}$$

Donde R_i es la frecuencia en el intervalo i en la primera muestra, S_i es la frecuencia para la segunda muestra. Se aplica la misma aproximación a la distribución de la estadística que en el caso de una muestra a través de la distribución χ^2.

2.7 Teoría estadística de la comunicación

La teoría estadística establece las medidas necesarias para cuantificar la información transmitida en un proceso de comunicación si esta está contenida en un mensaje codificable a través de un conjunto de símbolos. La información de acuerdo a la teoría estadística de la comunicación establecida por Claude Shannon [SHANNON48] puede ser medida a través de la incertidumbre implícita en un mensaje.

De acuerdo con Shannon un mensaje es una secuencia de símbolos que son generados en una fuente de información y que a través de un canal de comunicación llega a un receptor. En esta teoría no se busca la información en el sentido semántico, ya que no se asocia el contenido del mensaje con ningún mapeo a conceptos o significados. La información contenida en el mensaje es medida a través de los símbolos en los que está codificado el mensaje, de modo que la información contenida en un símbolo es dada por la sorpresa asociada con la aparición de ese símbolo. Esta información contenida en el símbolo S_i es definida como

$$I(S_i) = log(\frac{1}{p_i})$$

donde p_i es la probabilidad de transmitir el símbolo S_i en el mensaje.
La cantidad de información esperada en cada símbolo de un mensaje formado por un alfabeto Σ con n símbolos es definida por la entropía:

$$H(X) = -\sum_{i=1}^{n} p_i \ log(p_i)$$

Debido a que la información que se maneja está codificada en la base de datos esta teoría permite calcular una aproximación a la cantidad de información que realmente se puede obtener en el proceso de minería sobre ese conjunto de datos.

El valor de la entropía está relacionado con la longitud promedio de la codificación del conjunto de símbolos que pueden expresar de manera óptima un mensaje. Se entiende por óptima la codificación en la que el mensaje tiene una longitud mínima.

La codificación óptima se alcanza cuando se asigna una longitud mínima a los elementos que aparecen un mayor número de ocasiones y una longitud mayor a los símbolos menos frecuentes en el mensaje.

2.8 Simulación

La simulación ha sido una herramienta fundamental para la búsqueda de parámetros o distribución de funciones que no pueden ser resueltas teóricamente por su complejidad. Básicamente consiste en la creación de un modelo que permita imitar el funcionamiento de un sistema real para experimentar con él y obtener sus características con fines descriptivos o predictivos.

Una aplicación común de las simulaciones es la simulación estadística en los casos en los que la distribución de una población o de una estadística no puede ser obtenida por un análisis teórico tradicional o que sólo llevan a aproximaciones que carecen de precisión para un caso dado. Dado que en ocasiones el modelo a reproducir en una población para el análisis estadístico, es simplemente el muestreo, la simulación en estos casos consiste en la toma iterativa de muestras de la población que está siendo analizada. El conjunto de estadísticas obtenidas de cada muestra, por sí mismas, no suelen ser representativas de la población; sin embargo, en promedio al repetir suficientes veces el muestreo y el cálculo del estadístico, por ley de los números grandes, se puede llegar a obtener un parámetro de la población. Es decir que tomando M_1,M_2,...,M_B muestras de una distibución G, dada una estadística $T(M)$, se tiene que:

$$\overline{T(M)} = \lim_{B\to\infty} \frac{\sum_{i=1}^{B} T(M_i)}{B} \to \int T(m)dG(m) = E[T(Y)]$$

Este valor es llamado estimador de Monte Carlo puro o crudo [ZHU05]. Por este método se pueden obtener buenas aproximaciones de parámetros de cualquier tipo de distribución.
Por medio del análisis de la varianza del conjunto de los estimadores obtenidos se pueden generar intervalos de confianza. La creación de intervalos de confianza a través de simulación de Monte Carlo es llamada método de *bootstrap* y las pruebas de hipótesis basados en estimadores puros de Monte Carlo son llamados pruebas no paramétricas de Monte Carlo[ZHU05] que permiten realizar pruebas de hipótesis basadas en estadísticas con distribuciones arbitrarias.

2.9 Regresión y aproximación de funciones

El proceso de regresión es el estudio de la relación entre una variable de respuesta Y y un conjunto de variables X a través de una función f tal que:

$$Y = f(X)$$

Tal que $f(X)$ puede ser una función paramétrica.

Cuando un dato se encuentra en un espacio en el que es posible establecer una medida de distancia, entonces se puede encontrar un modelo $f(X)$ óptimo que minimice la distancia entre todos los puntos calculados a través del modelo y los valores de los datos. La medida de distancia definirá el proceso de búsqueda de la función de regresión y la calidad de cada función. Un ejemplo de medida de la calidad de la regresión es el coeficiente de correlación de Pearson analizado en la sección 2.4.

Se puede decir que incluso algoritmos de minería de datos incluyendo a las redes neuronales son un proceso de regresión en este sentido.

2.9.1 Regresión polinomial y teoría de la aproximación

La función con la que se modela la relación entre variables puede ser de distintos tipos. Un tipo de función común para funciones de regresión son los polinomios, ya que son capaces de aproximar una gran cantidad de funciones comunes. Teóricamente son capaces de aproximar cualquier función analítica [KRANTZ02]. Una función analítica se define como una función infinitamente derivable y por lo tanto se puede encontrar la serie de Taylor que la aproxima. Un polinomio de regresión de grado m en función de x tiene la siguiente forma

$$y_i = \sum_{j=0}^{m} c_j {x_i}^j + e$$

donde e es el error de aproximación que se trata de minimizar para todos los valores de las variables.

La teoría de la aproximación tiene por objetivo el desarrollo de algoritmos o procedimientos numéricos para la solución de problemas de ajuste de funciones que minimicen una medida de error entre el modelo y los datos reales. Este problema de manera general puede ser descrito de como sigue.

Dado el problema de obtener los parámetros del modelo $Y = f(X, \beta)$ donde β es el conjunto de parámetros a estimar, X es el conjunto de variables de entrada que puede ser una sola variable en el caso de regresión univariada, o más de una en el caso de regresión multivariada y Y una variable dependiente conocida, se debe minimizar una norma de error, por lo general basada en alguna de las medidas de distancia de Minkowski.

$$L_g = \sqrt[g]{\sum_{i=1}^{n} n\,(f(X) - Y)^g}$$

Donde L_2 corresponde a la suma de errores al cuadrado y L_∞ corresponde a la norma minimax:

$$L_\infty = \max(f(X) - Y)$$

La solución exacta a este problema usando la norma L_2, requiere la solución de un sistema de ecuaciones del tamaño del conjunto de datos, como sucede en el caso del algoritmo de mínimos cuadrados. Aunque otros algoritmos iterativos permiten reducir la cantidad de procesamiento para la obtención de un modelo (como Levenberg-Marquard [LEVENBERG44]), aún así resultan costosos computacionalmente en grandes conjuntos de datos y no existen generalizaciones para su aplicación en espacios de mayor dimensionalidad.

2.9.1.1 Polinomios minimax

Una alternativa son los polinomios que minimizan la norma L_∞, llamados polinomios minimax, ya que su solución involucra la solución iterativa de sistemas de ecuaciones de dimensiones menores como se presenta a continuación.

Para cada conjunto de $m + 1$ datos existe un polinomio compuesto por m monomios tal que minimiza el máximo error absoluto

$$\epsilon_i = |Y_i - P(X_i)|$$

El error de ajuste en la norma L_∞ en los *m+1* datos es igual a

$$\epsilon = \max(\epsilon_i)$$

Se puede mostrar a través del uso de la regla de Cramer (véase Apéndice C), que el valor mínimo de ϵ se alcanza cuando ϵ_i para todos los m+1 datos tienen un error de esa misma magnitud ϵ. El polinomio obtenido es el llamado polinomio minimax.

En un conjunto de n datos donde $n > m + 1$ el polinomio que mejor se ajusta en la norma minimax se obtiene a través de la búsqueda del conjunto de $m + 1$ datos, llamado conjunto interno, cuyo valor de ϵ sea el máximo entre todas las posibles combinaciones de elementos.

La importancia de este algoritmo radica en que no se requiere la solución de sistemas de ecuaciones del tamaño de todo el conjunto de datos, sino de subconjuntos no dependiente del tamaño del conjunto completo, sino del número de monomios utilizados, es decir del tamaño del conjunto interno. Adicionalmente existen algoritmos que permiten la optimización de los intercambios entre elementos de los conjuntos interno y el resto de los

datos, llamado conjunto externo, logrando que la búsqueda del conjunto de datos avance de manera constante a hacia la solución en caso de que no exista inestabilidad numérica por la precisión manejada. El algoritmo de ascenso (AA) [CHENEY98] presenta la manera en que se pueden estos intercambios garantizando la aproximación cada vez mejor al conjunto interno que cumpla con la condición minimax consistente en que el polinomio calculado minimiza el máximo error de ajuste en todo el conjunto externo e interno. Una descripción más detallada del cálculo de polinomios minimax es presentada en el Apéndice C, incluyendo la metodología para una implementación eficiente y el análisis de su complejidad algorítmica.

Dado que la función de aproximación se define como

$$p(x) = \sum_{i=1}^{m} c_i M_i(X)$$

El algoritmo de ascenso puede ser utilizado para el cálculo de los coeficientes c_i sin importar la función que sea $M_i(x)$. Si $M_i(x) = x^j$ en el caso con una sola variable dependiente, estos coeficientes generan el polinomio minimax. Este manejo general de la función de aproximación permite utilizar diferentes funciones $M(x)$, pudiendo hacer el cálculo de coeficientes de cualquier tipo de función, incluso multivariada de la forma $M(x_1, x_2, \dots, x_V)$ para V variables abriendo así la posibilidad de hacer un ajuste multivariado eficiente. Algunos ejemplos de las aproximaciones con polinomios de aproximación minimax con una y dos variables dependientes son presentados en las figuras 2.1 y 2.2.

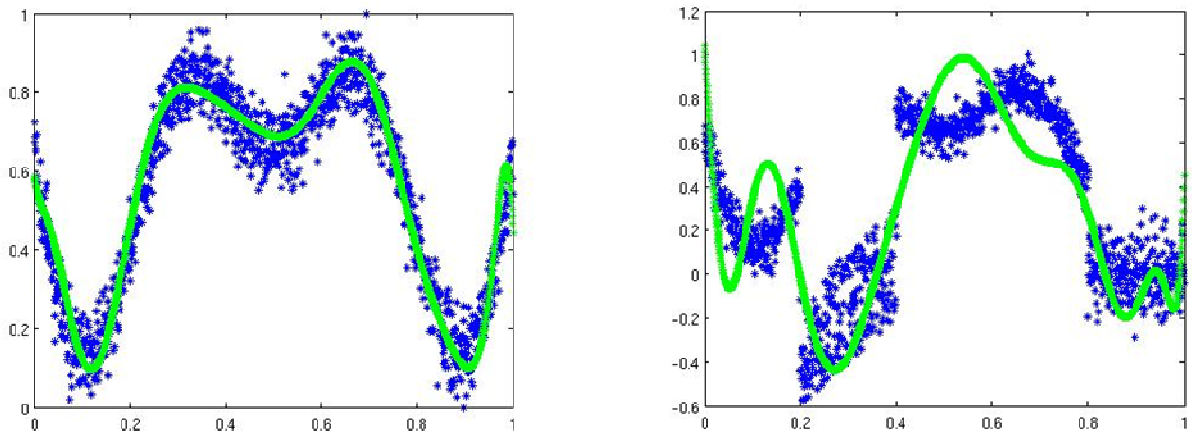

Figura 2.1. Aproximación minimax en conjuntos de dos dimensiones

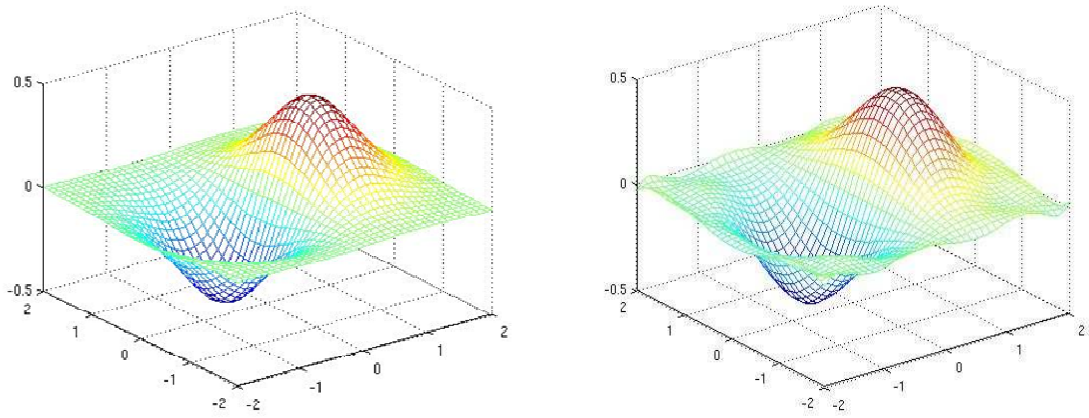

Figura 2.2. Aproximación minimax en conjuntos de tres dimensiones con máximo grado 9. A la izquierda se presenta el conjunto original. A la derecha la gráfica de la aproximación.

2.9.1.2 Selección de monomios y uso de un meta-algoritmo

La aproximación en muchas ocasiones no requiere del uso de ciertos monomios ya que estos no atrapan información respecto a las relaciones de variables en ese orden. Es por ello que no es necesario el cálculo de un polinomio que use todos los posibles monomios para las variables seleccionadas. Además, esto es conveniente cuando se tiene un número mayor de variables ya que el número de monomios crecerá exponencialmente con el grado máximo de las variables que se utilice, causando que el cálculo requiera la solución de grandes sistemas de ecuaciones. Es por ello que se puede hacer una búsqueda por medio de algoritmos de optimización combinatoria para la selección de los monomios más significativos obteniendo diferentes polinomios para evaluar su ajuste. Una opción adecuada para esta búsqueda es el uso de algoritmos genéticos.

2.9.1.3 Algoritmos genéticos

Los algoritmos genéticos son una interesante herramienta de optimización combinatoria que a partir de la posibilidad de codificación de una solución hace una búsqueda en el espacio de soluciones de manera paralela sobre conjuntos de posibles soluciones aprovechando la granularidad implícita en el problema de interés a través de la aplicación de operadores evolutivos. La nomenclatura y las ideas básicas de estos algoritmos tienen similitudes con los procesos evolutivos de la naturaleza, sin embargo es importante destacar que no se trata de una analogía exacta, pues es un proceso de una naturaleza más simple que aprovecha la posibilidad de codificación y la existencia de modularidad o granularidad en las soluciones codificadas.

El algoritmo genético básico consiste en lo siguiente:

1. Generación de un conjunto de soluciones iniciales. El conjunto de soluciones creado es llamado población, donde cada uno de las posibles soluciones codificadas es un individuo. El código que representa al individuo es llamado genotipo. La codificación usada para los individuos afectará los resultados del proceso pudiendo facilitar o complicar la búsqueda dependiendo de cuanto encapsule la modularidad de la solución.
2. Obtención de función de aptitud. Cada genotipo se expresa a través de un fenotipo, es decir que existe un mapeo del espacio de genotipos a un espacio que puede tener otra codificación y que le da características a cada individuo, relacionándose con una medida de aptitud (*fitness*). Esta aptitud es definida como una función del fenotipo en analogía con la capacidad de supervivencia de un individuo.
3. Selección. Los individuos que tuvieron una aptitud suficiente son seleccionados. Para determinar la suficiencia se puede establecer un umbral de aptitud o un número de individuos de modo que los *n* mejores sean seleccionados.
4. Cruza. Los individuos de la población combinan entre sí sus genotipos para dar paso a una nueva generación de individuos explotando la modularidad que pudiera haber en el espacio de soluciones en la codificación usada.
5. Mutación. Aleatoriamente se alteran partes de los genotipos de los individuos, esto permite explorar el espacio de búsqueda para evitar mínimos locales en la solución.

La importancia de los algoritmos genéticos radica en su capacidad de exploración y explotación en paralelo de espacios de búsqueda complejos asegurando la convergencia aunque sea en un tiempo infinito a un óptimo global. Existen variaciones a estos algoritmos para solucionar problemas de convergencia o la dependencia de la codificación de las soluciones, entre los cuales un algoritmo que ha probado dar buenos resultados es el algoritmo genético de Vasconcelos VGA [KURI02] que representa una aproximación práctica cercana al algoritmo genético idealizado que soluciona los problemas del algoritmo canónico.

El VGA presenta las siguientes características

- Conservación de mejores y peores elementos.
- Cruza determinística entre mejores y peores elementos, pero conservando a los individuos de la generación anterior, aumentando la variedad de individuos.
- Cruzamiento de tipo anular produciendo diferentes puntos de cruza variables y siendo independiente de la codificación de los individuos.

Capítulo 3. Descripción de la metodología

En este capítulo se presenta la descripción de la metodología llamada FDM, para la obtención de una muestra representativa para la minería de datos. En todo proceso de análisis de datos es fundamental la selección cuidadosa del conjunto de herramientas que se utilizarán, así como su uso adecuado, aprovechando tanto las capacidades de cada algoritmo así como la información contenida en el conjunto de datos analizado para obtener resultados eficientes, es decir, que se obtengan resultados útiles con tiempos de procesamiento relativamente cortos. En minería de datos esto es aún más importante ya que se presentan las condiciones que aumentan la complejidad tanto para la obtención de resultados exactos como la rápida ejecución de algoritmos. Es por esto que FDM consiste en un análisis de los datos para usar sólo el volumen necesario para la minería.

3.1 Enfoque y caracterización de la metodología *Fast Data Mining* (FDM)

Esta metodología está enfocada en seleccionar un subconjunto de datos que presente una distribución prácticamente equivalente a la de toda la base de datos, sin hacer uso de procesos costosos computacionalmente. Para simplificar la nomenclatura se usarán los siguientes términos:

P: Población con todos los datos disponibles en la base de datos
M: Muestra representativa de P
V: Número de atributos de P

Al hacer agrupamiento de datos, el objetivo de la búsqueda es una visión general de las características de los datos que permite relacionar grupos de datos por características comunes o cercanas. Este análisis depende de la distribución de los datos en un espacio n dimensional y por lo general los grupos de datos de mayor importancia son aquellos que contienen un gran número de elementos. Dado que los grupos más significativos son aquellos que tienen un número mayor de elementos la distribución de estos elementos puede ser conservada por un subconjunto menor que mantenga aproximadamente la proporción de elementos en cada uno de los grupos.

Esta idea lleva a la propuesta central de esta metodología, que consiste en la selección de una muestra M que mantenga la distribución de los datos en cada uno de los grupos identificables permitiendo hacer una búsqueda más rápida y eficiente.

El preprocesamiento y la validación de los resultados aquí propuestos corresponden a la cuarta etapa del proceso de descubrimiento de conocimiento en bases de datos, correspondiente a la etapa de transformaciones sobre el conjunto de datos, ya que se trata de una reducción de dimensionalidad, específicamente una reducción de número de instancias y su verificación.

Esta metodología puede ser aplicada en bases de datos de grandes dimensiones debido a que hace uso de algoritmos con una reducida sensibilidad al tamaño del conjunto de datos.

El obtener la muestra M puede ser de gran importancia práctica ya que los algoritmos orientados a grandes bases de datos requieren en su mayoría del uso de recursos de computo de mayor capacidad haciendo uso de paralelismo masivo o de altas capacidades de procesamiento. Uno de los objetivos de esta metodología es que el resultado de este preprocesamiento permita el uso de equipo de cómputo de capacidades regulares en la etapa de procesamiento.

FDM requiere de etapas previas que le proporcionen los datos en el formato y con las características para el análisis debido y genera una muestra que puede ser usada en el resto del análisis. A continuación se presenta cada una de las etapas previas requeridas y la presentación de la metodología para la reducción de instancias.

3.2 Preparación de los datos

FDM al igual que otras alternativas, requiere de una etapa de preprocesamiento de la información, con el fin de permitir:

- Organizar la información de manera que se tenga una vista de las características de interés.
- Mejorar la eficiencia al reducir espacios de búsqueda o permitir un manejo más eficiente de los datos (reducción de dimensiones).
- Conservar la estabilidad en los algoritmos para evitar problemas numéricos, o resultados incongruentes (normalización).
- Aumentar la versatilidad de la metodología, para poder aplicarla a un mayor número de casos (transformación a variables numéricas, procesos de discretización).

3.2.1 Obtención de vista para la minería

Una vez que se inicia el proceso de la minería de datos es necesario un proceso de organización de la información disponible en las bases de datos para el análisis. Las tablas de base de datos pueden estar estructuradas y organizadas de acuerdo al modelo de base de datos utilizado, como por ejemplo, estar organizadas de acuerdo a un proceso de normalización en el sentido de los modelos relacionales, o puede seguir otros modelos (como los orientados a los procesos de OLAP). Para fines de esta metodología se utiliza la información en forma de una sola tabla en la que se tienen todos los atributos relevantes para los objetivos planteados, es por ello que las tablas de la base de datos, de ser necesario, deben ser debidamente unidas y preprocesadas para obtener una vista simple de los datos de interés.

3.2.2 Corrección de inconsistencias

Una vez organizada la información disponible es conveniente un análisis rápido sobre los valores que se manejarán. Los datos pueden presentar defectos que obstaculizan el correcto desempeño del análisis, como datos faltantes, inconsistencias, errores de captura o de tipo, etc.

El problema de datos faltantes suele resolverse a través de los conocimientos que se tengan con respecto al área que se está manejando o por la distribución del resto de los datos. El objetivo de esta sustitución es fundamentalmente no afectar el proceso de análisis sin perder la información del resto de las variables. Las opciones más comunes son la sustitución de elementos por la media de la variable, la repetición del dato con los diferentes posibles valores que pudiera haber tomado o simplemente eliminar el dato que contiene. Para una referencia más completa sobre métodos para completar valores faltantes, se puede consultar [GRZYMALA05].

La detección de elementos atípicos (*outliers*) es compleja y requiere de un análisis más detallado de la población, las principales opciones para este proceso son las derivadas de la estadística, considerando atípicos aquellos elementos que se encuentren demasiado alejados de la media, basándose en la desigualdad de Tchebyshev que establece el límite inferior a la probabilidad de encontrar un elemento a una distancia dada del valor esperado de la población dada un varianza finita en la distribución. Otra opción es usar el resultado de un algoritmo de agrupamiento que permita identificar los elementos que no pertenecen a ningún grupo significativo. Otras técnicas para la detección de elementos atípicos pueden ser consultadas en [BEN05].

3.2.3 Transformaciones a los datos.

Los algoritmos propuestos en esta investigación trabajan con datos numéricos; es por ello que las variables no numéricas deben ser cuidadosamente codificadas o eliminadas de ser necesario. Si se realiza la transformación para realizar el cambio de tipo de variable se debe hacer un mapeo congruente a un valor numérico. Creando un diccionario de datos que permita recuperar la información mapeada.

Cabe destacar que estas transformaciones pueden presentan algunos problemas en la exactitud del resultado que puedan generar, pues variables de tipo no numéricas, no pueden ser fácilmente mapeadas a un número real ya que es difícil asignar una diferencia cuantitativa entre dos valores de estas variables. Debido a lo anterior, en el análisis aquí presentado se asume que la información es representativa de la característica que se desea analizar, es decir que fue seleccionada como información relevante y que esta pudo ser numéricamente codificada de manera eficiente y significativa. El análisis de este tipo de problemas es aún un campo abierto de investigación, y en ocasiones, al no poderse solucionar de manera óptima se utilizan algoritmos diseñados para este tipo de datos.

Es además conveniente aplicar un proceso de normalización en un intervalo a los datos por procesar. En este contexto normalización se refiere al mapeo de los valores numéricos a un rango acotado. En este caso se sugiere la normalización a valores en el intervalo [0,1] o [-1,1].

Esta normalización presenta diferentes ventajas, por una parte reduce la probabilidad de inestabilidad numérica al procesar la información en los algoritmos y por otra permite balancear el peso de las variables en cada dimensión ya que las variables originales pueden presentar diferentes escalas por lo que podrían existir grandes variaciones en dimensiones de grandes escalas que anulen la importancia de variaciones en otras dimensiones de menor escala.

3.3 Reducción de dimensionalidad

3.3.1 Selección de variables

La selección de variables o de características tiene la finalidad de obtener un conjunto reducido de variables que permitan obtener los mismos resultados que el conjunto original eliminando redundancia o seleccionando las variables más relevantes.

Dado que en esta etapa no se ha aplicado el muestreo, el tamaño de la población puede ser grande por lo que se sugiere un criterio sencillo para la eliminación de variables redundantes, como la eliminación de variables linealmente correlacionadas (véase sección 2.4) ya que aportan información que ya está siendo proporcionada por otra variable directamente para el análisis. Este método para selección de variables es práctico por mostrar directamente la redundancia al usar ciertas variables y requiere de procesamiento relativamente ligero.

3.3.2 Reducción de instancias

Una vez que se ha dado el debido preprocesamiento de la información se puede aplicar un método de selección de instancias. Esta es la propuesta central de este trabajo, tratando de lograr una reducción significativa, debidamente verificada para su uso para el análisis de bases de datos.

La selección de instancias puede realizarse por medio de distintas estrategias (véase sección 1.5.3), sin embargo los métodos de selección de instancias enfocados a algoritmos de agrupamiento resultan ser poco eficientes al requerir a su vez el uso de versiones simplificadas de estos algoritmos como *k medias*, que ya presentan deficiencias en exactitud de resultados o tiempos de procesamiento no despreciables. Es por ello que la opción más común y conveniente es realizar un muestreo aleatorio simple. Sin embargo no se han hecho suficientes análisis específicos respecto a este proceso, por lo que la selección de la muestra se hace con criterios poco apropiados, provenientes de análisis estadísticos sobre conjuntos de datos relativamente pequeños o con distribuciones sencillas ya que las pruebas estadísticas y teoría del muestreo básico se centran en un análisis de poblaciones que no tienen las dimensiones ni la heterogeneidad presente en los análisis de minería de datos.

Las decisiones respecto al tamaño de la muestra con base en criterios inadecuados pueden llevar a errores imperceptibles para el análisis realizado, pero importantes en el resultado final que pudieran indicar que el muestreo no es conveniente. En esta investigación se trata de mostrar que este muestreo es posible pero tomando en cuenta las características de la población de una forma más detallada y orientada al análisis que se quiere hacer.

Para no perder generalidad en este trabajo, se trata de manejar todas las validaciones y análisis, sin analizar la información de dominio del problema; sin embargo en caso de tener este tipo de información, que permite discriminar elementos en el muestreo no deja de ser conveniente para fines prácticos. La introducción del conocimiento que se tiene acerca de los datos manejados es también un campo de interés para este tipo de algoritmos.

3.4 Etapas FDM

FDM es una metodología con dos etapas:

1. Obtención del tamaño de la muestra para cada variable
2. Validación de la muestra obtenida en espacios multidimensionales.

3.4.1 Etapa 1: Obtención del tamaño de la muestra

Para la determinación del tamaño de la muestra se utilizan los conceptos básicos de teoría de la información por motivos que serán expuestos más adelante. En esta investigación se propone un muestreo aleatorio simple debido a que permite mayor versatilidad al definir sólo un tamaño de muestra mínimo que asegure un resultado aceptable sin requerir de conocimientos previos para no perder aplicabilidad en cualquier conjunto de datos. Otros muestreos probabilísticos podrían generar muestras de menor tamaño pero generan un conjunto fijo de objetos que deberá ser actualizado continuamente.

3.4.1.1 Descripción de muestreo

Analizando al conjunto de datos como una secuencia de símbolos que componen un mensaje de acuerdo con la teoría estadística de la información, cada conjunto de datos pertenecientes a un atributo representa un mensaje y cada valor del atributo representa un símbolo. De acuerdo a lo anterior se puede calcular la entropía contenida en cada uno de los mensajes que en este caso son los atributos. Esta entropía es una medida de la información promedio que contiene un símbolo en un mensaje (véase sección 2.7). En este caso la entropía es usada como una herramienta que asegure que la información presente en cada variable sea conservada en el proceso del muestreo, es decir que será la estadística de prueba que permita validar la representatividad de la muestra

.

La entropía en el mensaje puede ser aproximada a través del promedio de la proporción de aparición de cada símbolo en este mensaje, de acuerdo a lo siguiente:

$$-H(X) = \sum_{i=1}^{r} (p_i \log(p_i)) \cong \sum_{i=1}^{r} (\sum_{j=1}^{n} \{\frac{\delta(S_i, v_j)}{n}\} \log (\sum_{j=1}^{n} \frac{\delta(S_i, v_j)}{n}))$$

Donde X es el mensaje, p_i es la probabilidad de aparición del símbolo i, r es el número total de símbolos, s_i es el $i-ésimo$ símbolo, v_j es el valor del j-ésimo de n datos. Y

$$\delta(s, v) = \begin{cases} 1 & s = v \\ 0 & s \neq v \end{cases}$$

Dado lo anterior se puede aceptar una muestra que contenga un valor de entropía cercano al de la población, sin embargo el valor de la entropía de la población tampoco se conoce en

la práctica. El cálculo del valor de la entropía de las variables de la población P que puede ser de dimensiones muy grandes, no es conveniente ya que esto exige una gran cantidad de procesamiento al requerir el uso de todos los datos de la base. Por ello una alternativa es aproximar este valor aprovechando la consistencia de la estadística con la que se estimó la entropía. La consistencia estadística de la entropía consiste en que el valor de la estadística converge al del parámetro estimado, que en este caso es la entropía de P, al aumentar el tamaño de la muestra, esto se puede observar al calcular el valor de la entropía en muestras de diferentes tamaños como los mostrados en la figura 3.1.

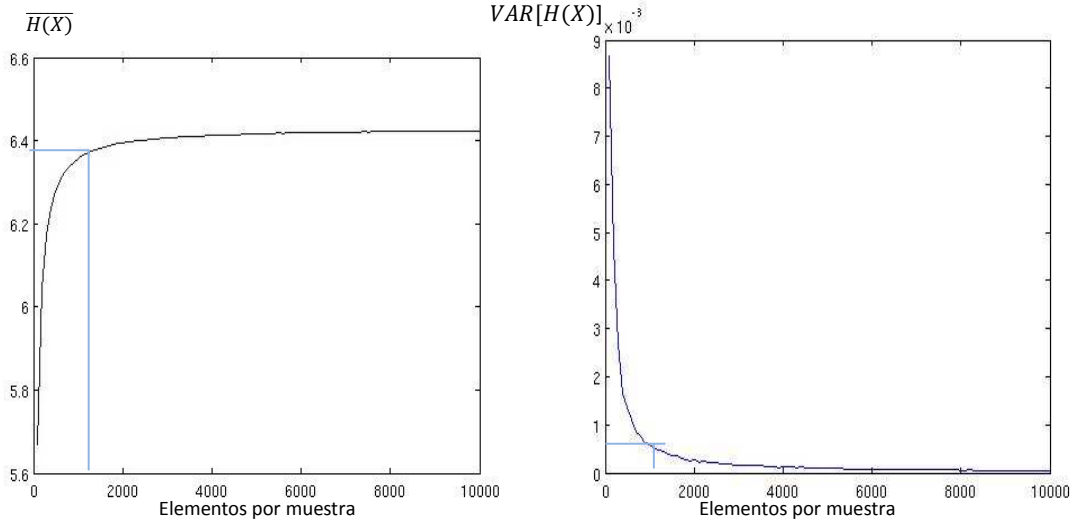

Figura 3.1. Comportamiento de la entropía para conjuntos de diferentes tamaños de muestras

A) Media de los valores observados en muestras de diferentes tamaños
B) Varianza de los valores observados en muestras de diferentes tamaños

A partir del comportamiento de esta curva se puede estimar una entropía y un tamaño de muestra en el que no se tengan grandes pérdidas tomando el tamaño en el que las variaciones de la entropía no exceden un valor umbral ΔH. El número de elementos que requiere para alcanzar ese punto varía de acuerdo a la distribución de los datos y es mayor en caso de existir símbolos con una menor probabilidad de aparición. Un ejemplo del tamaño de la muestra obtenible es mostrado en la figura3.1 con línea clara

Sea M la muestra representativa de toda la base P y $|M|$ el número de elementos de que componen a M, es decir, el tamaño de la muestra. El proceso del cálculo del tamaño de la muestra para el conjunto de datos X de n elementos y c atributos, es el siguiente.

Se tiene un conjunto de muestras M_1, M_2,...,M_V sin elementos y un conjunto de tamaños de muestra $|M_1|$,$|M_2|$,...,$|M_V|$ y se toma para cada atributo *t* una muestra M_t.
Se extrae de manera aleatoria con una probabilidad de selección uniforme un conjunto de ρ elementos del atributo t y se agrega a la muestra M_t. El tamaño de la muestra se incrementa agregando ρ elementos.

- En cada iteración i en el atributo t, la entropía de la muestra es calculada y comparada con la entropía de la iteración anterior.
- Si la diferencia entre la entropía de la iteración interior:

$$\Delta H = \frac{|H(i) - H(i-1)|}{H(i)}$$

  es menor que un valor umbral entonces se define el tamaño de la muestras para ese atributo $|M_t|$ y se analiza el siguiente atributo. Si no es así se repite el proceso desde el paso 2.
- Al terminar de determinar los tamaños para cada atributo $|M_1|, |M_2|, \ldots, |M_{V|}$, se obtiene el tamaño el máximo de estos tamaños

$$|M| = \max(|M_1|, |M_2|, \ldots, |M_V|)$$

Para evaluar la probabilidad de haber obtenido un valor de entropía sólo por aleatoriedad en el muestreo es conveniente estimar la distribución de la entropía para el tamaño de muestra dada. Esto se puede lograr por medio de una prueba no paramétrica de Monte Carlo (Véase sección 2.8). Esta prueba consiste en el cálculo de la estadística, que en este caso es la aproximación de la entropía, en un conjunto de muestras,

Se considera que el valor de la estadística calculada es representativo si al menos $\kappa\%$ de las estadísticas calculadas están dentro del intervalo aceptable, entonces la prueba no tiene argumentos para rechazar esta muestra; de lo contrario el muestreo debe ser rechazado y continuar la búsqueda del tamaño de muestra regresando al segundo paso del procedimiento.

3.4.1.2 Justificación del muestreo por entropía

El uso de la entropía como una medida de la conservación de la integridad de los grupos de datos presentes, es justificado siguiendo el siguiente razonamiento:

- La capacidad de identificar grupos en un conjunto de datos se da por la detección de similitudes en el conjunto de objetos y estas similitudes implican la existencia de un patrón, considerando en este contexto a un patrón como un conjunto de características que aparecen de repetidamente en los datos. Un ejemplo de un patrón observable en una base de datos se da cuando en un subconjunto de elementos, ciertas variables tienen un mismo valor o valores muy cercanos o cuando el valor de una variable determina el valor de otra en un subconjunto relativamente

grande de datos. En estos casos la existencia del patrón está relacionada con una similitud entre objetos.

- En todo conjunto de datos que no es totalmente aleatorio existen patrones que implican la posibilidad de la existencia de una codificación que pueda expresar el mensaje en una manera más compacta aprovechando la existencia de esos patrones.
- Por lo anterior existe una relación entre la existencia de estos patrones y la longitud de codificación posible, o en otras palabras existe una relación entre los grupos existentes y la entropía en la información.
- Es por ello que una reducción de información eficiente debe conservar la entropía de las variables para que la codificación óptima sea la misma o muy similar a la de la población original.

Esto sugiere que la entropía representa una buena opción como guía en el proceso de selección de elementos en una muestra.

3.4.1.3 Consideraciones en poblaciones de mayor complejidad

Dada la complejidad de los conjuntos que se puedan analizar, se pueden proponer algunas acciones adicionales en el proceso para realizar eficientemente el muestreo.

Aplicación en datos de tipo continuo dispersos

En caso de tener datos dispersos, probablemente provenientes de una variable de valores continuos, dado que la probabilidad de que dos valores sean iguales es reducida, el uso de un símbolo por cada valor de atributo no es apropiado ya que es probable que la distribución de los símbolos sea uniforme por diferencias poco significativas en los datos. En su lugar un intervalo puede ser utilizado como un símbolo.

El intervalo que se utilice debe ser de longitud constante para que cada símbolo conserve su distribución, ya que intervalos de longitud variable, como los que se pudieran usar para una prueba como la de χ^2, pueden ocasionar que la distribución de los símbolos sea más uniforme, ya que esos intervalos tratan de agrupar al menos un número mínimo de datos en su intervalo. Esto no es deseable en el análisis ya que la existencia de estos elementos con menor probabilidad de aparición indica que se requiere un tamaño de muestra mayor para que atrape la información de estos símbolos y si se uniformiza la distribución, símbolos con menor probabilidad se unirán con otros con mayor probabilidad de ser muestreados.

La longitud de los intervalos para cada símbolo depende de la dispersión de los datos y debe ser suficientemente pequeño para atrapar la información de cada símbolo pero

suficientemente grande para agrupar elementos con valores cercanos. Una manera de obtener un número de intervalos adecuado es hacer el análisis de la entropía en algunas muestras al dividir sucesivamente en mayor número de intervalos el espacio de los datos, de modo que si la entropía se conserva al hacer la división del intervalo entonces no es necesaria hacer esta separación. Otros métodos de discretización pueden ser consultados en [YAN05]

Consideraciones respecto al número de dimensiones

Por la forma en la que se calcula el tamaño de la muestra, el aumento en el número de dimensiones en el conjunto de datos no aumenta la complejidad del análisis ya que el proceso es aplicado variable por variable. Esto es una ventaja en procesamiento pero implica la suposición de que el análisis de cada una de las variables puede ser suficiente para conservar los patrones distribuidos en espacios multidimensionales,

Esto sugiere que el tamaño de la muestra debería ser calculado a través de la entropía conjunta del conjunto de variables como estadística de prueba. Sin embargo por razones prácticas se restringe el análisis al cálculo de la entropía de cada atributo, dado que es un análisis mucho más rápido y mantiene una relación con la entropía conjunta y las características de los grupos ya que el conservar suficientemente la distribución de cada variable aumenta significativamente la probabilidad de conservar los patrones multidimensionales más relevantes.

La complejidad del análisis de la entropía conjunta puede ser fácilmente observada en la cantidad de símbolos que se pueden considerar para su cálculo. El número de símbolos analizables para el cálculo de la entropía conjunta de V variables, donde cada una de esas variables v_i tiene n_i símbolos, es igual a $\prod_{i=1}^{V} n_i$. Dado que en el caso de tener variables continuas el número de símbolos puede ser alto, es claro que el análisis puede ser complejo, además de requerir de accesos más costosos a la base de datos para la extracción y almacenamiento temporal de grupos de tuplas mayores.

Es por ello que se considera en esta investigación que la conservación de la entropía de cada una de las variables puede ser suficiente para obtener una muestra representativa. Cabe resaltar que no es posible representar el comportamiento multivariado de un sistema por medio de un subconjunto de variable; sólo se está asumiendo que la muestra con suficiente información de cada variable, tendrá mayor probabilidad de conservar los patrones en espacios de dimensionalidad superior. Es por lo anterior que validaciones multidimensionales de la muestra deben también ser aplicadas.

3.4.2 Etapa 2: Validación multidimensional de la muestra

Para analizar el comportamiento de la muestra en un mayor número de dimensiones se pueden aplicar pruebas multivariadas que permitan asegurar que la probabilidad de error generada por el muestreo esté limitada y sea baja. Estas pruebas tienen el objetivo de validar que también la entropía conjunta se está conservando.

Para probar la conservación de las características de las variables de la muestra comúnmente se propone un análisis a través de pruebas estadísticas. Sin embargo las pruebas estadísticas no paramétricas comúnmente aplicadas no resultan apropiadas, al ser conservativas. es decir, aceptan la representatividad de muestras de tamaño pequeño sin ser necesariamente correctas, debido a la pérdida de información al pasar por un proceso de discretización, uso de rangos, o ser sólo medidas desviación respecto una medida central que no mantiene una relación con el proceso de búsqueda de patrones o grupos. Adicionalmente las generalizaciones a espacios multidimensionales, no son eficientes al no existir la posibilidad de lograr un ordenamiento natural de datos en espacios multidimensionales. El análisis y comparación de los resultados obtenibles con otras pruebas estadísticas son presentadas en el capítulo 4.

En vez de las pruebas estadísticas tradicionales en FDM se propone una metodología que capture las relaciones entre variables a través del cálculo de funciones de regresión para superar las limitantes mencionadas y hacer un análisis de relaciones de cualquier orden entre variables. En esta metodología se propone el uso del algoritmo de ascenso AA para la minimización de la norma L_∞ (véase sección 2.9.1).

La función obtenida es un estimador de la distribución de una variable dependiente, dadas las demás variables independientes, por lo que los valores de error de aproximación obtenidos tienen una relación con la información mutua entre las variables, la cual no se analizó al hacer el análisis univariado.

Si se define a $X = (x_1, x_2, \dots x_d)$ como el conjunto de variables independientes con d componentes, donde d es el número de variables independientes elegido para la aproximación, F la variable dependiente y $P(X)$ el polinomio de aproximación tal que

$$P(X_i) = F_i + e_i$$

El término e_i es el error de aproximación generado en el ajuste de F_i en función de X_i.
Al obtener el polinomio $P(V)$, esta misma función puede ser evaluada en otras muestras. Si los datos tienen una distribución cercana entre sí, entonces el error de aproximación debe

ser cercano al aplicarse a las diferentes muestras. El error de aproximación en L_2 es definido como

$$e = \sqrt{\sum_{i=1}^{n} (P(x_i) - F_i)^2}$$

A pesar de que el polinomio calculado a través del algoritmo de ascenso minimiza la norma L_∞, la validacion propuesta utliza L_2 ya que es una norma menos sensible a valores atípicos y toma en cuenta los errores de todos los elementos en la muestra.

Al obtenerse el error de ajuste en el conjunto de muestras se puede hacer una comparación entre el error de ajuste en ellos. No se conoce la distribución del error de ajuste en un conjunto de muestras pero se sabe que la muestra es aceptable si se conservaron las relaciones entre variables, por lo que el error de ajuste es cercano. Es por eso que a través de la proporción

$$R = e_{max}/e_{min}$$

Se verifica que la variación entre los errores de ajuste están acotados de manera que la muestra puede ser rechazada si

$$R < 1 + \gamma$$

donde γ es un parámetro para el algoritmo que indica la máxima variación aceptable al comparar los errores de ajuste. Se sugiere la asignación de un valor cercano a 0.1 para gamma, que supone que las muestras tienen una diferencia en el error de ajuste no mayor al 10% El análisis detallado de la distribución de los errores, también se puede realizar, en vez del uso de valores mínimos y máximos, pero esta resulta en una prueba más estricta de la conservación de los patrones.

Esta verificación de conservación de relaciones debe ser aplicada para cada conjunto de datos.

En caso de que la validación multidimensional encuentre diferencias significativas en el conjunto de datos el tamaño de la muestra deberá ser aumentado haciendo un análisis con un valor de ΔH menor.

3.4.2.2 Consideraciones por el número de dimensiones para la aproximación de funciones

Dado que el número de variables que pueden componer a la muestra puede ser alto, el cálculo de los polinomios de aproximación para todos los posibles conjuntos de variables

puede ser también un proceso lento. Es por ello que se puede hacer un muestreo de las posibles funciones a verificar bajo el siguiente criterio.

Si se quiere encontrar las funciones que aproximan cada una de V variables en función de otras d variables independientes. El número de posibles funciones N a obtener es el siguiente:

$$N = V \binom{V-1}{d}$$

Donde $\binom{V-1}{d}$ son las combinaciones de de $V-1$ variables en grupos de d elementos. En el caso particular de la aproximación por pares de variables se tiene que el número de variables independientes es 1, por lo que se tendrían que hacer $V(V-1)$ comparaciones.

Para hacer el análisis del número de funciones que deben obtener para validar una muestra con un cierto grado de confianza, se puede definir como N_α al conjunto de funciones que dieron resultados positivos en la validación y $|N_\alpha|$ su cardinalidad, la probabilidad de tomar un elemento del conjunto N_α al obtener la muestra M_f de tamaño S formada por un conjunto de funciones de aproximación para validar M a través de un muestreo aleatorio simple es

$$P(N_\alpha) = \left(\frac{|N_\alpha|}{N}\right)\left(\frac{|N_\alpha|-1}{N-1}\right)\left(\frac{|N_\alpha|}{N-2}\right)\ldots\left(\frac{|N_\alpha|-S-1}{N-S-1}\right) = \frac{|N_\alpha|\,!\,(N-S)!}{N!\,(|N_\alpha|-S)!}$$

si establecemos un valor de probabilidad $1-\tau$, podemos seleccionar un tamaño de muestra S tal que :

$$P(N_\alpha) < 1-\tau$$

Si S_{min} es el mínimo tamaño en el que se cumple esta desigualdad, se puede asegurar que una muestra de tamaño S_{min}, con elementos que hayan pasado la validación, asegura con una confiabilidad de $1-\tau$ que las validaciones fueron exitosas en las relaciones de variables analizadas.

Para evitar la solución de la desigualdad que implica este cálculo, S_{min} se puede obtener iterativamente, ya que los valores de T,d son conocidos. Esto se hace aumentando gradualmente el valor de S hasta que se cumpla la desigualdad y la probabilidad sea menor que $1-\tau$.

Cabe mencionar que aunque se considera deseable para el análisis el uso de polinomios que aproximen en espacios de muy alta dimensionalidad es complicado por el límite de precisión impuesto por el uso de las computadoras digitales. En el capítulo 4 se discutirá el número de variables más adecuado.

3.5 Aplicación de algoritmos de agrupamiento

El conjunto de datos resultante del muestreo de la base de datos debe ser usado finalmente en un algoritmo de agrupamiento de datos. En este trabajo se usan algoritmos de costo computacional no muy elevado, que no sacrifiquen la calidad de los resultados como los mapas auto-organizados y c-medias difusas. Es recomendado usar algoritmos cuya complejidad no llegue a ser cuadrática sobre el conjunto de datos, ya que aún cuando la muestra es reducida en dimensiones, el análisis puede ser lento y la diferencia en exactitud no da una ventaja suficiente en eficiencia.

La reducción aquí propuesta puede ser aplicada antes de cualquier algoritmo de agrupamiento incluyendo aquellos optimizados para grandes bases de datos haciendo una búsqueda sobre conjuntos más reducidos de una manera justificada y también es aplicable después de cualquier otro proyecto de compresión ya que su única restricción es el uso de valores numéricos.

Capítulo 4. Aplicación y análisis

La metodología FDM propuesta en el tercer capítulo incluye distintas hipótesis que vale la pena analizar y verificar a través de la implementación de este proceso además de hacer un análisis de la sensibilidad que se presenta a los diferentes parámetros. Es por ello que en este capítulo se presenta una descripción y análisis de las diferentes etapas de su aplicación, así como los resultados alcanzables.

Inicialmente se describen las ventajas que tiene el método de muestreo contra otras validaciones estadísticas similares. Posteriormente se describen los detalles de su aplicación a través de la descripción del proceso aplicado a diferentes conjuntos de datos y finalmente se hace un análisis de la eficiencia obtenida.

4.1 Experimentos realizados

Para el análisis de la metodología FDM se implementó cada una de las etapas que la componen así como algoritmos relacionados. El desarrollo de este proyecto involucró la investigación de diferentes alternativas para la validación estadística, la aproximación de funciones, y el agrupamiento de datos, seleccionando e implementando como resultado, aquellas herramientas que son adecuadas para los conjuntos de datos con las características antes mencionadas, analizando las ventajas y problemas que pudiera tener cada algoritmo propuesto .

Estos experimentos se hicieron en Matlab®, por facilidad para la visualización, eficiencia en el manejo matricial y conveniencia para el diseño de prototipos de programas para el análisis numérico. También se utilizó el software DataEngine® para la validación de resultados de algoritmos de agrupamiento.

Los programas implementados tienen las siguientes funciones:

a) Generación de grupos de datos de pruebas con distintas distribuciones no triviales como las que se pueden encontrar en casos reales.
b) Preprocesamiento de datos.
c) Comparación de resultados de pruebas estadísticas y análisis de errores estadísticos.
d) Cálculo de tamaño de la muestra de acuerdo con FDM.
e) Algoritmos para la aproximación de funciones en dos dimensiones
f) Algoritmos de aproximación de funciones en espacios multidimensionales,

incluyendo el uso de herramientas como el algoritmo genético.
g) Validación de las muestras por medio de aproximación de funciones.
h) Agrupamiento de datos y comparación de resultados.

En el apéndice A se incluye una descripción de los programas creados para estos fines y a continuación en las secciones 4.2 y 4.3 se presentan los resultados más importantes de las pruebas realizadas con estos programas.

4.2 Análisis del muestreo

Los tamaños de muestra obtenibles a través de fórmulas tradicionales basadas en la suposición de una distribución, como por ejemplo los obtenibles por medio de las fórmulas de Cochran [COCHRAN77], no pueden ser utilizadas en este tipo de análisis ya que en raras ocasiones estas suposiciones son acertadas para poblaciones heterogéneas de grandes dimensiones como las manejadas en la minería de datos. El análisis aquí propuesto sugiere el uso de pruebas con la capacidad de detectar una muestra representativa para un análisis de agrupamiento de datos sin perder potencia en la prueba o depender del conocimiento de la distribución de una estadística. A continuación se presenta un análisis de posibles pruebas para determinar el tamaño de una muestra basadas en suposiciones cercanas a la utilizada por la metodología aquí presentada.

4.2.1 Comparaciones en espacios unidimensionales

Si se acepta la suposición de que una muestra suficientemente representativa en cada una de las dimensiones, tiene una alta probabilidad de atrapar la información presente en un conjunto de datos perteneciente a un espacio de dimensionalidad mayor, entonces otras pruebas estadísticas no paramétricas unidimensionales podrían utilizarse en vez del criterio de la entropía como medio para identificar un tamaño de muestra suficiente.

Para una sola variable pueden existir un número infinito de estadísticas de prueba que pueden ser calculadas, pero sólo algunas estiman de manera adecuada las características que se desean analizar en el proceso. Las estadísticas de prueba usadas en las pruebas de hipótesis tratan de expresar a través de ellas, la información que pueda presentar evidencia suficiente para rechazar una hipótesis. En FDM la entropía fue propuesta para como estadística de prueba, sin embargo otras estadísticas de prueba (como las de las pruebas no paramétricas presentadas en la sección 2.6.1) pueden representar una alternativa práctica. Las pruebas paramétricas y los tipos de prueba no incluidas en estos resultados fueron

excluidas por no ser adecuadas para su aplicación en poblaciones multimodales y heterogéneas como las que se analizan en minería de datos.

A continuación se presenta un análisis de los resultados que generan a través de su aplicación en datos de distribuciones multimodales observables en casos prácticos. Estas distribuciones son mostradas en la figura 4.1 y corresponden a una mezcla de distribuciones normales y multinomiales.

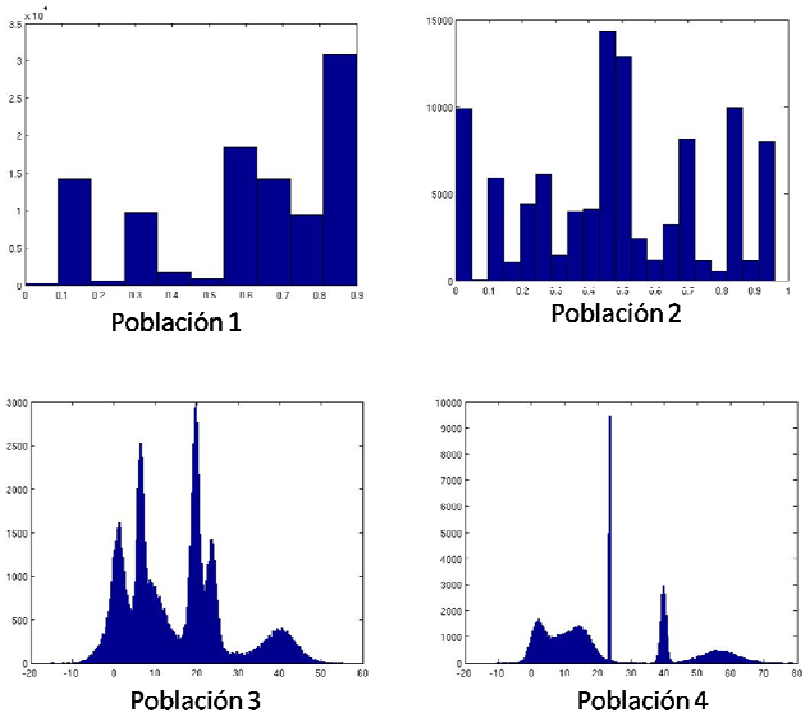

Figura 4.1. Distribuciones usadas para pruebas no paramétricas

Las pruebas no paramétricas siguientes fueron aplicadas en muestras con las distribuciones descritas, en búsqueda de un tamaño de muestra que asegure la representatividad de la muestra bajo la evidencia dada por las estadísticas de prueba.

- Prueba de Kolmogorov-Smirnov
- Prueba de Kruskal-Wallis
- Prueba de bondad de ajuste de χ^2 para dos muestras.

Descripción de la prueba

Dado que el tamaño de la población puede ser alto, se aplicaron las pruebas en conjuntos de muestras de diferentes tamaños iterativamente para identificar el tamaño mínimo aceptado por estas pruebas de manera similar a la que se propone en FDM. Para cada uno de los tamaños de muestra analizados que son mostrados en la tabla 4.1 se extrajeron 5

muestras de cada tamaño y se aplicaron las pruebas no paramétricas mencionadas sobre este conjunto de muestras. Se tomaron de 10 conjuntos de muestras para cada tamaño obteniendo como resultado un conjunto de valores p indicando la probabilidad de que las muestras provengan de una misma distribución en este caso a través de los valores de las distintas estadísticas de prueba. En la tabla 4.1 se reportan los valores máximos y mínimos del valor p obtenidos en las 10 simulaciones realizadas para los conjuntos de datos. Para cada tamaño de muestra se calculan la entropía máxima y mínima en cada muestra con el fin de analizar los tamaños en los que las variaciones de entropía entre muestras, es o no significativa.

Resultados generales de la prueba

El análisis de estos resultados lleva a las siguientes observaciones. En general todos los valores de probabilidad con la que rechazan o no rechazan la hipótesis nula presentan grandes variaciones generando resultados contradictorios, esto se puede ver en las grandes diferencias que existen entre los valores p mínimos y máximos tomados de un mismo tamaño de muestra, siendo fuente de errores de tipo I y II. Cómo se puede ver en todas las pruebas y para todos los tamaños de muestra, no se rechazó la hipótesis nula al menos en una simulación pues siempre el valor p máximo fue cercano a 1, dando evidencia de que frecuentemente se generan errores de tipo II. Si se ignoraran estas variaciones y se tomaran en cuenta sólo los mínimos de los valores obtenidos en las simulaciones, algunas de estas pruebas son capaces de rechazar las muestras no representativas, ya que basado en los valores de entropía presentes en la muestra, deben ser rechazadas, por lo que podrían generar un buen resultado. Sin embargo el uso de estos valores p mínimos aumenta significativamente la probabilidad de cometer un error de tipo I (falso positivo). A continuación se analizan los resultados observados para cada prueba

Tabla 4.1 Resultados de las pruebas no paramétricas

				Kruskal Wallis		Komogorov-Smirnov		Chi cuadrada	
a)Población 1	Tamaño de muestra	Entropía mínima	Entropía máxima	p-value minimo	p-value máximo	p-value minimo	p-value máximo	p-value minimo	p-value máximo
	50	2.5513	2.7703	0.027	0.8822	0.1546	0.9958	0.0628	0.9118
	100	2.4125	2.6802	0.0781	0.9994	0.6766	1	0.1379	0.8939
	200	2.5064	2.7183	0.0559	0.9314	0.1668	0.9996	0.0922	0.7119
	400	2.6297	2.6835	0.1108	0.9986	0.456	1	0.0374	0.9864
	800	2.5832	2.7308	0.138	0.8298	0.1373	0.9969	0.0487	0.9111
	1600	2.6187	2.6678	0.0257	0.9173	0.3614	1	0.0553	0.9264
	3200	2.6348	2.6675	0.07	0.9975	0.3239	1	0.0371	0.967
	6400	2.6325	2.6479	0.1162	0.879	0.233	1	0.0866	0.854
	12800	2.6241	2.6431	0.1423	0.9446	0.5626	1	0.2612	0.964
	25600	2.6408	2.6454	0.2897	0.9643	0.7493	1	0.3531	0.9852
	51200	2.6353	2.6434	0.6944	0.9933	0.8243	1	0.5085	0.9877
	99999	2.6394	2.6394	1	1	1	1	1	1
b)Población 2									
	50	3.5983	3.7574	0.0044	0.836	0.056	0.9541	0.092	0.9324
	100	3.6508	3.9549	0.1684	0.9855	0.3439	0.8938	0.0864	0.9409
	200	3.688	3.8514	0.0191	0.936	0.104	0.9154	0.1272	0.9573
	400	3.7814	3.9256	0.0332	0.9214	0.2348	0.9643	0.0685	0.9261
	800	3.7538	3.8008	0.0919	0.9465	0.0327	0.9931	0.0624	0.9695
	1600	3.7703	3.8429	0.052	0.9591	0.0476	0.9962	0.0882	0.9852
	3200	3.7889	3.8351	0.1451	0.9824	0.1479	0.9994	0.0644	0.974
	6400	3.8059	3.8243	0.5248	0.9516	0.3641	0.9145	0.0426	0.9661
	12800	3.7946	3.8023	0.212	0.9961	0.1764	0.922	0.4026	0.9719
	25600	3.8023	3.8193	0.0605	0.9305	0.4015	1	0.0654	0.9607
	51200	3.8078	3.8117	0.185	0.9755	0.2828	1	0.6786	1
	99999	3.8118	3.8119	1	1	1	1	1	1
c)Población 3									
	50	4.7513	5.0937	0.1532	0.896	0.1546	0.9541	0.3228	0.9766
	100	5.2564	5.4972	0.0271	0.8364	0.047	0.9921	0.1276	0.9858
	200	5.598	5.778	0.024	0.8956	0.0623	0.7787	0.008	0.8918
	400	5.7568	6.0351	0.1906	0.9281	0.1055	0.9996	0.2368	0.9368
	800	5.7865	5.9777	0.0142	0.8976	0.0845	0.8593	0.217	0.974
	1600	5.8176	6.0247	0.199	0.8767	0.296	0.8594	0.1781	0.9829
	3200	5.8849	5.9473	0.1924	0.8704	0.0313	0.987	0.1619	0.9778
	6400	5.7735	5.9278	0.0662	0.9506	0.1655	0.8238	0.5503	0.9987
	12800	5.7228	5.8639	0.1824	0.8642	0.0122	0.8096	0.1933	0.9974
	25600	5.725	5.8085	0.0888	0.9132	0.1799	0.9984	0.6109	0.9998
	51200	5.7069	5.7933	0.491	0.9759	0.3462	0.9916	0.9999	1
	99999	5.706	5.706	1	1	1	1	1	1
d)Población 4									
	50	4.4031	4.9317	0.0768	0.9655	0.0951	0.9541	0.132	0.8058
	100	4.9813	5.3021	0.2442	0.966	0.193	0.961	0.0234	0.9252
	200	5.2729	5.5858	0.1143	0.9175	0.0358	0.9996	0.1192	0.98
	400	5.4454	5.5638	0.0472	0.9601	0.0507	0.937	0.2903	0.9682
	800	5.4715	5.6151	0.0551	0.8518	0.1083	0.9619	0.014	0.9992
	1600	5.4703	5.5697	0.1329	0.923	0.0476	0.9756	0.1585	0.9425
	3200	5.4301	5.573	0.0497	0.9519	0.0477	0.8781	0.0918	0.9238
	6400	5.4063	5.6453	0.1437	0.9522	0.233	0.9145	0.122	0.909
	12800	5.375	5.5041	0.2397	0.9695	0.1967	0.9282	0.1161	0.9963
	25600	5.4185	5.5286	0.0594	0.9226	0.1306	0.9855	0.4815	0.9999
	51200	5.3316	5.4408	0.5808	0.9703	0.2898	0.9998	0.9939	1
	99999	5.4477	5.4477	1	1	1	1	1	1

Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov presenta el principal problema de suponer una distribución estrictamente continua de los datos para poder usar las aproximaciones a la distribución de la estadística. Tiene problemas principalmente al tener un conjunto de datos de tipo discretos como los de distribuciones multinomiales de la tabla 4.1.a y 4.1.b en donde no tuvo la potencia para rechazar las muestras de tamaños pequeños que no son representativas si se toma como valor límite del valor $p = 0.1$. Los valores p en las poblaciones 3 y 4 son menores, mostrando que la potencia de la prueba aumenta al manejar datos continuos, sin embargo en algunos tamaños de muestra pequeños, aún comete errores de tipo II como se muestra en la tabla 4.1.c donde sólo encuentra una probabilidad de 0.15 para rechazar una muestra, por lo que es necesario hacer un mayor número de pruebas para tomar un valor p mínimo del conjunto que haya presentado suficiente información para rechazar la muestra, aumentando con ello la probabilidad de errores de tipo I.

.

Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis y las pruebas cercanas a esta, como la prueba de Anderson-Darling, son basadas en el análisis de desviaciones de rangos, por lo que tiene la deficiencia de enfocarse sólo en la variación de las medianas en los conjuntos de muestras, esto lleva a que no sea una opción adecuada para el análisis de distribuciones multimodales en las que se quiere conservar la distribución. Como se puede observar en la tabla 4.1.a y 4.1.b, esta prueba presenta mejores resultados con las variables con distribución multinomial, sin embargo aún en estos conjuntos de tipo discreto, las variaciones del valor p indican que esta prueba genera valores contradictorios al rechazar muestras de tamaño 25,000 después de no haber encontrado evidencia para rechazar una muestra del tamaño anterior, de 12500 elementos.

Prueba de bondad de ajuste χ^2

La prueba de bondad de ajuste de χ^2 para dos muestras supone la distribución normal en las diferencias de cada intervalo. La forma de asegurar la normalidad de las diferencias en cada intervalo es tomando un número mínimo de datos por clase, que algunos autores [SHESKIN04] sugieren que debe ser superior a 5, sin embargo presenta problemas en una distribución como la presentada en la tabla 4.1.c y 4.1.d al ser conservativa en casos de distribuciones continuas o con un gran número de clases. Además en las distribuciones discretas mostró ser sensible a errores de tipo II al no aceptar muestras que no fueran mayores que la mitad de la población en el caso de la distribución 2.

Conclusiones de estas pruebas

En general estas pruebas tienen el problema de ser conservativas, es decir que no son capaces de rechazar un buen número de hipótesis falsas, ya que pierden potencia para tener la capacidad de hacer un manejo no paramétrico de la información y por la dificultad para encontrar la distribución de la estadística, además el hacer un mayor número de pruebas para evitar estos errores de tipo II trae como consecuencia un aumento en la probabilidad de errores de tipo I.

Enfoques alternos de estas pruebas que estimen la distribución de las estadísticas de prueba de una manera más eficiente, como sería la simulación estadística y los métodos de Monte Carlo, en vez de las distribuciones aproximadas, generan mejores resultados, haciendo un proceso similar al que se usa en FDM usando una estadística de prueba distinta en vez de la entropía, lo cual no representa ninguna ventaja con respecto a la metodología propuesta, ya que requeriría del mismo procesamiento sin mantener la relación directa que posee la entropía tiene una relación directa con los patrones en la población.

Por las razones mencionadas en la sección 3.4.1 y 3.4,2 un análisis unidimensional no es suficiente para validar la conservación de patrones en la muestra. Por ello se requiere un análisis multivariado de la distribución. A continuación se hace un análisis de las características de este análisis.

4.2.2 Análisis multivariado

El análisis de distribuciones multivariadas presenta diferentes dificultades:

- Al aumentar el número de dimensiones en el análisis los problemas de alta dimensionalidad descritos en el capítulo 1.3.
- Algunos análisis no paramétricos que requieren de la ordenación de los datos no son aplicables ya que no se tiene una noción clara del orden de los elementos en espacios multidimensionales, haciendo prácticamente imposible un análisis de rangos como en la prueba de Kruskal-Wallis o para datos ordenados como en la prueba de Kolmogorov-Smirnov.
- Los datos en espacios de mayor cantidad de dimensiones, se encuentran más disperso, haciendo difícil la definición de intervalos que contengan suficientes datos sin pérdidas de información, así como su almacenamiento.

La validación óptima del tamaño de las muestras para agrupamiento sería aquella que se hiciera a través de la obtención de los grupos de datos y comparación con los grupos obtenibles en la población. Esto se puede realizar en conjuntos de tamaño pequeño o

mediano pero no es posible en conjuntos de alta dimensionalidad, ya que el aplicar procesos de minería, incluso aplicándolo a un conjunto de muestras menores para hacer una validación cruzada, puede ser costosa computacionalmente, es por ello en este trabajo se propusieron métodos alternos que tratan de analizar las relaciones entre variables.

La aplicación de la teoría de la aproximación propuesta, es basada en el siguiente criterio. La función de aproximación es una estimación de la distribución de la población de una variable dadas otras variables, de ese modo se está haciendo un análisis de la información mutua de las variables, que no fue verificada en el cálculo de cada variable, buscando así asegurar la entropía conjunta de las variables. De manera similar a la que se hace en un análisis de χ^2, al medir un error de aproximación, se hace una comparación entre dos conjuntos de datos en intervalos, con la diferencia de que por medio del polinomio se utilizan intervalos de mínima longitud ya que no es necesaria la suposición de una distribución en los errores. De modo que el error de aproximación en la norma L_2 es usado como una estadística que mide la diferencia acumulada de los elementos con respecto a un modelo esperado que es el polinomio calculado en un espacio multidimensional, sin requerir la definición de intervalos u orden en los datos.

4.3 Aplicación de FDM

El proceso descrito en la sección 3.4 correspondiente al muestreo y validación del muestreo a través del método FDM, fue implementado y los resultados y detalles de su aplicación son presentados a continuación

4.3.1 Conjuntos de datos de prueba

Se aplicó FDM en diferentes conjuntos de datos representativos para analizar sus resultados. Los conjuntos de datos utilizados para estas pruebas fueron conjuntos de datos de dos y tres dimensiones con distintas distribuciones no triviales. En la figura 4.2 se muestran gráficas con muestras de estas distribuciones. Cada una de estas distribuciones está compuesta por 50,000 instancias. Estos corresponden a grupos de datos cuya representación espacial genera distintas formas a través de la mezcla de distribuciones de la que provienen.

Para verificar su aplicabilidad en conjuntos de datos reales también se presenta el análisis de un conjunto de datos provenientes de una empresa real con 65000 instancias y 37 dimensiones. Las 37 variables están descorrelacionadas linealmente ya que se le aplicó previamente la técnica de reducción de dimensionalidad basada en el análisis correlacional

entre variables. Este conjunto contiene 4 variables con distribuciones binomiales, 4 variables continuas y 29 variables con distribuciones multinomiales.

Todas las variables están codificadas numéricamente en un intervalo de 0 a 1 en el caso de las distribuciones generadas artificialmente y en un intervalo entre -1 y 1 en el caso del conjunto de datos proveniente de una base de datos real, la razón de el uso de un intervalo distinto en el conjunto de 37 dimensiones será discutido después al hacer la validación a través de polinomios de aproximación, ya que permite reducir inestabilidad numérica al usar polinomios.

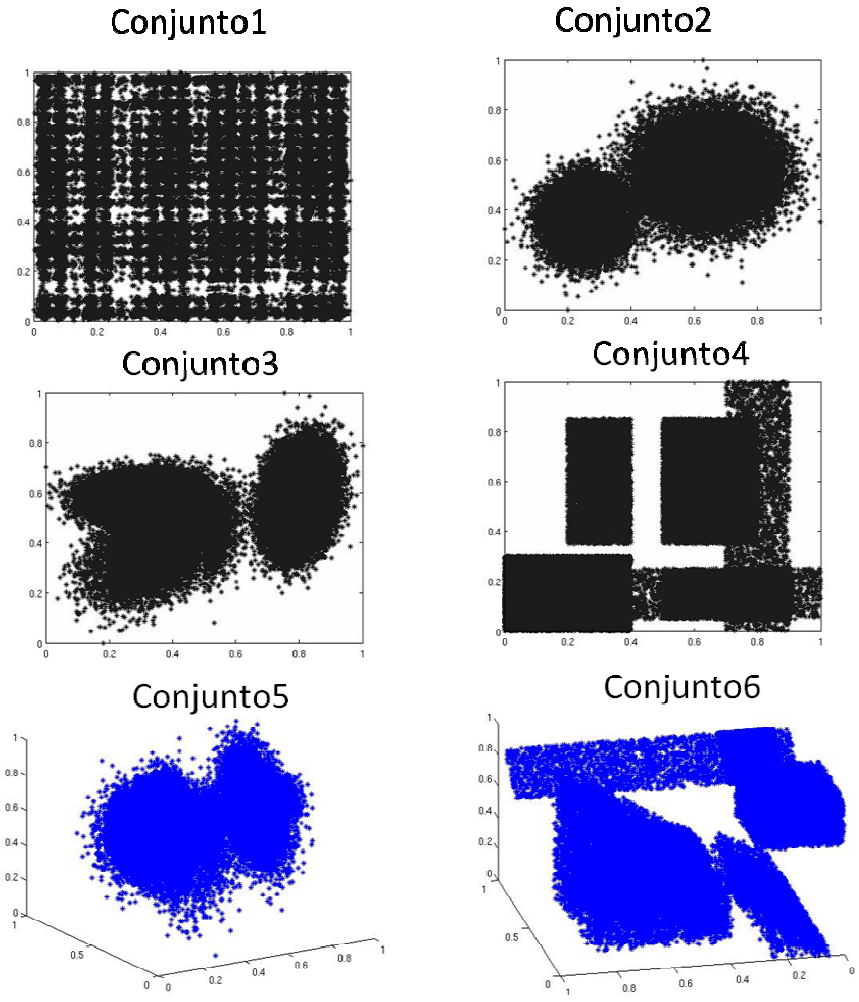

Figura 4.2.Conjuntos de datos generados de dos y tres dimensiones

4.3.2 Parámetros del muestreo y validación

El parámetro ΔH usado fue determinado experimentalmente con base al siguiente criterio. El correcto valor de este umbral para la obtención del tamaño de la muestra $|M|$ es aquel

que permita tener una alta probabilidad selección de muestras con entropía cercana a la de la población. Para la elección de un valor conveniente se debe tomar en cuenta el comportamiento de la entropía calculada para cada tamaño de muestra (figura 3.1). Si se decrementa iterativamente el valor de ΔH y se analiza el tamaño de muestra que exige este umbral, se puede observar que a partir de cierto valor, el tamaño requerido aumenta exponencialmente. Tomar un valor cercano a este punto es una buena opción ya que permite una significativa reducción sin sacrificar demasiada información como se observará en los resultados de las pruebas.

En la etapa de validación de relaciones entre variables existen una mayor cantidad de parámetros para la metodología. Estos parámetros establecen las características de la aproximación que se utilizará, así como el margen de tolerancia que se dará en las diferencias observadas en los resultados. El grado de los polinomios y número de variables usadas en la prueba son discutidos en la siguiente sección.

Los algoritmos usados requieren también usar un número de grupos a priori. En este caso fue fijado sin hacer un análisis exhaustivo de las características del agrupamiento obtenido ya que el objetivo en esta investigación es la validación de los resultados del muestreo con respecto a la población, sin embargo existen criterios experimentales con buenos resultados como criterio del codo de Bezdek [BEZDEK74] que en casos prácticos permite la obtención de un número adecuado de grupos a través de el análisis de la calidad de los grupos formados, a través de una medida de qué tan compacto son los grupos, llamado coeficiente de partición, y una medida de borrosidad en la asignación a cada uno de ellos, llamada entropía de la partición.

4.3.3 Etapa 1: Cálculo del tamaño de muestra

Se calculó para cada uno de los conjuntos de datos de prueba el tamaño de muestra, obteniendo los resultados mostrados en la tabla 4.2. Esto representa una importante reducción de entre el 87% y el 97% para los conjuntos analizados.

Tabla 4.2.Tamaños de muestra calculados para los conjuntos de prueba

Conjunto de datos	Tamaño de la población	Tamaño de la muestra	Eficiencia del muestreo
1	50000	1500	97%
2	50000	1500	94%
3	50000	3000	97%
4	50000	2000	96%
5	50000	2200	96%
6	50000	4400	91%
37 dimensiones	65535	9100	86%

4.3.4 Etapa 2: Validación del tamaño de muestra

4.3.4.1 Validación en conjuntos de dos y tres dimensiones

Las variables en los conjuntos de dos dimensiones conservaron sus relaciones después del muestreo, de acuerdo al criterio establecido para un valor de γ =0.1, como se observa en la tabla 4.3 utilizando un polinomio de grado 9. El grado del polinomio se fijó haciendo la aproximación con polinomios de distintos ordenes y se tomó el grado mínimo en el que el aumentar el grado de los polinomios no provoca una mejora en el ajuste realizado en este caso.

Tabla 4.3. Validación en conjuntos bidimensionales

Conjunto de datos	Valor de R para X1=f(X2)	Valor de R para X2=f(X1)
1	1.07	1.096
2	1.02	1.09
3	1.03	1.095
4	1.06	1.087

Para el caso de los conjuntos 5 y 6 que son tridimensionales, dado que el número de monomios es más elevado, se utilizó el algoritmo genético del modo indicado en la sección 2.9.1 para la búsqueda de monomios significativos para la aproximación polinomial codificando en el genoma de cada individuo el grado de las variables de cada monomio. Se utilizó una codificación binaria de longitud fija incluyendo todos los posibles exponentes de las variables que componen al polinomio, de modo que si el exponente es cero, entonces la variable no es tomada en la evaluación del polinomio. Se manejaron polinomios de grado máximo 16 en la búsqueda y los resultados del coeficiente R, se presentan en la tabla 4.4, mostrando que también las relaciones tridimensionales también fueron conservadas en los conjuntos.

Tabla 4.4. Validación en datos tridimensionales

Conjunto de datos	Valor de R para X1=f(X2,X3)	Valor de R para X2=f(X1,X3)	Valor de R para X3=f(X1,X2)
5	1.04	1.04	1.07
6	1.02	1.025	1.039

4.3.4.2 Aplicación en el conjunto de dimensionalidad mayor

En el conjunto de 37 dimensiones, el análisis de las relaciones es más complejo. Se analizó la muestra haciendo aproximaciones con distintos números de variables a través de un muestreo de conjuntos de variables (véase sección 3.4.4.2 para detalles del número de grupos muestreados). El número de variables que se puede utilizar simultáneamente se encuentra limitado debido a la precisión finita que se tiene en una computadora, ya que por la forma del aproximador que es un polinomio, los valores de todas las variables elegidas son elevadas a ciertas potencias y multiplicadas entre sí, provocando que los datos para hacer la aproximación, después de haber sido evaluados en los polinomios, sean cada vez más cercanos al cero al estar normalizados. Esta es la razón de usar una normalización en un rango de -1 a 1, ya que de ese modo pueden encontrarse menos valores demasiado cercanos a cero. Esto es un efecto de la forma del aproximador utilizado y queda abierta la investigación sobre diferentes formas del aproximador que permitan reducir la sensibilidad al número de variables sin perder calidad en el ajuste. El uso de redes neuronales es motivado en parte para solucionar este tipo de problemas en la aproximación en espacios de alta dimensionalidad, pero sigue siendo interesante la forma de una función de ajuste adecuada, que además, a diferencia de las redes neuronales, genere un modelo que refleja más claramente las interacciones entre variables.

Análisis bidimensional

En la tabla 4.5 se presenta el resultado de la validación propuesta para pares de variables a través de polinomios de distintos órdenes. Gracias a la eficiencia del algoritmo de ascenso usado para el cálculo de los polinomios fue posible tomar todas las posibles parejas de variables

Tabla 4.5 Resultados de validaciones en dos dimensiones.

Grado del polinomio	Número de pares comparados	Número de pares aceptados	Proporción
5	1369	1367	99.9%
7	1369	1365	99.7%
11	1369	1325	96.8%

El resultado de esta tabla indica que las relaciones entre pares de variables se están conservando casi totalmente, sólo con diferencias pequeñas en algunos casos. Haciendo un análisis de estos valores obtenidos se puede observar que la prueba rechaza la muestra en algunos casos cuando se usan grados de polinomio elevados como en el caso de polinomios de grado 11 en el que se rechazó a un 3% de las parejas aproximadamente. Esto se da a

causa de la sensibilidad de los polinomios minimax, ya que presentan fuertes oscilaciones en algunos intervalos que pueden provocar la inestabilidad observada.

El grado del polinomio de aproximación está asociado con la posibilidad de tener un mejor ajuste a los datos. Sin embargo un mayor ajuste a los datos, a pesar de ser algo deseable en la mayoría de los casos, debe ser analizado cuidadosamente ya que el uso de polinomios de orden mayor puede causar fuertes oscilaciones en el aproximador, sobreajustándose sólo a los datos con los que se calculó.

Una de las ventajas que el muestreo propuesto en FDM es que a través del muestreo aleatorio simple algunos valores poco comunes en la población son ignorados; sin embargo de existir un sobreajuste como el mencionado, puede provocar que la prueba altamente sensible a valores atípicos, provocando el rechazo de tamaños de muestra que podrían ser útiles como puede observarse claramente en el ajuste por pares de variables.

Una vez que se obtiene el polinomio de aproximación y se hace el cálculo de la proporción de errores que muestre un error significativo, aún hace falta el análisis de las variables que intervienen en el ajuste en caso de rechazarse la muestra.

El uso de polinomios entre dos variables resultó de gran utilidad a pesar de ser un análisis más limitado dimensionalmente ya que por el uso de algoritmos eficientes, es posible analizar las relaciones de un gran número de combinaciones y en caso de detectarse un rechazo de la muestra, es posible analizar la variable no ajustada. En el caso del conjunto de datos estudiado, se analizaron las relaciones rechazadas por la prueba, coincidiendo en contener variables que presentaban pocas variaciones y algunos valores atípicos aportando poca información para el agrupamiento. La prueba de relaciones entre las variables fue rechazada, pues al usar un polinomio de alto grado sin tomar en cuenta este valor atípico, el valor de la variable dependiente que predice el aproximador está totalmente fuera de rango, causando grandes diferencias en el error calculado. El uso de un polinomio de grado menor disminuye este efecto como se observó en la tabla 4.5, ya que el reducir el grado trajo como consecuencia la aceptación de la muestra.

Análisis con mayor número de variables

La mejor validación de las relaciones se obtiene al aumentar el número de variables que se están aproximando; sin embargo el incremento de variables en el proceso tiene algunas consecuencias no deseadas. Como se explicó antes, por un lado al aumentar el número de variables puede causar inestabilidad numérica por la forma del aproximador, además la forma del aproximador requiere de una mayor cantidad de parámetros y la búsqueda de los parámetros más importantes se vuelve más difícil al ampliarse el espacio de búsqueda. Es por ello que se utilizó un algoritmo de optimización combinatoria eficiente para

solucionar este problema, en este caso se usó un algoritmo genético codificando los exponentes de cada variable en cada uno de los monomios participantes.

Tabla 4.6 Resultado tomando 3 variables independientes

Grado máximo de polinomios	Número de grupos de variables muestreados	Número de muestras aceptadas	Proporción
7	57	47	0.8245614
15	57	40	0.70175439

Tabla 4.7. Resultados de validación tomando 5 variables independientes

Grado máximo de polinomios	Número de grupos de variables muestreados	Número de muestras aceptadas	Proporción
7	59	27	0.45762712
15	59	10	0.16949153

Experimentalmente en el conjunto analizado en este trabajo, se notó que el aumento del número de variables dependientes a más de 6 no es adecuado para un equipo de cómputo común, ya que la forma del aproximador polinomial utilizado exige la multiplicación de un gran número de variables elevadas a potencias que pueden ser altas provocando un problema de precisión al requerir equipos de cómputo con una precisión mucho mayor que la manejable con números de punto flotante de doble precisión.

Como se muestra en la tabla 4.6 Y 4.7 el resultado de la aproximación multivariada indica que las muestras no son aceptadas en un buen número de ocasiones. Dado lo anterior se podría llegar a la conclusión de que las muestras del tamaño calculado no son representativas; sin embargo el análisis hecho con los algoritmos de agrupamiento llevó a encontrar un modelo muy similar como resultado del agrupamiento en la muestra y en la población, como se verá en la siguiente sección. Por lo tanto un análisis más detallado de los resultados es requerido.

El error de aproximación calculado, depende del polinomio usado, el cual puede presentar fuertes oscilaciones como se presentó en los polinomios de orden 11 al hacer la validación bivariada. Sin embargo esas oscilaciones están limitadas por los elementos que se usaron para la aproximación, indicando que al existir una dispersión mayor en los datos al aumentar la dimensionalidad existe una cantidad mayor de elementos atípicos que no se tomaron en el cálculo y al ser calculado el error en otras muestras, estos provocan grandes

variaciones, tomando en cuenta que el grado real del polinomio usado para la aproximación multivariada es la suma de los grados a los que se eleva cada una de las variables, su sensibilidad es similar a la de un polinomio de alto grado.

Estos datos más dispersos no son fundamentales en el agrupamiento, por lo cual la aproximación a través del polinomio minimax en espacios de mayor dimensionalidad resultó más sensible al muestreo que un algoritmo de agrupamiento.

A pesar de lo anterior se puede observar que en grupos de 3 y 4 variables la prueba es superada en un buen número de comparaciones reforzando la validez estimada en el análisis bivariado.

4.4 Reducciones obtenidas al tamaño de la muestra

La proporción de elementos necesarios para la muestra (tabla 4.2) indican una gran eficiencia en el muestreo ya que se conserva aún en el caso del conjunto de datos más complejos sólo un poco más del 10% de los datos y en otros casos más sencillos requiere de menos del 5% de la población.

Sin embargo esta reducción sólo resulta en una mejora de la eficiencia del proceso si el uso de la muestra M permite obtener prácticamente los mismos resultados que en el conjunto de datos total. Para verificar esta eficiencia se realizan las siguientes pruebas consistentes en la aplicación de algoritmos de agrupamiento.

Los algoritmos utilizados fueron c-medias difusas y mapas auto-organizados. Esta verificación consiste en lo siguiente:

- Aplicación del algoritmo de agrupamiento a la población de datos completa para etiquetar cada uno de los elementos con el número del grupo al que pertenecen. El modelo del agrupamiento obtenido se conserva.
- De la misma manera se aplican los algoritmos sobre la muestra y se etiquetan los elementos, conservando el modelo obtenido.
- El modelo de agrupamiento obtenido en la muestra es aplicado a los datos de la población para reetiquetar los elementos.
- El modelo obtenido del conjunto de datos completo es aplicado en la muestra.
- Se evalúa la proporción de elementos que son asignados a los mismos grupos por ambos modelos tanto en la población como en la muestra.

En la tabla 4.8 se muestran los resultados obtenidos de esta prueba, los conjuntos de datos de prueba de dos y tres dimensiones fueron agrupados usando el algoritmo de c-medias

difusas, mostrando que en estos casos la aplicación del muestreo de FDM resultó altamente eficiente.

La aplicación en el conjunto de 37 dimensiones de c-medias difusas no generó una clasificación útil, por lo que se presenta el resultado generado por mapas autoorganizados que fue capaz de distinguir grupos en este caso usando una red de 2x2 neuronas. El resultado de la comparación de los resultados de la aplicación del mapa autoorganizado en la población P y en la muestra M, puede observarse al final de la tabla 4.8.

Tabla 4.8.Diferencias en clasificación usando la muestra y la población

Conjunto de datos	Porcentaje de elementos incorrectamente clasificados
1	2.1%
2	0.46%
3	0.1%
4	1.5%
5	1.4%
6	1.13%
Conjunto de 37 dimensiones	0.94%

4.4.1 Reducciones en uso de memoria

Una de las premisas del uso de esta metodología es el utilizar una cantidad limitada de recursos durante el proceso para asegurar la eficiencia. Por recursos en este caso se refiere a la cantidad de memoria necesaria para almacenamiento temporal en cada una de las etapas de la metodología, ya que esto reducirá el número de accesos a memoria secundaria y facilitará la distribución de la carga del trabajo realizado en el proceso. Se tomará como un valor constante D la cantidad de memoria utilizada por cada dato.

4.4.1.1 Almacenamiento en la etapa de muestreo

Dado que el análisis se hace variable por variable, durante el proceso de muestreo se requerirá el almacenamiento de a lo más $|M|(D)$ que es el tamaño de muestra representativo.
El cálculo de la entropía requiere además el almacenamiento de un conjunto del tamaño del número de símbolos r en la muestra correspondientes a la probabilidad de aparición de cada uno de los símbolos, utilizando así $(D)r$.

La prueba de Monte Carlo requerida para validar la conservación de la entropía en las muestras de tamaño $|M|$, requiriendo así del almacenamiento de conjuntos del mismo tamaño que en la etapa de cálculo de la entropía y determinación de $|M|$.

4.4.1.2 Almacenamiento en la Etapa de validación

Se requiere mantener en memoria el conjunto interno de datos para el cálculo del polinomio minimax, siendo de tamaño NM, correspondiente al número de monomios del aproximador utilizado.

Se requiere un registro en el que se almacene temporalmente un dato del conjunto externo que se actualizará iterativamente en búsqueda de los elementos que no satisfacen la condición minimax.

El algoritmo genético requiere del almacenamiento de un N conjuntos de NM monomios codificados así como su valor de aptitud, donde N es el número de individuos elegidos para la búsqueda.

4.4.1.3 Almacenamiento al ejecutar algoritmos de agrupamiento

El espacio requerido por estos algoritmos es presentado a continuación.

- Para c-medias difusas
  -Matriz de datos de tamaño $|M|VD$
  -Matriz de centros de *clusters* de tamaño VD
  -La matriz de membresías U de tamaño $c|M|D$
- Mapas autoorganizados de Kohonen
  -Matriz de datos de tamaño $|M|VD$
  -Matriz de pesos de tamaño $mnVD$ para una red de dos dimensiones con $m \times n$ neuronas
  -Valores de radio de vecindad, tasa de aprendizaje y época en el proceso

En esta metodología ninguna etapa requiere el uso del conjunto completo de datos, la etapa en la que se requiere una mayor cantidad de memoria es en la de agrupamiento, sin embargo, esa es la etapa objetivo de la reducción, y se requiere el acceso a un conjunto de datos M en vez de la población P resultando en una gran ventaja en cantidad de accesos a memoria secundaria.

4.4.2 Reducción en tiempo de procesamiento

La reducción obtenida es útil si el tiempo que requiere para su procesamiento es menor que el que se requeriría al aplicar al conjunto de datos sin ninguna reducción. Para verificar esta reducción es necesario comparar la cantidad de operaciones que se requiere para hacer el análisis de la población, contra la cantidad de operaciones necesarias para aplicar la reducción aquí propuesta y su posterior análisis. Para esto se analizará la complejidad en cada una de las etapas que corresponden a la reducción de número de instancias así como la reducción en tiempo de procesamiento en los algoritmos, debido a la reducción de número de datos que se requiere procesar.

Se supondrá que la obtención de un elemento de la base de datos toma un tiempo constante, al igual que el tiempo cálculo de un número aleatorio o el efectuar cualquier operación matemática simple.

Etapa de muestreo

El cálculo de la muestra requiere las siguientes etapas
El muestreo de N elementos bajo los supuestos antes especificados requiere de un tiempo de orden $O(N)$.
El cálculo de la entropía en cada muestra de tamaño N requiere de la asignación de cada uno de los elementos al símbolo que le corresponde, requiriendo un procesamiento de $O(Nr)$, donde r es el número de símbolos.
El número de veces que se requiere hacer el cálculo de la entropía depende de $|M|$ y del tamaño de los incrementos ρ que se utilicen. Por lo que el procesamiento para encontrar un tamaño de muestra es de orden

$$O((\frac{|M|}{\rho})(|M| + |M|r))$$

La validación de la entropía requiere el cálculo en un conjunto de Q muestras por lo que tiene una complejidad de $O(Q(|M|r + |M|))$

Etapa de pruebas multivariadas

El cálculo de los polinomios de aproximación toma un tiempo de procesamiento de $O(I(m+1)^2)$ donde I es el número de intercambios realizados al calcular el polinomio minimax, lo cual resulta altamente eficiente dado que m tiene un valor relativamente bajo al tratarse del número de monomios utilizados e I no es grande ya que el intercambio se realiza de manera que siempre se está aproximando a la solución general.

En caso de hacerse la búsqueda de monomios significativos a través del algoritmo genético la búsqueda requiere del cálculo de un polinomio por cada uno de los $Nind$ individuos de

la población de cada generación por lo que requiere un procesamiento de orden $O(GNind|M|(m+1)^2)$ por cada función calculada.

Los polinomios de aproximación deben ser calculados nf veces, donde nf es el número de funciones a validar para asegurar la representatividad de las relaciones en las muestras. Resultando en una complejidad total de orden $O(nf\text{GNind}|M|(m+1)^2)$. En el caso de dos dimensiones en el que no es necesaria la selección de monomios tiene una complejidad de $O(2|M|(m+1)^2)$

Etapa de agrupamiento

El procesamiento a través de un algoritmo de agrupamiento como c-medias difusas que es de orden $O(|M|Vc^2)$ (véase apéndice 2 para el análisis de complejidad), para c grupos, T variables y muestras de tamaño $|M|$, dada una reducción como la obtenida en los conjuntos de prueba cercana al 90%, implica 90% menos procesamiento, lo cual es aún más significativo al aumentar la complejidad del problema, es decir aumentar el número de grupos o de variables analizadas. El mismo resultado se presenta con el uso de algoritmos como los mapas autoorganizados de orden $O(eNcT)$, el tiempo de procesamiento es reducido en un factor igual que la reducción de dimensionalidad obtenida sin sacrificar la calidad del agrupamiento.

Esto quiere decir que si se alcanzó una reducción cercana al 90% del tamaño original, obteniendo una pérdida de precisión que no alcanzó el 5% en los casos analizados, entonces cada dato está proporcionando alrededor de 10 veces más información por cada dato ya que el resultado al que se llega es prácticamente el mismo.

Es importante resaltar además que en ninguna etapa se requirió el procesamiento de la población completa, pudiendo ser aplicada incluso en bases en las que un algoritmo de orden lineal sería inadecuado.

Esta significativa reducción de tiempo de procesamiento proporcional a la reducción en el espacio de almacenamiento es aún más significativa tomando en cuenta factores como los tiempos de accesos a memoria, ya que una muestra reducida reduce el número de accesos a memoria secundaria, que requieren tiempos del orden de milisegundos, manejando una mayor parte del proceso con accesos a memoria principal que requiere tiempos de acceso del orden de nanosegundos.

Esta reducción puede permitir incluso el uso de algoritmos que pudieran ser de $O(n\,logn)$ como DBSCAN o de orden $O(n^2)$ como los algoritmos jerárquicos que requieren mayor procesamiento dado que la cantidad de datos por procesar después de la reducción es

considerablemente menor.

Capitulo 5. Conclusiones

El proceso de minería de datos se enfrenta al manejo de grandes volúmenes de datos con distribuciones difíciles de analizar aún con los equipos de cómputo actuales. Particularmente el proceso de búsqueda de agrupamientos de datos, básico en la minería de datos, que analiza el conjunto de datos en búsqueda de patrones importantes es un proceso de alta complejidad y es por ello que se han propuesto un gran número de posibles soluciones basadas en diferentes heurísticas, y conceptos de diferentes áreas. Una metodología para simplificar este análisis como la propuesta en esta investigación es relevante como guía para futuros análisis.

En los grandes volúmenes de información frecuentemente existe una gran cantidad de redundancia en a información. La información contenida en cada una de las variables es sumamente importante y puede ser suficiente para conservar los patrones y relaciones en espacios de mayores dimensiones que son formados a través de estas variables.

Es importante analizar el detalle con el que se llevará a cabo el análisis, pues en esencia el proceso de agrupamiento de datos es un análisis que busca una descripción más “gruesa” de los datos por lo que demasiado detalle o redundancia en el conjunto de datos por analizar no son necesarios. Es por ello que no toda la información disponible es necesaria en estos algoritmos y se pueden hacer simplificaciones significativas como la reducción de dimensionalidad descrita en esta metodología.

Como en todo análisis computacional además de buscar una solución, se busca que esta solución sea alcanzable en un tiempo lo más corto posible, por lo que cada algoritmo debe ser eficiente, proporcionar resultados útiles y se debe explotar la información de manera suficiente. El muestreo y su validación tratan de usar la información que pudiera ser útil en un análisis de la manera más eficiente posible.

Es interesante resaltar el papel del análisis unidimensional como preparación para un análisis multidimensional como es el agrupamiento, ya que es este análisis el que permite hacer esta reducción del espacio de búsqueda a través de la cantidad de información de los símbolos de cada variable. Este enfoque es una clara demostración de que en una gran cantidad de datos un análisis más sencillo es el que puede proporcionar las características más relevantes en un proceso.

Las hipótesis principales analizadas en este trabajo son las siguientes:

- De existir redundancia en los datos, esta puede ser aprovechada para obtener una muestra de tamaño menor que conserve la información necesaria para el reconocimiento de los grupos de datos.
- Las medidas de teoría de la información como la entropía representan suficientemente la existencia de patrones en el conjunto de datos.
- La conservación de la información en cada una de las variables, puede ser suficiente para conservar la información contenida en los datos de forma multivariada.
- Existen medios para la evaluación y análisis de las relaciones entre las variables para verificar la conservación de relaciones entre variables.

Los resultados observados indican que para distribuciones no triviales como las analizadas, estas hipótesis son justificadas, sin embargo una validación multivariada con suficiente capacidad de aproximación y menor sensibilidad que los polinomios multivariados minimax sería de gran utilidad.

5.1 Resultados de la metodología

El muestreo a través de los métodos aquí utilizados da como resultado

- Reducción del volumen de datos a cerca de un 10% en los casos analizados
- Reducción significativa del número de iteraciones en el algoritmo de agrupamiento proporcional a la reducción en el volumen de datos utilizado.
- Se identificaron prácticamente los mismos grupos de datos en los conjuntos de datos utilizados para comparación de los resultados aplicados a toda la población.
- El cálculo y verificación del tamaño de la muestra en ninguna de las etapas requiere del análisis de toda la población de datos.

5.1.1 Reducción de dimensionalidad y complejidad obtenida

Las reducciones de dimensionalidad que se lograron en los conjuntos de prueba permiten

- Mejorar el desempeño del algoritmo de minería de datos al analizar una menor cantidad de información redundante.
- Reducir tiempo de ejecución, tanto por la reducción de número de iteraciones y por el espacio en memoria y cantidad de accesos a memoria secundaria requeridos.
- Las diferentes validaciones realizadas en el conjunto de datos puede dar luz acerca de características útiles de los datos como lo son variables poco informativas, totalmente aleatorias, o señalar en su caso cuales son aquellas que pueden presentar dificultad en análisis futuros.

5.1.2 Mecanismos de validación

Los algoritmos para la validación utilizados son aplicables con buenos resultados para tipos de datos de distribuciones arbitrarias sin presentar las limitaciones de las pruebas estadísticas tradicionales al no requerir la suposición de una distribución en ninguna etapa del muestreo o validación.

Además de su importancia para la validación de muestras como se presentó, los algoritmos de aproximación de funciones, usados para obtener una medida de similitud entre las relaciones de variables, presentan por sí mismos resultados aprovechables respecto al comportamiento y relaciones entre variables.

5.2 Trabajos a futuro

Queda abierta la investigación para varios aspectos que son manejados en este trabajo. Algunos de ellos son los siguientes

- La aplicación para otro tipo de algoritmos de minería de datos.
- Al hacer la aproximación multivariada a través del algoritmo de ascenso, se pueden explorar diferentes formas del aproximador de manera que se no se tengan limitaciones por la precisión en el equipo de cómputo que se utilice, la forma óptima para un aproximador de este tipo queda pendiente de investigar.
- La exactitud que puede lograrse en la aproximación de acuerdo al grado de los elementos de los monomios seleccionados para la aproximación así cómo un análisis de la distribución de los errores.
- Un análisis más detallado de los monomios seleccionados en el proceso de aproximación ya que estos monomios, obtenidos a través del proceso de optimización combinatoria a través del algoritmo genético, son los que capturan mayor cantidad de información respecto a la relaciones no lineales entre las variables.
- Integración con sistemas de minería de datos para la toma de decisiones o integración con otros algoritmos.
- Determinación de valores adecuados de los parámetros de la metodología, estableciendo la probabilidad de error ocasionado por la reducción realizada. Los parámetros principales a analizar son la diferencia de entropía aceptable ΔH, el número de elementos a usar al hacer validaciones, el grado del aproximador y número de monomios usados en caso de usar un polinomio como aproximador.

Apéndice A. Programas implementados

Nombre de programa	Tipo	Descripción
Adj	Auxiliar para aproximación de funciones	Cálculo de la matriz adjunta
AG	Auxiliar para aproximación de funciones	Implementación del algoritmo genético de Vasconcelos con modificaciones para la aproximación de funciones
AplicValid	Validación de muestras	Aplicación de la segunda etapa de FDM a través de la aproximación de funciones y el cálculo de errores de aproximación haciendo uso del algoritmo de ascenso.
CaracMuestreo	Cálculo de tamaño de muestra	Pruebas para el análisis u caracterización de la etapa 1 de FDM
Chi2test	Pruebas estadísticas	Implementación de la prueba de χ^2 para dos muestras
closest	Auxiliar agrupamiento	Encuentra el más vector más cercano de un conjunto a un vector proporcionado como parámetro
Cofact	Auxiliar para aproximación de funciones	Cálculo de cofactores de una matriz
compCtrSom.m	Validación de agrupamiento	Permite la comparación de los grupos generados por diferentes modelos si estos modelos son descritos a través de centroides
DBSCAN	Agrupamiento	Implementación del algoritmo de agrupamiento DBSCAN
DemoMNMX	Aproximación de funciones	Aplicación de la aproximación de funciones con polinomios minimax para algunos conjuntos de datos
distancia	Auxiliar para agrupamiento	Calculo de distancia euclidiana entre dos vectores
distribDatos	Pruebas estadísticas	Aplicación de pruebas estadísticas no paramétricas a algunos conjuntos de datos generados
entropia	Auxiliar cálculo de tamaño de muestra	Hace el cálculo de la aproximación de la entropía del conjunto univariado dado
evalMultiPoly	Auxiliar aproximación de funciones	Evaluación de polinomios en conjuntos de datos.
fuzzycmeans	Agrupamiento	Implementación del algoritmo de c medias difusas
generaFuncDep	Auxiliar para generación de datos	Auxiliar para la creación de conjuntos de datos dependientes de otro conjunto

generaMultinomiales	Auxiliar para generación de datos	Herramienta para la generación de conjuntos de datos con distribución multinomial con cualquier número de clases parámetros de proporción de elementos en la clase
getFirstIndex	Auxiliar para aproximación de funciones	Auxiliar para aumentar la eficiencia de los elementos elegidos para el primer conjunto interno al hacer la aproximación en L_∞
getMinimaxSign	Auxiliar para generación de datos	Cálculo de los signos para la solución del polinomio minimax
minimax	Aproximación de funciones	Cálculo del polinomio de aproximación minimax con una variable dependiente usando el método de intercambio de Remez
minimaxdd	Aproximación de funciones	búsqueda del polinomio minimax por el método de diferencias divididas para una variable dependiente
minimaxSS	Auxiliar para aproximación de funciones	tamaño de muestra necesario para la evaluación del error externo en la aproximación de funciones minimax
MMNMXE	Aproximación de funciones	Aplicación de ajuste de funciones con selección de monomios a través de un algoritmo genético
MuestreaEntrD.m	Cálculo de tamaño de muestra	Implementación de la primera etapa de FDM correspondiente al cálculo de la muestra con respecto a la entropía de las muestras y su validación a través de simulación.
muestrear	Auxiliar muestreo	toma aleatoriamente un conjunto de elementos del vector
multiminimax	Aproximación de funciones	Cálculo de polinomios minimax para una o más de una variable independiente
normaliza	Preprocesamiento	Mapea valores de un vector a un rango [0,1].
normalizaAB	Preprocesamiento	Mapea valores de un vector a un rango [a,b].
plotClusters	Visualización	Herramienta para la visualización de los grupos generados por el algoritmo de *c medias difusas*
plotCrispClusters	Visualización	herramienta para graficar resultado de un agrupamiento nítido
PruebasBidim	Validación de muestras	Cálculo de tamaño de muestra y validación del conjunto de datos de 37 dimensiones con aproximaciones de dos dimensiones tomando polinomios hasta grado 11, usado para la publicación [LOZANO10]
repetidos	Auxiliar aproximación de funciones	auxiliar para eliminar individuos que representen funciones repetidas en una población para el algoritmo genético implementado

selectedNumberFunction	Auxiliar validación de muestras	Indica el número de funciones que es necesario comparar para asegurar con un porcentaje de confiabilidad que un porcentaje de los elementos de un conjunto poseen una característica. En ese caso es usado para calcular el número de funciones a comparar requerido para asegurar la conservación de las relaciones entre variables
solveminimadd	Auxiliar aproximación de funciones	Solución del polinomio minimax del conjunto interno al hacer aproximación con una variable dependiente por el método de diferencias divididas
SOM	Agrupamiento	Implementación de mapas auto-organizados de Kohonen para el análisis de su complejidad y desempeño.
testDataSets	Auxiliar para generación de datos	Creación de conjuntos de 2 y 3 dimensiones para posteriores pruebas de la metodología FDM.
unsort	Auxiliar para muestreo	Desordena los datos de un vector

Apéndice B. Análisis de complejidad de algoritmos de agrupamiento

Complejidad *c-medias difusas*

A continuación se presenta un análisis del tiempo de procesamiento para el algoritmo de c-medias difusas
La primera etapa del algoritmo consiste en el cálculo de los centros de cada grupo

$$c_j = \frac{\sum_{i=1}^{n} {u_{ij}}^m \cdot x_j}{\sum_{i=1}^{n} {u_{ij}}^m}, \forall j$$

La sumatoria $\sum_{i=1}^{n} {u_{ij}}^m \cdot x_j$ requiere de un procesamiento de orden $O(nV)$ donde n es el número de datos y V el número de variables.
La sumatoria $\sum_{i=1}^{n} {u_{ij}}^m$ requiere de un procesamiento de orden $O(n)$
En conjunto esta operación es de orden $O(c(nV + n)) = O(c\,nV)$,ya que se ejecuta para cada uno de los c grupos

El cálculo de la función de membresía se hace a través de la siguiente operación

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d(x_i, c_j)}{d(x_i, c_k)})^{\frac{2}{m-1}}}$$

Cada cálculo de distancia es de orden $O(V)$, y se realiza para cada grupo por lo que el denominador es de orden $O(Vc)$. Este valor es calculado para cada dato y cada variable por lo que resulta en un procesamiento de orden $O(nVc^2)$.

Si este proceso se repite en i iteraciones la complejidad del algoritmo es de orden

$$O\big(i(cnV + nVc^2)\big) = O(inVc^2)$$

Complejidad mapas autoorganizados

De acuerdo con el entrenamiento presentado en la sección 2.5.2 la complejidad del entrenamiento de la red neuronal es la siguiente:

- Inicialización de pesos : $O(lV)$ para l neuronas y V variables
- Búsqueda de la neurona ganadora tal que

  $$i(x) = \min\big(d(x, w_j)\big), j = 1,2, \dots, l$$

  Lo cual requiere de un procesamiento de orden $O(Vl)$
- Actualización de los pesos de las neuronas

  $$w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(x(n) - w_j(n))$$

  requiere de un procesamiento de orden O(V) para cada neurona por lo que resulta de orden $O(lV)$

- Estas actualizaciones se deben hacer en una época con todos los n datos por lo que es requiere de un procesamiento de orden $O(nlV)$
- El cálculo de la función de vecindad
  $h_{j,i(x)}(n) = e^{-\frac{d(i(x),j)^2}{\sigma^2}}$
  requiere del cálculo distancias entre cada posible par de neuronas lo cual es de orden $O(Vl^2)$

- Estos procesos se repiten un número E de épocas por lo que la complejidad del entrenamiento es de orden $O(E(nlV + Vl^2))$

Apéndice C. Aproximación multivariada a través de polinomios minimax

El problema de aproximación de funciones para minimizar la norma L_∞, al aproximar una variable F en función de un conjunto de variables $X = \{x_1, x_2, x_3, .., x_V\}$, donde V es el número de variables independientes, consiste en minimizar el error de ajuste:

$$L_\infty = \max\left(F - p(x_1, x_2, \dots, x_V)\right) \quad (1)$$

La función p(x) puede tenerla forma

$$p(x) = \sum_{i=1}^{m} c_i M_i(X) \quad (2)$$

Donde $M_i(X)$ es una función de las variables dependientes y c_i es el conjunto de coeficientes que permiten la aproximación.

En el caso de hacer una aproximación polinomial cada monomio tiene la forma

$$M_i(X) = \prod_{j=0}^{V} x_j^{d_j} \quad (3)$$

donde d_j es el exponente al que se desea elevar la variable j en el monomio i.

A través de la selección adecuada de los coeficientes $c_1, c_2, \dots, c_m$, se puede obtener un modelo de aproximación que satisface la condición minimax, la cual consiste en minimizar el error de ajuste en la norma L_∞. Estos coeficientes son obtenidos a partir de un conjunto de datos D consistente en N tuplas de V, elementos donde cada uno de estos corresponde a una variable x_i.

Dado un polinomio univariado en el caso de que $V = 1$ o un polinomio multivariado en el caso en el que $V > 1$ compuesto por m monomios el polinomio de aproximación puede ser obtenido a partir de m+1 datos:

$$\begin{aligned} F_1 - p(X_1) &= e_1 \\ F_2 - p(X_2) &= e_2 \\ &\vdots \\ F_{m+1} - p(X_{m+1}) &= e_{m+1} \end{aligned} \quad (4)$$

Si se expresa en forma matricial en términos de los m monomios $M_{i,j}$ resultantes de evaluar el dato i en el monomio j y los coeficientes c_i $\forall\, i \in [1, \dots, m]$ este sistema

puede ser expresado como:

$$\begin{bmatrix} e_1 & M_{1,1} & M_{1,2} & \dots & M_{1,m} \\ e_2 & M_{2,1} & M_{2,2} & \dots & M_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{m+1} & M_{m+1,1} & M_{m+1,2} & \dots & M_{m+1,m} \end{bmatrix} \begin{bmatrix} 1 \\ c_1 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_{m+1} \end{bmatrix} \quad (5)$$

Se busca minimizar $e_{max} = \max(|e_1|, \dots, |e_{m+1}|)$. Denotando a los errores de aproximación $e_i = \alpha_i e_{max} \mid 0 \leq \alpha_i \leq 1 \quad \forall i \in [1, \dots, m+1]$. el error e_{max} puede ser calculado de la siguiente manera:

$$e_{max} = \frac{\begin{vmatrix} F_1 & M_{1,1} & M_{1,2} & \dots & M_{1,m} \\ F_2 & M_{2,1} & M_{2,2} & \dots & M_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{m+1} & M_{m+1,1} & M_{m+1,2} & \dots & M_{m+1,m} \end{vmatrix}}{\begin{vmatrix} \alpha_1 & M_{1,1} & M_{1,2} & \dots & M_{1,m} \\ \alpha_2 & M_{2,1} & M_{2,2} & \dots & M_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{m+1} & M_{m+1,1} & M_{m+1,2} & \dots & M_{m+1,m} \end{vmatrix}} \quad (6)$$

Por expansión en cofactores

$$= \frac{\begin{vmatrix} F_1 & M_{1,1} & M_{1,2} & \dots & M_{1,m} \\ F_2 & M_{2,1} & M_{2,2} & \dots & M_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{m+1} & M_{m+1,1} & M_{m+1,2} & \dots & M_{m+1,m} \end{vmatrix}}{\alpha_1 \begin{vmatrix} M_{2,1} & M_{2,2} & \dots & M_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m+1,1} & M_{m+1,2} & \dots & M_{m+1,m} \end{vmatrix} + \dots + (-1)^{n+1} \alpha_{m+1} \begin{vmatrix} M_{1,1} & M_{1,2} & \dots & M_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m,1} & M_{m,2} & \dots & M_{m,m} \end{vmatrix}} \quad (7)$$

Dado que se requiere minimizar el máximo error e_{max} y los valores del numerador son fijos al corresponder a los datos, sólo se pueden elegir los valores de α_i adecuados. Para minimizar el error, el denominador debe ser maximizado por lo que se deduce que $|\alpha_i| = 1$ por lo que $|e_1| = |e_2| = \dots = |e_{m+1}|$

Y el signo de cada α_i, $\sigma_i = signo(\alpha_i)$ deben ser elegidos de modo que todos los términos tengan el mismo signo. Una vez obtenido el error absoluto y los dignos de cada término el sistema de ecuaciones puede ser resuelto obteniendo el polinomio minimax del conjunto de datos elegidos.

En un conjunto de N datos donde $N > m + 1$ el polinomio que mejor se ajusta en la norma minimax se obtiene a través de la búsqueda del conjunto de datos cuyo error sea el máximo entre todas las posibles combinaciones de elementos.

Definiendo dos conjuntos, un conjunto interno que contiene a los $m + 1$ elementos que participan en el cálculo de la función de aproximación y un conjunto externo con el resto de los datos disponibles, la forma eficiente de obtención del conjunto que genera el máximo error consiste en hacer intercambios entre el conjunto interno y externo de modo que se garantice la obtención de un error más cercano al máximo buscado. Para esto, una vez que se han calculado los coeficientes de aproximación en un conjunto interno se evalúa el modelo obtenido en los datos del conjunto externo. El dato que genere el máximo error se incluirá en el conjunto interno para acelerar la búsqueda. El dato con el cual se intercambiará en el conjunto interno será aquel que sea más cercano al dato seleccionado en el conjunto externo y con un error del mismo signo para garantizar el avance en la búsqueda.

Una implementación eficiente de este proceso es presentada a continuación.

1. En primer lugar es necesario hacer un mapeo de los datos a los valores de los monomios M_i de modo que la matriz de datos manejada no será la definida por el conjunto D de $V + 1$ dimensiones sino una matriz A de $m + 1$ dimensiones tal que

$$F = AC$$

   Tal que C es un vector de coeficientes, A es la matriz con los monomios evaluados en cada dato y los valores de α_i incluidos y F es un vector con los valores de la variable dependiente. Si se requiere el uso de un polinomio con todos los posibles grados se deben incluir suficientes monomios para incluir todas las posibles potencias a las que se pueda elevar una variable
2. Dado que se requerirá la solución de sistemas de ecuaciones dependencias lineales deben ser eliminadas, una opción para esto es transformar el conjunto de datos como sigue.

$$X_j{}^* = \begin{cases} X_j(1 + \rho\delta_h)\ si\ \ X_j \neq 0 \\ \rho\delta_h \qquad\quad si\ \ X_j = 0 \end{cases}$$

   Donde ρ es una variable aleatoria uniformemente distribuida en un intervalo [0,1] y $\delta_h = 10^{-6}$ un valor suficientemente pequeño para no alterar demasiado la aproximación.

3. Calcular los signos minimax. Para lograr que el denominador de la ecuación (7) sea máximo los signos de todos los términos deben ser iguales por lo que el signo de cada cofactor debe ser calculado. Una posible opción es calcular cada determinante y tomar su signo, sin embargo esto puede ser costoso computacionalmente ya que requiere un procesamiento de $O(m^4)$. Una opción más eficiente es hacer uso del teorema del cofactor:
   Si los elementos de la columna un determinante son multiplicados por los cofactores de una diferente columna y sumados el resultado es cero.

Por medio de este teorema sólo es necesario el cálculo de un cofactor para deducir el resto de ellos a través de un sistema de ecuaciones lo cual requiere un procesamiento de orden $O(m^3)$.

Una tercer opción es el cálculo de la matriz inversa ya que la primer fila de la matríz inversa es proporcional a los cofactores de la matriz por lo que los signos de estos elementos correponden a σ_i. Esta opción requiere también un procesamiento de orden $O(m^3)$ en caso de calcularse la matriz inversa, sin embargo es la más adecuada para este análisis ya que, como se verá en el paso 8, a partir de la segunda iteración la matriz inversa es directamente obtenible a partir de la matriz anterior, por lo que alcanza una complejidad $O(m^2)$ a partir del segundo intercambio.

4. Obtener los coeficientes minimax. Esto implica la resolución del sistema de ecuaciones obtenido, es decir calcular.
$$C = BF$$
donde $B = A^{-1}$

5. Evaluar el modelo obtenido p(X) en cada uno de los datos del conjunto externo.($O((N-m)m)$).

6. Tomar el dato que presenta un mayor error, si el máximo error es menor que el obtenido en el conjunto interno entonces los coeficientes obtenidos son la solución. Si no, entonces se intercambiará algún elemento por este dato.

7. Intercambiar el dato encontrado. Se busca hacer el intercambio por el elemento que conserve el signo del error encontrado en el conjunto externo y aumente el error en el siguiente polinomio calculado en el conjunto interno, esto quiere decir que se cambiarán los términos del denominador de la ecuación (7) , es decir los cofactores, de modo que se obtenga un denominador mayor. Esto se expresa de la siguiente manera.

   Sea v el dato con error máximo en el conjunto externo con el signo de su error σ_ϕ en su primer componente, y un vector λ tal que
$$\lambda = vB$$
   λ relaciona los valores de la matriz A con el elemento que se desea intercambiar. dado que el primer renglón de B es proporcional a los cofactores, se busca sustituir por aquel dato que maximice la proporción $\sigma_\phi \frac{\lambda_j}{B_{1,j}}$ donde $B_{1,j}$, es el elemento del renglón 1 de la columna j de la matriz inversa de A, ya que este maximizará el cambio en el error y conservará el signo de los errores de cada dato. El índice j que haya maximizado la proporción es el índice del elemento que deberá ser sustituido. Este intercambio requiere un tiempo de procesamiento de orden $O(m)$

8. Cálculo de la siguiente matriz inversa. El proceso antes usado puede ser repetido cada intercambio, sin embargo algunas técnicas de algebra lineal pueden ser

utilizadas para reducir el tiempo de procesamiento. En vez de calcular la matriz inversa del nuevo conjunto interno, puede obtenerse la nueva matriz inversa a través de la matriz original como sigue. Sea A una matriz no singular, B su inversa y $B_1, B_2, \dots, B_m$ sus columnas y sea $\bar{A}$ la matriz obtenida por remplazar el $\beta - ésimo$ renglón por un vector v si todos los componentes de $\lambda = vB$ son diferentes de 0 entonces $\bar{A}$ es una matriz no singular y las columnas de su inversa están dadas por $\overline{B_\beta} = \frac{B_\beta}{\lambda_\beta} \quad y \quad \bar{B} = B_j - < v, B_j > \overline{B_\beta} \; \forall j \neq \beta$

Esto permite obtener la siguiente matriz inversa en tiempo de $O(m^2)$ y adicionalmente evita el cálculo de los signos en la siguiente iteración ya que estos están dados por el primer renglón de la matriz $\bar{B}$.

9. Regresar al paso 4 y repite hasta encontrar el conjunto que produce el máximo error.

Este proceso tiene una complejidad en tiempo de procesamiento de orden $O(I(m^2Nm))$ donde I es el número de intercambios de datos realizados donde m por lo general es un número pequeño, siendo así un algoritmo altamente eficiente al no tener que solucionar sistemas de grandes dimensiones.

Adicionalmente otras optimizaciones pueden ser usadas para reducir el número de intercambios. Algunas de ellas son las siguientes.

- Establecer un umbral para aceptar el modelo generado por un conjunto interno. En vez de buscar el conjunto interno que genere el máximo error entre todos los elementos, se puede establecer un umbral K tal que si e_θ es el error en el conjunto interno y e_ϕ es el error en el conjunto externo se pueda aceptar el modelo en el que $1 - e_\phi / e_\theta \leq$ K.
- Selección inicial de puntos dispersos. Consiste en la búsqueda de elementos iniciales que se encuentren lo más separados entre sí para iniciar la búsqueda del conjunto minimax con un avance inicial mayor.
- Muestreo en el conjunto externo. La parte de mayor procesamiento puede ser la evaluación del modelo en el conjunto externo entero. Una opción aceptable es calcular el error en un subconjunto que asegure con cierta confiabilidad que se está analizando a una parte suficiente del conjunto externo para hacer una buena aproximación.

Referencias

AGGARWAL99	Aggarwal, C. C., Procopiuc, C. M., Wolf, J. L., Yu, P. S., Park, J. S. Fast algorithms for projected clustering. Proceedings of the ACM International Conference on Management of Data (SIGMOD). 1999
AGRAWAL94	Agrawal, R., Srikant, R. Fast algorithms for mining association rules. ACM International Conference on Management of Data (SIGMOD) 1994
ASSENT07	Assent, I., Krieger, R., Muller E, Seidl, T. DUSC: Dimensionality unbiased subspace clustering. Proceedings of the 7th International Conference on Data Mining (ICDM).2007
BANFIELD93	Banfield J., Raftery A. Model Based Gaussian and non Gaussian Clustering. Biometrics Vol 43, 1993
BARLETT01	Bartlett J., Kotrlik J. K., Higgins C.C. Organizational Research: Determining Appropriate Sample Size in Survey Research. Information Technology, Learning, and Performance Journal, Vol. 19, No. 1, Spring 2001
BEN05	Ben-Gal I. Outlier Detection. Data Mining and Knowledge discovery Handbook, Chapter 7, 2005
BERKHIN06	Berkhin, P. A Survey of Clustering Data Mining Techniques. In Grouping Multidimensional Data. Springer-Verlag: Berlin/Heidelberg, 2006.
BEZDEK74	Bezdek J.C. Cluster validity with fuzzy sets. Journal of Cybernetics 3, 58–72, 1974
BOHM04	Bohm, C., Kailing, K., Kriegel, H.-P., Kroger, P. Density connected clustering with local subspace preferences. Proceedings of the 4th International Conference on Data Mining (ICDM), 2004
BRADLEY98	Bradley P., Fayad U. Refining Initial Points for K-Means Clustering. Proceedings of the 15th International Conference on Machine Learning (ICML98)
CABANES09	Cabanes, G.; Bennani, Y.; Dufau-Joel, F Mining Customers' Spatio-Temporal Behavior Data Using Topographic Unsupervised Learning. Machine Learning and Applications, 2009. ICMLA '09.
CASSELA02	Cassela G., Berger R. Statistical Inference. Duxbury Tompson Learning, 2a edición 2002
CHEESEMAN96	Cheseman P., Stutz J. Bayesian Classification (AutoClass): Theory and results. Advances in Knowledge Discovery in Data Mining 1996
CHENEY98	Cheney E.W. Introduction to Approximation Theory.. AMS Chelsea Publishing (1998).
CHENG06	Cheng, D., Kannan, R., Vempala S., Wang G. A divide-and-merge methodology for clustering.. ACM Trans. Database Syst., Vol. 31, No. 4. (2006), pp. 1499-1525.
CHENG99	Cheng C. H., Fu A. W. Zhang Y. Entropy-Based subspace clustering for mining numerical data. Proceedings of the 5th ACM International Conference on Knowledge Discovery and D.M.1999
COCHRAN77	Cochran, W. G.. Sampling techniques . New York: John Wiley & Sons. (1977). 3rd ed.
DEMPSTER77	Dempster A.P., Laird, N. Rubin D.B. Maximum Likelihood from Incomplete Data Using EM algorithm. Journal of the Royal Statistical Society, 1977
DHILLON01	Dhillon I., Modha D. Concept Decomposition for Large Sparse Text Data using Clustering. Springer, Machine Learning 42 pp. 143-175, 2001.
DOMENICONI07	Domeniconi C., Gunopulos D. Locally adaptive metrics for clustering high dimensional data. Data Mining and Knowledge Discovery. Vol 14, 2007
DUNN73	Dunn J. C. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separated Clusters. Journal of Cybernetics, 3, 32-57. 1973
ESTER96	ESTER, M., KRIEGEL, H.-P., SANDER, J., AND XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd ACM InternationalConference on Knowledge Discovery and Data Mining (KDD).1996
FANG06	Fang Li; Mehlitz, M.; Li Feng; Huanye Sheng; Web Pages Clustering and Concepts Mining: An approach towards Intelligent Information Retrieval. Conference on Cybernetics and Intelligent Systems IEEE, 2006
FAYAD96	Fayyad,U., Piatetsky-Shapiro, G., Smyth, P From Data Mining to Knowledge Discovery in databases. MIT Press, AI magazine(1996)
FISSHER87	Fisher, Douglas H. Improving inference through conceptual clustering. Proceedings of the 1987 AAAI Conferences
FLAKE70	Flake R.H. Computer-aided classification techniques for biological taxonomy. Adaptive Processes (9th) Decision and Control, 1970.
FRALEY1998	Fraley C., Rafteri A. How many clusters? Which clustering method? Answers Via Model Based Cluster Analysis. Technical Report 329 Department of Statistics University of Washington, 1998
FRIEDMAN04	Friedman J. H., Meulman, J. J. Clustering objects on subsets of attributes. Royal Statist. Soc. Series B (Statistical Methodology) 66, 4, 825–849. 2004
GRZYMALA05	Grzymala-Busse J., Grzymala-Busse W. Handling Missing Attribute Values. Data Mining and Knowledge discovery Handbook, Chapter 3, 2005

GUHA98	Guha S., Rastogi R. CURE. An Efficient Clustering Algorithm for Mining Association Rules in Large Databases. ACM SIGMOD International Conference on Management Data, 1998
GUO04	Guo H,, Hou W..C.,Yan, F., Zhu, Q. A Monte Carlo Sampling Method for Drawing Representative Samples from Large Databases. Proceedings of the 16th International Conference on Scientific and Statistical Database Management (2004)
GUYON03	Guyon, I. Elisseeff, A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3. (2003)
HAM01	Ham, J., Kamber, M Data Mining: Concepts and Techniques. Morgan Kaufman Publishers, 2001.
HAYKIN99	Simon Haykin Neural Networks. A Comprehensive Foundation, Capítulo 9. Prentice Hall(1999)
HOPPNER00	Hoppner F., Klawoon F. et al. Fuzzy Cluster Analysis. Wiley, 2000
HUAN02	Huan Liu, Hiroshi Motoda. On Issues of Instance Selection. Data Mining and Knowledge Discovery. ACM, Vol 6. Issue 2 (2002), 115-13
JAGADISH01	Jagadish, H.V., Larkshmanan, L.V.., Srivastava, D. Snakes and Sandwiches: Optimal Clustering Strategies for a Data Warehouse. ACM Proceedings: International Conference on Management of Data, Philadelphia, USA, pp.37–48. (2001)
JENSEN01	Jensen, S Mining Medical Data for Predictive and Sequential Patterns. PKDD Volume: 9 2001
JIANWEI98	Jiangwei Han Towards On-line Analytical Mining in Large Databases. ACM Sigmod Record 27, 1998
JOLLIFFE02	Jolliffe I.T. Principal Component Analysis.. Springer Series on Statistics, 2 nd Edition (2002)
KAILING04	Kailing, K., Kriegel, H., Kroger P. Density-Connected subspace clustering for high-dimensional data. Proceedings of the 4th SIAM International Conference on Data Mining,2004
KOHONEN90	Kohonen T. The self-organizing map. Proceedings of the IEEE. Vol 8, No.9, 1990
KRANTZ02	Krantz, S.; Harold R., Parks, A Primer of Real Analytic Functions. Second ed., Birkhäuser,2002
KRIEGEL09	Kriegel, H.P. Kroger, P., Zimekm, A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering.. ACM Trans. Knowl. Discov. Data. Article 1 (2009),
KRUSKAL52	Kruskal W. H., Walis A. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 1952
KURI02	Kuri-Morales Á. A Metodology for the Statistical Characterization of Genetic Algorithm. Springer Verlag, Págs 79-88, 2002
KURI07	Kuri-Morales Á. Rodriguez E. F. A Search Space Reduction Methodology for large Databases: A Case Study. 7th Industrial Conference on Data Mining. Springer LNAI (2007).
KURI10	Kuri-Morales Á. Aldana-Bobadilla E. Finding irregularly shaped clusters based on Entropy. Springer-Verlag LNAI 6171 ICDM 2010
LENTH01	Russel V. Lenth Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, Vol. 55, No. 3 (Aug., 2001), pp. 187-19
LEVENBERG44	Levenberg K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. The Quarterly of Applied Mathematics 2: 164–168, 1944
LIU00	Bing Liu, Yiyuan Xia, Philip S. Yu Clustering through decision tree construction. Proceedings of the ninth international conference on Information and knowledge management, 2000
LOZANO10	Alexis Lozano, Angel Kuri-Morales Sampling for Information and Structure Preservation when Mining Large Data Bases. Proceedings of IBERAMIA 2010
MAHALANOBIS36	Mahalanobis P. On the generalized statistics. Journ. Asiat. Bengal, 1936
MAIMON05	Maimon O., Rockach Introduction to Knowledge Discovery in Databases. Data Mining and Knowledge discovery Handbook, Chapter 1, 2005
NAGESH01	Nagesh H., Choudary A. Adaptive grids for clustering massive data sets. Proceedings of the 1st SIAM International Conference on Data Mining (SDM)
NG02	Ng, R.T. ; Jiawei Han ; CLARANS: a method for clustering objects for spatial data mining. Knowledge and Data Engineering, IEEE Transactions on 2002
OLKEN94	Frank Olken, Doron Rotem Random Sampling from Databases - A Survey. Statistics and Computing, Vol 5, 1994
PALMER00	Palmer, P., Floutsos, C Density Biased Sampling: An Improved Method for Data Mining and Clustering. Proceedings of ACM SIGMOD International Conference on Management of Data, 2000
PENTTI04	Pentti Minkkinen Practical applications of sampling theory. Chemometrics and Intelligent Laboratory Systems 74, Elsevier, 2004
PITT07	Pitt, E., Nayak, R. The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset.. Proc. 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM 2007).
POWELL08	Powell, N. ; Foo, S.Y. ; Weatherspoon, M. Supervised and Unsupervised Methods for Stock Trend Forecasting. System Theory, 2008. SSST 2008

ROKACH05	Rokach L., Maimon Oded Clustering methods. Data Mining and Knowledge discovery Handbook, Chapter 15, 2005
SAUNDERS89	Saunders, I. Restricted stratified random sampling. International Journal of Mineral Processing, Vol. 25, No. 3-4., pp. 159-166. (1989)
SHANNON48	Shannon C. E. A Mathematical Theory of Communication. Bell System Technical Journal, vol. 27, pp. 379-423, 1948
SHESKIN04	Sheskin D. Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC, 2004
SLAGLE75	Slagle, J. R., Chang, C. L. and Heller, S. A Clustering and data-reorganization algorithm. IEEE Trans. on Systems, Man and Cybernetics, (1975)
TANG05	Tang, W. & Mao, K. Feature Selection Algorithm for Data with Both Nominal and Continuous Features. Advances in Knowledge Discovery and Data Mining. Springer Berlin / Heidelberg.(2005)
WALLACE94	Wallace C., Dowe D.L. Intrinsic classification by mml-The snob program. 7th Australian Joint Conference on artificial Intelligence 1994
WOO04	Woo, K.-G., Lee, J.-H., Kim, M.-H., Lee, Y.-J FINDIT: A fast and intelligent subspace clustering algorithm using dimension voting. Inf. Softw. Technol. 46, 4, 255–271. 2004
YAN05	Yan Y., Webb G., Wu X. Discretization Methods. Data Mining and Knowledge discovery Handbook, Chapter 6, 2005
YIP05	Yip, K. Y., Cheung, D. W., Ng, M. K. On discovery of extremely low-dimensional clustersusing semi-supervised projected clustering. Proceedings of the 21st International Conference on Data Engineering (ICDE).2005
YU04	Yu Lei, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, 2004
ZADEH65	Zadeh L. Fuzzy Sets. Information and Control; 8: 338–353. 1965
ZAHN71	Zahn C. T. Graph-Theoretical methods for detecting and describing gestalt clusters. IEEE Trans. on Computer,1971
ZHANG09	Zhang, Y., Zhang J., Ma, J., Wang, Z. Fault Detection Based on Data Mining Theory.. Intelligent Systems and Applications, pp 1-4. (2009).
ZHANG96	Zhang T., Ramakrishnan R., Livny M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD 1996
ZHU05	Zhu, L. Nonparametric Monte Carlo Tests and Their Applications. Springer Science+Business Media, Inc. (2005).