Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

MINERÍA DE DATOS
UNA HERRAMIENTA PARA LA TOMA DE DECISIONES

Neftalí de Jesús Calderón Méndez

Asesorado por el Ing. Edgar Mauricio Lone Ayala

Guatemala, abril de 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA

MINERÍA DE DATOS
UNA HERRAMIENTA PARA LA TOMA DE DECISIONES

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA

POR

NEFTALÍ DE JESÚS CALDERÓN MÉNDEZ

ASESORADO POR EL ING. EDGAR MAURICIO LONE AYALA

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, ABRIL DE 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	
VOCAL II	Ing. Amahán Sánchez Álvarez
VOCAL III	Ing. Julio David Galicia Celada
VOCAL IV	Br. Kenneth Issur Estrada Ruiz
VOCAL V	Br. Elisa Yazminda Vides Leiva
SECRETARIA	Inga. Marcia Ivonne Véliz Vargas

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Sydney Alexander Samuels Milson
EXAMINADOR	Inga. Virginia Victoria Tala Ayerdi
EXAMINADOR	Ing. Pedro David Tzoc Tzoc
EXAMINADOR	Ing. César Augusto Fernández Cáceres
SECRETARIO	Ing. Carlos Humberto Pérez Rodríguez

HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

MINERÍA DE DATOS
UNA HERRAMIENTA PARA LA TOMA DE DECISIONES

Tema que me fuera asignado por la Coordinación de la Carrera de Ingeniería en Ciencias y Sistemas en febrero de 2004.

Neftalí de Jesús Calderón Méndez

Guatemala, abril de 2006

AGRADECIMIENTOS

A Dios

Por estar presente en todos los momentos de mi vida, iluminándome el camino y por la oportunidad que me ha brindado de seguir mi formación profesional

A mis padres

Felipe de Jesús Calderón Pérez
Ethelvina Amanda Méndez Ramírez de Calderón

Por ser la fuente de mi inspiración y motivación para superarme cada día más y por inculcar en mí el sentido de responsabilidad.

A mis catedráticos

Por su dedicación y orientación académica, les agradezco sus enseñanzas.

Al Ingeniero Edgar Mauricio Lone

Gracias por su apoyo en la elaboración de este trabajo.

A mis amigos

Gracias por su amistad incondicional, por el apoyo y por estar siempre en las buenas y en las malas.

DEDICATORIA A:

Guatemala

Que esta profesión y mis conocimientos adquiridos puedan ser de utilidad para tu crecimiento.

Mis abuelos

Porque desde el espacio infinito han sido testigos de mi esfuerzo, empeño, alegrías y desesperaciones que hoy culminan con este acto.

Mi padre

De quien no tengo otro sentimiento que el orgullo de ser su hijo, porque con ejemplos me ha enseñado a vivir y enfrentar los problemas siempre con la frente en alto.

Mi madre

Ejemplo de constancia, perseverancia y fortaleza, que siempre con su amor y ternura me ha apoyado y acompañado en mis noches de desvelo.

Mis hermanos

Por su constante compañía, su apoyo incondicional, por las alegrías y tristezas que juntos hemos convivido,

Mis sobrinos

Que este triunfo sea un ejemplo para ustedes, y que un día no muy lejano yo sea espectador de sus propios logros académicos.

Mis compañeros

Para que sigan adelante y logren llegar a la meta que todos nos hemos propuesto desde que nos iniciamos en esta carrera universitaria.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

RESUMEN

Tratar de encontrar patrones, tendencias y anomalías es uno de los grandes retos de vida moderna. Código de barras, automatización de procesos en general, avances tecnológicos en almacenamiento de información y abaratamiento de precios en memoria, son algunos de los factores que han contribuido a la generación másiva de datos.

Las técnicas tradicionales de análisis de información no han tenido un desarrollo equivalente y la velocidad en que se almacenan datos es muy superior a la velocidad en que se analizan. Se cree que se está perdiendo una gran cantidad de información y conocimiento valioso que se podría extraer de los datos.

La minería de datos es un conjunto de herramientas y técnicas que por medio de la identificación de patrones extrae información de las bases de datos, una gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y data warehousing, con técnicas de aprendizaje automático y de estadística.

Para descubrir conocimiento de la información se pueden utilizar varias formas de análisis por medio de las cuales se puede llegar a identificar patrones y reglas en los datos para luego crear escenarios, esta información se puede representar por medio de modelos matemáticos sobre datos históricos y con esto se crea un modelo de minería de datos. Después de haber creado un modelo de minería de datos, se puede examinar nueva información a través del modelo evaluando si se apega a los patrones o reglas definidos.

OBJETIVOS

GENERAL

Evaluar el uso de la Minería de Datos como una herramienta que sirva para la toma de decisiones a nivel gerencial.

ESPECÍFICOS

1. Determinar en que consiste la minería de datos.

2. Identificar las técnicas de minería de datos para la exploración de los mismos.

3. Establecer si el proceso de minería de datos permite generar conocimiento de la información analizada, para la toma de decisiones a nivel gerencial.

INTRODUCCIÓN

La minería de datos es una nueva tecnología muy poderosa con un gran potencial para ayudar a las compañías a enfocarse en la información más importante en sus bases de datos o almacenes de datos. Las herramientas de minería de datos predicen comportamientos, permitiendo a los gerentes y empresarios ser más eficientes en la toma de decisiones y el manejo del conocimiento.

La perspectiva automatizada de análisis que ofrece la minería de datos va mas allá del análisis de eventos pasados y puede responder a preguntas gerenciales que antes consumían demasiado tiempo responder.

La tecnología actual como los códigos de barras, la automatización de procesos, los avances en técnicas de almacenamiento de información y los precios bajos de los dispositivos de almacenamiento, permite capturar y almacenar grandes cantidades de información.

En la actualidad, alrededor del mundo, se ha estimado que el crecimiento de los datos almacenados en las bases de datos se duplica cada 20 meses, mientras que la técnicas de análisis de información no han tenido un desarrollo equivalente, dicho en otras palabras, la velocidad en que se almacena la información es muy superior a la velocidad en que se analizan.

Existe un gran interés comercial por explotar los grandes volúmenes de información, pero no saben de qué forma se puede transformar toda esa información en conocimiento o sabiduría que apoye, efectivamente, la toma de decisiones, especialmente, a nivel gerencial.

La minería de datos es un conjunto de procesos y técnicas o algoritmos que permiten extraer el conocimiento a partir de la información almacenada en grandes bases de datos. Las bases de datos fueron desarrolladas para almacenar datos y más datos significa, más información, de manera que mientras exista más información se puede obtener más conocimiento.

1. MARCO TEÓRICO

1.1 Minería de datos

1.1.1 Definición

La minería de datos es un conjunto de herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información interesante, novedosa y potencialmente útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones.

Si se analiza la definición anteriormente descrita, se dice que la minería de datos es un conjunto de herramientas y técnicas, una gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y *data warehousing*, con técnicas de aprendizaje automático y de estadística.

Para descubrir conocimiento de la información se pueden utilizar varias formas de análisis por medio de las cuales se puede llegar a identificar patrones y reglas en los datos para luego crear escenarios, esta información se puede representar por medio de modelos matemáticos sobre datos históricos y con esto se crea un modelo de minería de datos. Después de haber creado un modelo de minería de datos, se puede examinar nueva información a través del modelo evaluando si se apega a los patrones o reglas definidos.

1.2. Datos, información y conocimiento

1.2.1 Datos

Los datos son en esencia números o texto que puede ser procesado en una computadora, en la actualidad las organizaciones acumulan grandes cantidades de datos en distintos formatos y en distintas bases de datos, entre las que se incluyen datos operacionales o transaccionales en las que se almacenan costos, ventas, inventarios, contabilidad, etc.

1.2.2 Información

Los patrones, asociaciones o relaciones entre los datos proporcionan información, por ejemplo el análisis de transacciones de un punto de venta nos pueden dar información sobre que cantidad de productos se han vendido y durante cuanto tiempo.

1.2.3 Conocimiento

La información puede ser convertida en conocimiento partiendo de patrones históricos

1.3 Origen de la minería de datos

Las técnicas de minería de datos son el resultado de un largo proceso de investigaciones. Esta evolución se inicia cuando se empieza a almacenar la información de organizaciones en las computadoras, continúa con mejoras en el

acceso de datos y más recientemente se han generado tecnologías que los usuarios naveguen por la información en tiempo real. La minería de datos toma este proceso evolutivo y va más allá del acceso retrospectivo de los datos y navegación analizando la información para luego mostrar resultados.

La minería de datos está lista para la aplicación en la comunidad de los negocios ya que ahora cuenta con un soporte de tres tecnologías que la hacen suficientemente madura:

- Recopilación de datos de forma masiva
- Computadoras poderosas con multiprocesadores
- Los algoritmos de minería de datos

Las bases de datos comerciales están creciendo a proporciones sin precedentes y conjuntamente la necesidad de motores de búsqueda mejorados. Los algoritmos de minería de datos personalizan técnicas que han existido por lo menos desde hace 10 años, pero que hasta ahora han podido ser implementadas de tal forma que se consideran herramientas confiables y que consistentemente ejecutan los métodos estadísticos más antiguos.

En la evolución de la forma en que se manejan los datos de los negocios, cada nuevo paso que se construye sobre la base la uno previo, por ejemplo, el acceso dinámico de datos es crítico para las aplicaciones de navegación de datos y la habilidad de almacenamiento de grandes bases de datos es crítica para la minería de datos.

Desde el punto de vista de los usuarios, el manejo de la información ha evolucionado y ha permitido responder nuevas preguntas sobre los negocios de manera rápida y exacta, a continuación se muestran los pasos evolutivos que se han dado:

Tabla I. Pasos evolutivos de la minería de datos

Paso evolutivo	Pregunta
Recopilación de datos (1960s)	Cual ha sido mi ingreso en los últimos 5 años?
Acceso de datos (1980s)	Cuales han sido las unidades de venta en Nueva Inglaterra el mes pasado?
Data Warehousing y soporte de decisión (1990s)	Cuales han sido las unidades de venta en Nueva Inglaterra el mes pasado? Durante el entrenamiento en Boston
Minería de datos	Que puede suceder el mes que viene en las ventas de Boston? Y por que?

Los componentes que conforman el núcleo de la minería de datos han estado en desarrollo durante décadas, áreas tales como estadística, inteligencia artificial y máquinas de aprendizaje. Hoy la madurez de estas técnicas acopladas al alto desempeño de los motores de búsqueda de bases de datos relacionales y los esfuerzos de integración de datos, hacen estas tecnologías prácticas para los entornos del Data warehousing

1.4 Áreas relacionadas con la minería de datos

1.4.1 Minería de datos y almacenes de datos (data warehouse)

Frecuentemente los datos que serán minados se extraen del data warehouse de una empresa. Existe un beneficio real si los datos son parte ya de un data warehouse esto es porque el proceso de depuración de datos para un data warehouse y para la minería de datos son similares.

Si la mayor parte de los datos ya han sido limpiados para un data warehouse, es muy probable que no se necesite una limpieza adicional para minar los datos.

La base de datos que se va a minar puede ser un subconjunto lógico en lugar de uno físico del data warehouse, esto sólo si el DBMS del data warehouse soporta los recursos de demanda de minería de datos.

Si, el data warehouse no posee soporte de minería de datos, entonces se tendrá que separar las base de datos a minar en:

- Geográfica
- Data Mart
- Análisis
- Data mining
- Data source
- Data warehose

El data warehouse no es un requerimiento indispensable para la minería de datos, la configuración de un data warehouse de múltiples datos, resuelve problemas de integridad de datos y una consulta de los datos por medio de un query, puede ser una tarea enorme que puede tomar mucho tiempo y a un costo elevado. Sin embargo para minar los datos de una base de datos operacional o transaccional se puede cargar la información a una base de datos de sólo consulta, este proceso es muy parecido al de un data mart.

1.4.2 Minería de datos y OLAP

Una de las preguntas más frecuentes entre los profesionales en informática es ¿cuál es la diferencia entre la minería de datos y OLAP? Pues son dos herramientas muy diferentes, pero que se pueden complementar mutuamente.

OLAP es un espectro de las herramientas de soporte de decisiones. Las consultas tradicionales en sql y las herramientas para reportes solo muestran lo que está en una base de datos, OLAP va un poco más allá y responde porqué ciertas cosas son verdaderas, el usuario forma una hipótesis sobre alguna relación y la verifica por medio de una serie de queries en los datos,

Por ejemplo: un analista desearía determinar los factores que llevan a las omisiones de un préstamo. El analista puede crear una hipótesis que dice que la gente con pocos ingresos corren el riesgo de un mal crédito. Luego analiza la base de datos con OLAP para verificar esta suposición. Si la hipótesis no puede ser comprobada por los datos, el analista puede asumir que su hipótesis es correcta.

En otros términos el analizador de OLAP genera una serie de modelos hipotéticos que por medio de queries trata de verificar, el análisis de OLAP es esencialmente un proceso deductivo, pero qué sucede cuando la cantidad de variables a analizar crece, se vuelve un poco más difícil y se necesita de mucho tiempo para encontrar una hipótesis que pueda ser comprobada por el sistema.

La minería de datos es diferente de OLAP porque en lugar de hacer verificaciones por medio de modelos hipotéticos utiliza los datos para encontrar dichos modelos, este es esencialmente un proceso inductivo.

Por ejemplo, supongamos que el analista desea determinar los factores de riesgo por medio de una herramienta de minería de datos, la minería de datos

podría determinar que las personas con deudas e ingresos bajos son de mal crédito, pero adicionalmente también puede encontrar un patrón que el analista no tomó en cuenta y es que la edad es un factor determinante de riesgo.

Aquí es donde la minería de datos y OLAP se pueden complementar mutuamente, el analista necesita saber las implicaciones financieras que representa dar un crédito, la herramienta OLAP puede permitir al analista responder ese tipo de preguntas, además OLAP también es complementario en las fases iniciales del descubrimiento del conocimiento porque puede ayudar a explorar los datos enfocando atención en variables importantes, identificar excepciones o hallazgos.

1.4.3 Minería de datos y estadística

Las técnicas usadas en la minería de datos, cuando son exitosas, son igualmente exitosas de la misma forma en que técnicas estadísticas son exitosas. Y en su mayor parte las técnicas son usadas en los mismos lugares para los mismos tipos de problemas (predicción, descubrimiento de clasificación). En realidad algunas de las técnicas que son definidas como clásicas de la "minería de datos" son como CART y CHAID, estas tienen su origen en técnicas estadísticas.

Las técnicas mineras clásicas de datos tal como CART, redes neuronales y técnicas de vecino más cercano tienden a ser más robustas para ser usadas por usuarios menos expertos. Pero esa no es la única razón. La otra razón es que el espacio y el tiempo son limitados. Debido al uso de computadoras para almacenamiento y generación de datos, ahora existen grandes cantidades de información que están a disposición de los usuarios. Los dispositivos de almacenamiento han aumentado su capacidad de forma dramática por lo que la

capacidad de almacenar y procesar la información hace que las técnicas mineras sean más poderosas.

La línea fundamental sin embargo, de un punto de vista académico al menos, es que existe una diferencia muy pequeña entre las técnicas estadísticas y algunas técnicas clásicas de la minería de datos.

1.5 Minería de datos, máquinas de aprendizaje y estadística

La minería de datos toma ventaja de los avances en los campos de la inteligencia artificial (AI) y estadística. Ambas disciplinas han estado trabajando en problemas de reconocimiento de patrones y clasificación. Ambas comunidades han hecho contribuciones excelentes a la comprensión y aplicación de las redes neuronales y los árboles de decisión.

El desarrollo de la mayoría de las técnicas estadísticas estaba, hasta ahora basado en una teoría elegante de métodos analíticos eso ha funcionado bastante bien para analizar cantidades pequeñas de datos. La minería de datos no reemplaza técnicas estadísticas tradicionales, más bien, es una extensión de los métodos estadísticos, que es en parte el resultado de un cambio en la comunidad estadística.

El poder aumentado de computadoras y su bajo costo, complementado con la necesidad de analizar enormes cantidades de datos con millones de filas, ha permitido el desarrollo de nuevas técnicas basadas en una exploración de fuerza bruta para obtener posibles soluciones.

Nuevas técnicas incluyen algoritmos relativamente recientes como redes neuronales y árboles de decisión y nuevos acercamientos a algoritmos más viejos

tal como análisis discriminante. En tal virtud aprovechar la potencia de los ordenadores en los enormes volúmenes de datos disponibles, estas técnicas pueden aproximar casi cualquier forma o interacción funcional en ellas mismas.

Las técnicas estadísticas tradicionales dependen en el modelo para especificar la forma e interacciones funcionales.

El punto clave es que la minería de datos es la aplicación de estas técnicas estadísticas y de otras de inteligencia artificial para resolver los problemas más comunes de los negocios hasta cierto punto esas técnicas se acercan a la habilidad del conocimiento de un trabajador tan bien como un profesional de estadísticas adiestrado.

La minería de datos es una herramienta para crecer la productividad de personas tratando de construir modelos de predicción.

1.6 Aplicaciones de la minería de datos

La minería de datos es cada vez más popular debido a la contribución substancial que puede hacer. Puede ser usada para controlar costos así como también para contribuir a incrementar las entradas.

Muchas organizaciones están usando minería de datos para ayudar a manejar todas las fases del ciclo vital del cliente, incluyendo la adquisición de nuevos clientes, aumentando los ingresos con clientes existentes y manteniendo bien a la clientela.

Determinando características de clientes buenos (trazado de perfil), una compañía puede determinar conjuntos con características similares. Perfilando

clientes que hayan comprado un producto en particular, ello puede enfocar la atención en clientes similares que no hayan comprado ese producto (de venta cruzada).

Cerca de perfilar clientes que se han ido, lo que la compañía hace para retener los clientes que están en riesgo de alejarse, porque es normalmente un poco menos caro retener un cliente que conseguir uno nuevo.

Las ofertas de minería de datos se valoran a través de una amplia efectividad de industrias. Las compañías de telecomunicaciones y tarjeta de crédito son dos de los conductores para aplicar minería de datos, para detectar el uso fraudulento de sus servicios.

Compañías aseguradoras y bolsas de valores se interesan también al aplicar esta tecnología para reducir el fraude. Las aplicaciones médicas son otra área fructífera; la minería de datos se puede utilizar para predecir la eficiencia de procedimientos quirúrgicos, las pruebas médicas o medicaciones.

Las compañías activas en el mercado financiero, usan minería de datos para determinar el mercado y características de industria así como para predecir el comportamiento de las compañías individuales y mejorar el sistema de inventarios.

Los minoristas están haciendo mayor uso de la minería de datos, para decidir que productos en particular deben mantener en inventario para no abastecerse de productos innecesarios, así como para evaluar la eficacia de promociones y ofertas.

Las firmas farmacéuticas poseen grandes bases de datos de los compuestos químicos y de material genético en las cuales hay substancias que pueden muy buenas ser candidatas para minar, esto con el objetivo de determinar como se

pueden desarrollar nuevos agentes para los tratamientos de determinadas enfermedades.

1.7 Retos de la minería de datos

Nadie duda que la experiencia es el elemento fundamental del conocimiento y la sabiduría. La asimilación de hechos pasados permite enfrentar al futuro con más posibilidades de éxito, sin tener que recordar todos los detalles del pasado. Esto es claro en las personas, pero, ¿cómo puede aplicarse a las corporaciones?

¿Qué promete la minería de datos?

La tecnología informática es infraestructura fundamental de las grandes organizaciones y permite, hoy en día, registrar con lujo de detalle, los elementos de todas las actividades con asombrosa facilidad. La tecnología de bases de datos permite almacenar cada transacción y muchos otros elementos que reflejan la interacción de la organización con todos sus interlocutores, ya sean otras organizaciones, sus clientes, o internamente, sus divisiones, sus empleados, etcétera. Tenemos pues, un registro bastante completo del comportamiento de la organización. Pero, ¿cómo traducir ese voluminoso conjunto de datos en experiencia, conocimiento y sabiduría corporativa que apoye efectivamente la toma de decisiones, especialmente al nivel gerencial que dirige el destino de las grandes organizaciones? ¿Cómo comprender el fenómeno, tomando en cuenta grandes volúmenes de datos?

Ilustremos algunas ideas: ¿Cuántas transacciones realiza un banco? Hoy en día, un banco cuenta con, posiblemente, cientos de agencias. Mas aún, con otros miles de medios en los que se realizan transacciones, que van desde autorizaciones de compra de tarjeta de crédito, hasta manejo de inversiones e

instrumentos vía mercados electrónicos. A diferencia de los primeros bancos de la historia, actualmente los tesoreros no pueden supervisar en persona cada transacción y registrar en su memoria o en un pequeño libreto, el estado de cuentas de su negocio. Tienen que confiar en sistemas de información basados en computadoras. Estos sistemas, no sólo permiten la operación del negocio, sino también producen resúmenes, reportes, estadísticas e información generalizada, lo que posibilita imaginar el estado de cuentas de la institución. Otra serie de reportes, tal vez indican o sugieren estrategias a futuro, o bien condiciones del nicho de mercado en que se encuentra el negocio y proveen elementos para la planeación.

En esta masividad de datos, ¿qué sabe, por ejemplo, la institución de su mercado?, es decir, ¿qué conoce sobre cada uno de sus clientes?. Tal vez todo y tal vez nada. Nada en el sentido de que, cuando un cliente de varios años se presenta en una agencia diferente a la que acostumbra visitar, lo único que puede obtener es información escueta, directamente relacionada con su estado de cuenta, pero extremadamente impersonal.

El empleado bancario que representa a la institución no tiene ningún elemento para darle un trato personal a los clientes. La vecina, podría preguntarle, cómo le va al hijo en la escuela; pero tal vez, los empleados del banco saben si el cliente tiene un hijo en la escuela, qué escuela es, qué tan a menudo hace pagos de colegiatura, qué tan a menudo hace compras de artículos escolares con tarjeta de crédito, qué tan a menudo usa el cajero automático que está frente a la escuela, qué seguro médico u hospital utiliza, etcétera, simplemente del historial de transacciones del fiel cliente. ¿Acaso no podrían inferirse elementos que permitieran dar un trato más personal a cierto tipo de clientes? Por ejemplo, aquellos que tienen hijos pequeños, seguramente esperan servicios muy distintos que aquellos que son pensionistas.

La Minería de Datos ha surgido del potencial del análisis de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoye la toma de decisiones y que pueda construir una experiencia a partir de los millones de transacciones detalladas que registra una corporación en sus sistemas informáticos.

La Minería de Datos parece ser más efectiva cuando los datos tienen elementos que pueden permitir una interpretación y explicación en concordancia con la experiencia humana. Lo anterior se facilita mucho si estos elementos son el espacio y el tiempo. Afortunadamente, se estima que el 80% de los datos registrados en una base de datos tiene la posibilidad de geo-referenciarse y, el 100%, de puntualizarse temporalmente. ¿Qué quiere decir esto? En primer lugar, que en la mayoría de los casos es posible asociar un punto en el espacio, un domicilio, unas coordenadas geográficas con la entidad que representa el dato, una fecha o punto en el tiempo. En segundo lugar, que los patrones o inferencias sobre los datos son usualmente interesantes, en la medida en que son patrones en el tiempo o en el espacio. Por ejemplo, qué productos se comercializan mejor en la temporada navideña, en qué regiones es productivo sembrar café, qué áreas de una zona urbana incrementarán su demanda de escuelas primarias.

La tecnología promete analizar con facilidad grandes volúmenes de datos y reconocer patrones en tiempo y espacio que soportarán la toma de decisiones y construirán un conocimiento corporativo de alto nivel.

1.7.1 Primer reto

La tecnología de Minería de Datos parece robusta y lista para su aplicación, dado el gran crecimiento de empresas que comercializan software con diferentes técnicas. Más aún, gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y "data warehousing", con técnicas de

aprendizaje automático y de estadística. Sin embargo, la tecnología enfrenta aún varios retos.

El primero de estos retos, es la facilidad con que se puede caer en una falsa interpretación; para explicarlo, basta reconocer que las primeras y más maduras técnicas para el análisis de datos, con el fin de modelar un fenómeno, provienen de la estadística. Todos saben que existe la posibilidad de ser engañados por la estadística; no todos tienen un sólido entendimiento de la matemática, los supuestos y el modelado para entender a la perfección el riesgo o margen de error en un ejercicio de inferencia estadística, pero todos operan y funcionan con resúmenes e indicadores estadísticos generalmente muy simples. Cuando se dice que una gran decisión se basó en la información disponible, típicamente es una serie de promedios y estimadores estadísticos que presentan una generalización de un gran volumen de datos, donde se hace una inferencia.

Nótese la facilidad y el poder que proporciona la estadística. Volviendo al ejemplo del tesorero bancario, la decisión de cerrar agencias o reducir empleados no se basa en una revisión de los miles de datos posibles, pero sí en indicadores (estadísticos), como el valor de la transacción promedio, el salario promedio, etcétera.

La estadística es una herramienta poderosa, y es elemento crucial en el análisis de datos. Sin embargo, a veces se enfrentan problemas muy serios en la interpretación de sus resultados. El ejemplo típico es que, usualmente, no se recuerdan que estos resultados se aplican a grupos (poblaciones) y no a individuos. Estos peligros se ven amplificados en el uso de software de Minería de Datos. Dichas herramientas informáticas pueden poner a disposición de un "analista" (o minero de datos), la posibilidad de crear fácilmente indicadores, resúmenes, gráficas, y aparentes tendencias, sin un verdadero entendimiento de lo que se está reflejando. Es decir, resulta más fácil hacer creíble una falsedad,

posiblemente porque la produjo una computadora, con muchas gráficas y con base en muchos datos, eso sí, en un instante.

Así que el reto es doble. ¿Cómo hacer las herramientas de minería de datos accesibles a cualquiera, hasta aquel que no sabe lo más mínimo de estadística, pero que sus resultados e interpretaciones sean válidos? Nótese que es importante que la herramienta tenga un gran elemento de accesibilidad para que su producción sea rentable. Un ejemplo de esto son las bases de datos relacionales, pues su diseño, modelado, y las herramientas alrededor de los manejadores, han hecho posible que no se requiera de una gran especialización para tener una gran cantidad de usuarios y que, por lo tanto, el mercado sea extenso para mantener a los que producen manejadores de datos.

Naturalmente, es importante que las inferencias sean válidas. Esto trae un segundo punto crítico, o segundo reto**.** ¿Si con la estadística se enfrenta el problema de que es relativamente fácil equivocarse, existe la posibilidad de equivocarse con la Minería de Datos?

¿Por qué es más fácil equivocarse con la minería de datos?

La primera razón es porque, aun con la estadística, hallar una correlación (estadísticamente significativa) no significa haber encontrado una relación causa-efecto. El contraejemplo clásico lo constituyen los datos anuales de edad, de las personas fallecidas en los Estados Unidos, por Estados. Los análisis estadísticos más abundantes encuentran que el Estado de Florida tiene, año tras año, la edad promedio más avanzada en que la gente fallece, y con todo el rigor (significancia estadística) que se desee. ¿Es acaso esto un indicador de que nacer en Florida garantiza longevidad? ¿Se vive más si se muda uno a la península? De ninguna manera; la verdadera explicación es que Florida alberga a una gran cantidad de pensionistas, retirados, etcétera. La gente se va a morir a Florida, pero para

entonces, ya es muy mayor. Si se muere antes de ser pensionista, se muere en su lugar de origen, forzando el promedio de su estado a bajar; si vive mucho, le alcanza para mudarse a Florida y subir allí el promedio.

El software de Minería de Datos esta diseñado para hallar correlaciones, para olfatearlas. Su tarea consiste en encontrar aquella proyección de los datos, aquella perspectiva donde aparece una correlación y, lamentablemente, en muchos casos, presentarla como una relación causa-efecto. Esto es especialmente cierto en los sistemas que generan reglas de asociación, de tal forma, "SI ESTADO = FLORIDA, ENTONCES, EDAD - AL - FALLECIMIENTO = ANCIANA".

Esto se deriva de que la Minería de Datos sigue una filosofía muy diferente a como se hace la ciencia. La ciencia, generadora del conocimiento y fundamento sorprendente de la tecnología, opera con base en el método científico. Este método postula que la hipótesis se genera con antelación a la colección de los datos. La Minería de Datos genera hipótesis a partir de los datos. No es catastrófico que se generen hipótesis a partir de los datos. En realidad, el formular creencias a partir de una experiencia finita y limitada es un elemento fundamental del aprendizaje, pero el otro elemento crucial consiste en la revisión de las hipótesis a la luz de nuevos datos y nuevas experiencias.

La Minería de Datos es una herramienta explorativa y no explicativa. Es decir, explora los datos para sugerir hipótesis. Es incorrecto aceptar dichas hipótesis como explicaciones o relaciones causa-efecto. Es necesario coleccionar nuevos datos y validar las hipótesis generadas ante los nuevos datos y después descartar aquellas que no son confirmadas por los nuevos datos.

Pero la Minería de Datos no puede ser experimental. En muchas circunstancias, no es posible reproducir las condiciones que generaron los datos (especialmente si son datos del pasado, y una variable es el tiempo).

Afortunadamente, existen algunas técnicas para resolverlo, pero se requiere cierta madurez estadística para su comprensión.

1.7.2 Tiempo y espacio

El modelar tiempo y espacio en computadoras son problemas complejos, especialmente para hacer inferencias. Esto hace que las técnicas de aprendizaje automático enfrenten mayores dificultades cuando abordan los temas que parecen más interesantes, de descubrimiento de patrones.

A esto se añaden varios tipos de problemas. El primero, es el de Minería de Datos con relaciones en el tiempo. Es muy posible que se deseen hacer inferencias y análisis de datos sobre un periodo determinado, pero que durante dicho periodo no se haya registrado el mismo número de variables, o que éstas no tengan la misma precisión, o carezcan de la misma interpretación. En ciertos casos puede que se haya hecho un ejercicio de Minería de Datos en el pasado y que los datos se hayan descartado o destruido, pero que se desee hacer una comparación con datos más recientes. Nótese que un ejercicio de Minería de Datos puede traer a la luz relevancia de variables y factores, pero que sea imposible recopilar estas variables y completar adecuadamente conjuntos de datos del pasado. Otros problemas de análisis de datos con relación al tiempo, son asociados a la discontinuidad de los datos con respecto al tiempo. En este sentido, no se conocen todos los datos en el continuo del tiempo. Por ejemplo, si se hacen recopilaciones mensuales, es imposible hacer una predicción semanal.

Desde el punto de vista geográfico o espacial, resulta complejo identificar las esferas de influencia y las distribuciones en espacio que reflejan la realidad. Esto es una observación que a veces parece paradoja --todo está relacionado con todo-- pero es mayormente influenciado por lo que tiene más próximo. Se puede identificar una relación explícita de cómo decrece una esfera de influencia en el espacio, y existen ideas que, incluso, proponen que los fenómenos espaciales no son modelables (como la teoría del caos).

1.7.3 Privacidad

Finalmente, quisiera tocar el punto de lo que es privacidad. Cuando la Minería de Datos era aún emergente, se llegó a pensar que no presentaba ningún peligro o riesgo para la privacidad de los clientes. Hoy en día, se piensa todo lo contrario, sin embargo, no existe un marco jurídico que haya mantenido el paso con el avance tecnológico. Esto es, hoy en día, las corporaciones comercializan con millones de perfiles personales, sin que aquellos a que se refieren los datos intercambiados, estén en posibilidad de intervenir. Cada llamada telefónica, cada transacción bancaria, cada compra en un supermercado, es registrada en una computadora, y si la compañía de teléfonos, el banco y el club de supermercados combinan sus bases de datos, están en condiciones de elaborar un perfil muy completo. Este perfil definiría a más de una persona (y no como los que están en condiciones de conocer a sus vecinos). Si a esto añadimos qué sitios de WEB visita, qué y dónde se compró con la tarjeta de crédito, etcétera, no existe ninguna privacidad. El problema va desde las definiciones de qué constituye privacidad y quién es el propietario de los datos, hasta qué tanto de un individuo, está en posibilidad real de compilarse.

1.8 DMQL

Las herramientas emergentes de minería de datos y los sistemas llevan naturalmente a la demanda de un lenguaje de minería de datos de consulta poderoso, encima de que muchas interfaces de usuarios interactivas y gráficas pueden desarrollarse. Esto motiva a diseñar un lenguaje de consulta de minería de datos, DMQL, para diferentes tipos de conocimiento en bases de datos de relacionales.

En este sentido, el éxito de los sistemas deben acreditarse en parte a la normalización de lenguajes de consulta. Las actividades de normalización recientes en sistemas de base de datos, tal como el trabajo relacionado con el SQL-3, el OMG y el ODMG 3, muestra de nuevo la importancia de una base de datos estandar.

Esto motiva a examinar lo que deben ser las primitivas de un lenguaje de minería de datos. Las actividades de R y D mineras actuales de datos muestran que la minería de datos cubre un espectro ancho de tareas, de reducción de datos a minería, reglamentos de asociación, clasificación de datos. Esto hace el diseño de un lenguaje de minería de datos comprensivo.

Corrientemente, existen muchas interfaces de usuarios gráficas para diversas tareas de la minería de datos. Sin embargo, es importante comprender los mecanismos fundamentales de los métodos de minería de datos.

2. EL PROCESO DE MINERIA DE DATOS

2.1 Cómo trabaja la minería de datos

¿Cómo es exactamente que la minería de datos es capaz, de extraer información importante que no se sabe que va a suceder después? La técnica que se utiliza para ejecutar estos hechos en la minería de datos se llama modelado de datos. El modelado datos es simplemente el acto de construir un modelo en una situación donde se sabe la respuesta y luego ésta se aplica a otra situación en la que no se sabe la respuesta. Por ejemplo, si se está buscando un tesoro de un galeón español en el mar lo primero que se debe hacer es investigar en el pasado cuando otros barcos encontraron tesoros. Se puede notar que estos buques a menudo tienden a encontrarse de las costas de Islas Bermudas y existen ciertas características de las corrientes de océano y ciertas rutas que probablemente fueron tomadas por los capitanes de los barcos en esa época. Si se toman estas similitudes y se construye un modelo que incluye las características que son comunes a las localizaciones de estos tesoros sumergidos, existe una gran posibilidad de que con estos modelos se pueda navegar en busca de otro tesoro, el modelo construido indica el lugar más probable donde pueda darse una situación similar al pasado. Si se ha construido un modelo adecuado, la probabilidad de encontrar un tesoro es bastante grande.

El proceso de construcción de un modelo es algo que las personas han estado haciendo durante mucho tiempo, desde antes del advenimiento de las computadoras. Lo que pasa con las computadoras, no es muy diferente de la forma en que las personas construían sus modelos. Las computadoras son cargadas con una gran cantidad de información sobre una variedad de situaciones donde se cuenta con una respuesta conocida y entonces el software de minería de

datos en la computadora debe analizar a través de esos datos y determinar las características que se deben examinar por medio del modelo.

Una vez que haya sido construido el modelo entonces éste puede ser usado en situaciones similares donde no se cuenta con una respuesta. Por ejemplo, si se supone que el director de comercialización para una compañía de telecomunicaciones desea adquirir nuevos clientes que utilizan el teléfono para llamadas de larga distancia. De forma aleatoria se pueden enviar por correo cupones a la población general pero en ningún caso se logrará los resultados que se desean, ya que no toda la población realiza llamadas de larga distancia pero se puede utilizar la experiencia almacenada en la base de datos para construir un modelo.

Como el director de comercialización tiene acceso a una gran cantidad de información sobre todos sus clientes: edad, sexo, historia crediticia y uso de llamadas a larga distancia, se debería concentrar en aquellos usuarios que tienen un uso continuo de llamadas de larga distancia. Esto se puede realizar construyendo un modelo.

La meta al explorar los datos en busca de respuestas es hacer cierto cálculo sobre la información de los datos conocidos, el modelo que se construye va de la información general del cliente hacia información particular. Por ejemplo, un modelo simple para una compañía de telecomunicaciones podría ser:

El 98% de mis clientes que tiene un ingreso anual de más de Q.60,000.00 y gasta más de Q.80.00/mes en llamadas de larga distancia.

Este modelo se podría aplicar entonces a los datos para tratar de decir algo sobre la información con que se cuenta en la compañía de telecomunicaciones a la que normalmente no se tiene acceso. Con este modelo nuevos clientes pueden ser selectivamente fichados.

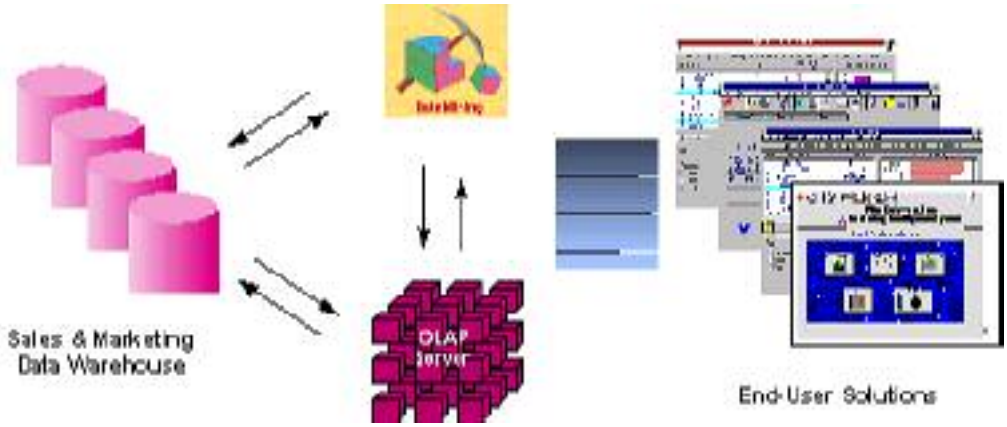
El data warehouse es una excelente fuente de datos para este tipo de modelado. Minar los resultados de un data warehouse de prueba que representa una muestra grande pero relativamente pequeña de elementos que puede proporcionar fundamentos para identificar nuevos patrones en el data warehouse completo.

2.2 Una arquitectura para la minería de datos

Para obtener el mejor resultado de estas técnicas avanzadas, deben estar enteramente integradas con un data warehouse así como herramientas de análisis de negocio interactivas flexibles. Muchas herramientas de minería de datos operan corrientemente fuera un data warehouse, requiriendo de un paso para extraer, importar, y analizar los datos. Además, cuando nuevas perspicacias exigen ejecución operacional, la integración con el data warehouse simplifica la aplicación de la minería de datos.

El resultado del análisis de un data warehouse junto con la minería de datos puede aplicarse para mejorar los procesos de un negocio en una organización, en áreas como administración de campañas promocionales, detección de fraudes, rotamiento de un producto, etc. En la. Figura 1 se ilustra una arquitectura para la minería de datos.

- Figura 1 arquitectura de minería de datos integrada **FUENTE:** Kurt Thearling, Ph.D. Director, Advanced Data Mining http://thearling.com



El punto de partida ideal es un data warehouse que contiene una combinación de datos internos siguiendo los registros de los contacto de los clientes acoplados a los datos externos sobre la actividad de los competidores. La información sobre clientes potenciales proporciona también una base excelente para explorar terreno que puede ser minado. Este data warehouse puede ponerse en práctica en una variedad de los sistemas de bases de datos relacionales: Sybase, Oracle, o cualquiera que sea flexible y de rápido acceso de datos.

El servidor OLAP (proceso analítico en línea) habilita un modelo de negocio un poco más sofisticado para ser aplicado al navegar por el data warehouse. Las estructuras multidimensionales permiten al usuario analizar los datos de la forma que ellos desean mirar su negocio ya sea por línea de productos, región, y otras perspectivas claves de su negocio. El servidor de minería de datos debe ser integrado con el data warehouse y con el servidor de OLAP para obtener el análisis de negocio enfocado directamente en esta infraestructura. Una plantilla de metadata avanzada, céntrica del proceso define los objetivos de la minería de datos para asuntos de negocio específicos como dirección de campaña, prospección, y optimización de promociones. La integración con el data

warehouse habilita decisiones operacionales para ponerse en práctica directamente. Conforme vaya creciendo el data warehouse se obtienen nuevas decisiones y resultados, la organización puede continuamente minar las mejores prácticas y aplicarlas a decisiones futuras.

Este diseño representa un cambio fundamental de los sistemas de soporte de decisión convencionales. Antes que dar simplemente datos al usuario final a través de queries y reportes, el servidor de análisis avanzado aplica modelos de casos del negocio del usuario directamente al data warehouse y retorna un análisis proactivo de la información más pertinente. Estos resultados mejoran la metadata en el servidor de OLAP proporcionando una capa de datos dinámicos que represente una vista filtrada de los datos. Reporteadores, herramientas de visualización, y otras herramientas de análisis se pueden aplicar entonces para planear acciones futuras y confirmar el impacto de esos proyectos.

2.3 Minería de datos predictiva

La meta de la minería de datos es producir nuevo conocimiento que el usuario pueda utilizar. Esto se realiza construyendo un modelo del mundo real basado en los datos de una variedad de fuentes que puede incluir transacciones corporativas, historias de cliente e información demográfica, datos de control de proceso, y bases de datos externas pertinentes tal como información de agencia que suministra información acerca del crédito de presuntos clientes o datos del tiempo.

El resultado del modelo construido es una descripción de patrones y relaciones en los datos que se puede usar confiadamente para predicción.

Para evitar confundir los aspectos diferentes de la minería de datos, es necesario imaginar una jerarquía de las elecciones y decisiones que necesita hacer antes empezar:

- Meta de negocio
- El tipo de la predicción
- Tipo ejemplar
- Algoritmo
- Producto

El nivel más alto, es la meta de negocio: ¿cuál es el propósito de la minería de datos? Por ejemplo, buscando los modelos en sus datos para ayudarle a retener los clientes buenos, podría construir un modelo para predecir la rentabilidad de un cliente y un segundo modelo para identificar los clientes que probablemente pueden alejarse del negocio.

El conocimiento de las necesidades y objetivos de la organización facilitará la creación de metas en los modelos.

El próximo paso está decidiendo el tipo de la predicción que es más apropiado.

La clasificación, consiste en predecir en que categoría se clasifica cada caso.

La regresión consiste en predecir el valor numérico que tendrá la variable (si es una variable que varia en el tiempo, ello se llama la predicción de serie cronológica).

En el ejemplo anterior, se podría usar regresión para predecir la cantidad de rentabilidad, y clasificación para predecir que clientes poseen una tendencia para alejarse del negocio. Ahora puede escoger el tipo ejemplar: una red neuronal para

ejecutar la regresión, tal vez, y árbol de decisión para la clasificación. Existen también modelos estadísticos tradicionales para escoger, como logística regresión, análisis discriminante, o los modelos lineales generales.

Muchos algoritmos son disponibles para construir sus modelos. Se puede construir la red neuronal usando backpropagation o funciones de base radiales. Para el árbol de decisión, se podría escoger entre CART, C5.0, rastree, o CHAID.

Cuando se selecciona un producto de minería de datos, se debe tener presente que cada uno ha sido implementado con un algoritmo particular aún cuando lo identifican con el mismo nombre. En Estas implementaciones las diferencias pueden afectar características operacionales tal como uso de memoria y almacenamiento de datos, así como características de ejecución tal como velocidad y exactitud.

Muchas metas de negocio deben ser construidas por modelos múltiples utilizando una variedad de algoritmos. Posiblemente no pueda ser capaz de determinar que tipo modelo es mejor hasta que ha probado varios acercamientos.

2.3.1 Clasificación

La clasificación de los datos sirve para identificar las características que indican el grupo al que cada caso pertenezca. Este modelo puede ser usado para comprender los datos existentes y para predecir cómo se comportará algún nuevo caso. Por ejemplo, si se quisiera poder predecir si los individuos pueden ser clasificados como posibles candidatos para responder a una solicitud de correo directo, o si es vulnerable al cambio de un servicio de larga distancia de teléfono, o si es un buen candidato para un procedimiento quirúrgico.

La minería de datos crea los modelos de clasificación examinando datos que a su vez se encuentran ya clasificados como casos, e inductivamente encuentra un modelo que predice. Estos casos ya clasificados pueden venir de una base de datos histórica, tales como personas que en algún momento ya se han sometido a un tratamiento médico particular o personas que se cambian de algún tipo de servicio a otro. Estos datos se pueden extraer de una muestra experimental de la base de datos. Por ejemplo, como muestra puede ser enviada una lista con una oferta, y los resultados serán enviados para desarrollar un modelo de clasificación para ser aplicado a toda la base de datos.

En algunas ocasiones un experto clasifica una muestra de la base de datos, y luego esta clasificación se utiliza para crear el modelo que se aplicará a toda la base de datos.

2.3.2 Regresión

La regresión utiliza valores existentes para pronosticar qué valores son los que se obtendrán más adelante. En un caso simple de regresión se utilizan técnicas estadísticas como la regresión lineal, desafortunadamente muchos problemas de la vida real no son simples proyecciones lineales de los valores previos. Por ejemplo los rangos de fallo en volúmenes de ventas de un determinado stock de productos son bastante difíciles de predecir porque dependen de la interacción de múltiples variables de predicción.

2.4 El proceso de minería de datos

El único propósito del proceso de minería de datos es extraer conocimiento de grandes bases de datos, esto se logra utilizando algoritmos. Los pasos del proceso de minería de datos son los siguientes:

- Selección de datos
- Depuración de datos
- Enriquecimiento de datos
- Transformación de datos
- Preparación de un conjunto de casos
- Construcción del modelo de minería de datos

2.4.1 Selección de Datos

La selección de datos para el proceso de minería de datos se divide en dos partes, la primera parte consiste en la localización de los datos, esto tiende a ser un poco más mecánico en comparación con la segunda parte, que consiste en la identificación de los datos, esta segunda parte requiere de la ayuda de un experto en los datos que se desean analizar (una persona que se familiarice con los propósitos del negocio y los datos que se van a examinar).

2.4.2 Localización de los datos:

La minería de datos puede ser implementada en casi cualquier base de datos, pero muchas bases de datos generalmente no traen soporte para ambientes de negocios.

Las bases de datos que se recomiendan para la minería de datos son:

Data Warehouse: Por varias razones el data warehouse es ideal para la minería de datos, los procesos que ya posee el data warehouse seleccionan, limpian, enriquecen y transforman los datos, estos procesos son muy parecidos a

los utilizados por la minería de datos. El data warehouse ha sido diseñado para hacer queries que manejan un alto volumen de información que es representada en un formato dimensional, lo que facilita la identificación de escenarios específicos.

Data Mart: El data mart es un subconjunto del data warehouse encapsulado para propósitos específicos del negocio. Por ejemplo un data mart de ventas y comercialización contendría una copia de las tablas que se encuentran en el data warehouse pero las tablas en el data mart solo contienen la información necesaria para satisfacer la investigación de ventas y comercialización.

Como los data mart se modelan según las necesidades de los usuarios de las empresas, la mayor parte de los data mart no son adecuados para la minería de datos. Sin embargo se puede construir un data mart diseñado específicamente para la minería de datos. Las bases de datos OLAP frecuentemente son modeladas como data mart, ya que su uso y funcionalidad son similares a otros data mart

Bases de datos OLTP: Las bases de datos OLTP, también conocidas como de bases de datos operacionales, no han sido optimizadas para el tipo de recuperación que la se requiere en la minería de datos; Los impactos de ejecución como el acceso y velocidad de transacción se pueden dar en otras aplicaciones que depende de la optimización de actualización de alto volumen de tales bases de datos. La falta del pre-agregado puede impactar también el tiempo necesario para el tratamiento de los modelos de minería de datos basados en bases de datos OLTP, debido a muchas uniones y cantidad de registros que se recuperan en la ejecución de quieries en las bases de datos OLTP.

Bases de datos operacionales (ODS) : Las bases de datos operacionales han crecido en popularidad ya que se usan para procesar y consolidar volúmenes

grandes de datos típicamente manejados por bases de datos OLTP. El concepto que se tiene de una base de datos operacionales es fluida, pero las bases de datos operacionales son típicamente usadas como un buffer de datos entre los datos de OLTP y aplicaciones que requieren acceso a tales datos, pero es necesario que esta información esté aislada de la base de datos de OLTP por razones de ejecución de los queries.

Mientras la minería de datos, en bases de datos operacionales (ODS) pueden ser útiles, las bases de datos operacionales son conocidas por los cambios que se realizan rápidamente. El modelo de minería de datos entonces se convierte en un lente en el cual se pueden obtener resultados rápidamente, y el usuario nunca se da cuenta de que el modelo de minería de datos refleja exactamente una vista histórica de los datos.

La minería de datos es una búsqueda de datos que se basa en la experiencia, no una búsqueda que pretenda obtener inteligencia de los datos. Porque al analizar esta experiencia se obtiene una vista ancha de datos históricos, las bases de datos transaccionales más volátiles deben evitarse.

Al localizar los datos para la minería de datos, idealmente se debería documentar toda la información, para tener acceso fácilmente; muchos de los pasos en el proceso de la minería de datos suponen el acceso libre y directo a los datos. Los niveles de seguridad, las comunicaciones entre módulos, las limitaciones físicas de la red, y otros aspectos pueden restringir el acceso libre a datos históricos. Tal acceso libre se considera esencial del proceso de diseño para la implementación de una solución de minería de datos.

2.4.3 Identificación de datos

Este paso es uno de los más importantes de todos los pasos en el proceso de la minería de datos. La calidad de los datos escogidos determina finalmente la calidad de los modelos de la minería de datos. El proceso de identificar datos para su uso en la minería de datos va en paralelo con el proceso de selección de datos utilizado en los data warehouse.

Al identificar los datos que serán de utilidad, se debe cuestionar las tres preguntas siguientes:

¿Estos datos cumplen con los requerimientos según el escenario propuesto? Los datos no solo deben coincidir con el propósito del escenario, sino también su nivel de detalle. Por ejemplo, si se desea modelar la información sobre el desempeño de un producto, es necesario representar cada producto de forma individual, ya que cada producto se convierte en un caso o en un conjunto de casos.

¿Están completos los datos? Los datos deben tener todos los atributos necesarios para describir exactamente un escenario. Hay que recordar que una falta de los datos es información desconocida; la falta de información sobre un producto en particular puede indicar una tendencia de desempeño positiva en una familia de productos; el producto puede ejecutar tan bien que ningún cliente ha relatado ningún asunto de ejecución con el producto.

¿Los datos contienen los atributos necesarios para obtener el resultado deseado? Al realizar un modelo predictivo, los datos utilizados para construir el modelo de minería de datos debe contener atributos que proporcionen el resultado deseado. A veces, para satisfacer este requerimiento, un atributo temporal es construido para proporcionar un valor de resultado discreto para cada

caso; esto puede ser hecho en los pasos de enriquecimiento de datos y transformación de datos.

Los datos que pueden satisfacer inmediatamente estas preguntas,se son los mejores datos para ponerse en marcha la minería de datos, sin embargo no se está limitado a tales datos. Los pasos de enriquecimiento de datos y transformación de datos permiten modelar los datos hacia un formato más útil para la minería de datos, y en algunos casos los datos considerados marginales se convierten en datos útiles por esta manipulación.

2.4.4 Depuración de datos

La limpieza o depuración de datos es el proceso en el que se aseguran los propósitos de la minería de datos, los datos son uniformes desde el punto de vista del uso de la llave y los atributos. Identificar y corregir información perdida, limpiar los registros, son aspectos de la depuración de datos.

La depuración de datos se separa del enriquecimiento y transformación de datos porque los intentos de depuración son para corregir el mal empleo de datos o atributos incorrectos en los datos existentes. El enriquecimiento de datos, por el contrario, añade nuevos atributos, mientras que la transformación de datos cambia la forma o estructura de atributos de los datos existentes para satisfacer los requerimientos de la minería de datos.

Típicamente, la mayor parte de la minería de datos ya ha sido procesada por los procesos de data warehouse. Sin embargo, ciertas orientaciones generales para la depuración de datos son útiles para situaciones en que un data warehouse bien diseñado no es disponible, y para aplicaciones en que las necesidades de negocio requieren la limpieza de tales datos.

Cuando se depura la información en un data warehouse, el mejor lugar para empezar está en la casa; es decir, empezar a limpiar los datos en la de base de datos OLTP, antes que sean importados los datos malos hacia el data warehouse. Esta regla también se aplica a minería de datos, especialmente si se piensa construir un data mart para los propósitos de minería de datos. Siempre hay que tratar de limpiar datos a la fuente, antes de tratar de modelar datos que proporcionarán un resultado poco satisfactorio. La parte del "close loop" en el proceso de apoyo a la toma de decisiones debe incluir los mejoramientos de calidad de datos, tales como orientaciones de entrada de datos y optimización de reglas de validación para datos OLTP, de manera que la depuración de datos proporcione la información necesaria para tales mejoramientos.

Idealmente, un área de almacenamiento temporal puede estar acostumbrada a manejar la depuración de datos, enriquecimiento de datos, y pasos de transformación de datos. Esto permite la flexibilidad para no solo cambiar los propios datos, sino también los datos de meta que formulan los datos. Enriquecimiento de datos y transformación en particular, especialmente para la construcción de nuevas llaves y relaciones o conversión de tipos de datos, pueda beneficiarse de este acercamiento.

Limpiar datos para los propósitos de la minería de datos normalmente requiere los pasos siguientes:

2.4.4.1 Verificación de consistencia de las llaves

Verificar que los datos tecleados son consistentes a través de todos los datos pertinentes. Ello puede ser más probable se esta acostumbrado a identificar casos o atributos importantes.

2.4.4.2 Verificación de las relaciones

Verificar que las relaciones entre casos se conforman con reglamentos de negocio definidos. Las relaciones que no soportan reglamentos de negocio definidos pueden sesgar los resultados de un modelo de minería de datos, extraviando el modelo en construir modelos y reglamentos que no puede aplicarse a un escenario definido.

2.4.4.3 Uso de atributos y verificación de alcance

Generalmente, la calidad y exactitud de un atributo está en proporción directa a la importancia de los datos. La información, para un negocio industrial que crea partes y productos para la industria aeroespacial, es crucial a la operación exitosa del negocio, y generalmente es más exacto y de calidad más alta que la información de contacto de los vendedores que suministra el inventario.

Verificar que los atributos utilizados se están usando tal y cual es la información que se tiene en la base de datos, y que el alcance o campo de los atributos seleccionados tienen el significado al escenario para ser modelado.

2.4.4.4 Análisis de datos

Verificar que los valores almacenados en los atributos son adecuados según el escenario a evaluar. Los datos de los atributos menos críticos típicamente exigen más limpiadores que los atributos vitales para la operación exitosa del negocio.

Siempre se debe ser precavido al momento de excluir o sustituir valores en los atributos ya que puede perder la información original. Los datos perdidos no

siempre son calificados como información perdida. La falta de datos para un racimo específico en un escenario puede revelar mucha información al realizar las preguntas correctas. Por lo tanto, se debe ser precavido al excluir atributos o elementos de datos de un conjunto de casos.

Los esfuerzos en la depuración de los datos directamente contribuyen al éxito o fracaso del proceso de minería de datos. Este paso nunca se debe pasar por alto, sin importar los costos o tiempo. Aunque los servicios de análisis trabajan bien con cualquier clase de datos, los resultados son mejores cuando los datos son consistentes y uniformes.

2.4.5 Enriquecimiento de datos

El enriquecimiento de datos es el proceso de añadir nuevos atributos, tales como campos calculados o datos de fuentes externas, a los datos ya existentes.

La mayor parte de las referencias en la minería de datos tienden a combinar este paso con transformación de datos. La transformación de datos supone la manipulación de datos, pero el enriquecimiento de datos supone añadir información a los datos existentes. Esto puede incluir la combinación de los datos internos con los datos externos, que se obtiene de los distintos departamentos, compañías.

El enriquecimiento de datos es un paso importante si está intentando minar datos. Se puede añadir información a tales datos, de las fuentes de industria de partes externas normalizadas para hacer que el proceso de la minería de datos sea más exitoso y confiable, o proporcionar atributos derivados adicionales para una comprensión mejor de relaciones indirectas. Por ejemplo, los data warehouse

frecuentemente proporcionan pre-agregación a través de las líneas de negocio que comparten atributos comunes para propósitos de análisis de venta cruzados.

De igual forma que la depuración y transformación de datos, este paso se maneja mejor en un área de almacenamiento temporal. El enriquecimiento de datos, en particular la combinación de fuentes de datos externas con los datos para minarse, puede requerir varias actualizaciones en los datos y meta-datos, y tales actualizaciones generalmente no son aceptables en un data warehouse establecido.

2.4.6 Transformación de datos

La transformación de datos, desde el punto de vista de minería de datos, es el proceso de cambiar la forma o estructura de los datos existentes.

Las orientaciones para la transformación de datos son similares en la minería de datos y los data warehouse, y una cantidad grande del material de referencia existe para la transformación de datos en los entornos de los data warehouse.

Una de las formas más comunes de la transformación de datos usadas en la minería de datos es la conversión de atributos continuos en atributos discretos.

Muchos algoritmos de minería de datos funcionan mejor al trabajar con un número pequeño de atributos discretos, tales como rangos de sueldos, antes que atributos continuos, tales como sueldos reales. Este paso, como con otros pasos de transformación de datos, no añade información para los datos, ni limpia los datos; en vez de ello hace que los datos sean fáciles de modelar. Ciertos proveedores de algoritmos de minería de datos puede transformar los datos, en datos discretos de forma automática, usando una variedad de los algoritmos

diseñado para crear rangos discretos basadas en la distribución de datos dentro de un atributo contínuo.

Demasiados valores discretos dentro de un atributo sencillo pueden sobrecargar ciertos algoritmos de minería de datos. Por ejemplo, usando códigos postales de las direcciones de cliente para categorizar clientes por la región son una técnica excelente si se propone examinar una región pequeña. Si, por el contrario, se planea examinar los modelos de los clientes para un país entero, usando códigos postales pueden llevar a 50,000 o los valores más discretos dentro de un atributo sencillo; se debe usar un atributo con un alcance más ancho, tal como la ciudad o la información de estado suministrados por la dirección.

2.5 Preparación de un conjunto de casos

El conjunto de casos es usado para construir el conjunto inicial de reglas y patrones que sirven de base para un modelo de de minería de datos. Preparar un conjunto de casos de prueba es esencial para el éxito del proceso de la minería de datos. Generalmente, varios modelos diferentes de minería de datos se construirán del mismo conjunto de casos, como parte del proceso de minería de datos. Existen varias orientaciones básicas que se utilizan para la selección de un conjunto de casos de prueba.

Por lo general se desea tener tantos casos de prueba como sea posible cuando se crea un modelo de minería de datos, asegurándose de que el conjunto de casos represente la densidad y distribución del conjunto de casos de producción. Se recomienda seleccionar el conjunto más grande posible de casos, para alisar la distribución del conjunto de casos. El proceso de crear tal conjunto representativo de datos, es mejor realizando la selección de los registros completamente al azar.

En teoría, tal muestreo aleatorio debe proporcionar una vista verdaderamente imparcial de los datos.

Sin embargo, el muestreo aleatorio no siempre es el adecuado para construir un escenario específico, y un gran conjunto de casos no siempre puede ser lo mejor. Por ejemplo, si está intentando modelar una situación que se encuentra aislada dentro de sus datos, y el propósito es asegurar que la frecuencia de ocurrencias para la situación deseada es según la estadística bastante alta para proporcionar información de tendencia.

La técnica de aumentar la densidad de ocurrencias raras en una muestra se denomina overampling, influye en la información estadística transportada por el conjunto de casos de prueba. Tal influencia puede ser de beneficio al intentar modelar casos muy raros o aislados, los casos sensitivos donde la confirmación positiva de la existencia de un caso de éstos debe ser el primero en operarse, o cuando los casos para modelarse pueden ocurrir dentro de un período de tiempo muy corto.

Por ejemplo, si analizamos los casos de fraude con tarjeta de crédito, el que una transacción con tarjeta de crédito fraudulenta puede ocurrir sin el uso de una tarjeta de crédito, representa el 0.001 por ciento de todas transacciones con tarjeta de crédito almacenadas en un conjunto de datos. La muestra podría retornar teóricamente 1 caso de fraude por 100,000 transacciones mientras que el modelo podría proporcionar abrumadora información sobre transacciones exitosas, porque la desviación estándar para los casos de fraude puede ser inaceptablemente alta para el modelo propuesto. El modelo de minería de datos tendría el 99.999 por ciento de exactitud, pero también es completamente inútil para el escenario propuesto encontrar patrones de fraude en las transacciones de ninguna tarjeta.

En vez, el oversampling podría ser utilizado para proporcionar un número más grande de casos fraudulentos dentro del conjunto de casos. Un número más alto de casos fraudulentos puede proporcionar mejor perspicacia en los modelos detrás de transacciones fraudulentas. Existen unos cuantos inconvenientes con el oversampling, sin embargo, no es así, si se utiliza esta técnica cuidadosamente. La evaluación de un modelo de minería de datos creado con los datos de oversampled deber ser manipulado diferentemente debido al cambio en relaciones entre ocurrencias raras y comunes en el conjunto de casos.

Por ejemplo, en el fraude de tarjeta de crédito anteriormente descrito, el conjunto se construye de cinco años de datos de transacción, o aproximadamente 50 millones de registros. Esto significa que, fuera del convertidor de señal entero para minarse, sólo existen 500 registros fraudulentos. Si el muestreo aleatorio fue utilizado para construir un conjunto de casos con 1 millón de registros, sólo 10 casos podrían ser incluidos. Así, el conjunto de casos estaba oversampled, de modo que los casos fraudulentos representarían un 10 por ciento del número total de casos. Se extraen los 500 casos fraudulentos, así unos 4,500 casos adicionales se escogen para construir el conjunto de casos con 5,000 casos, del cual el 10 por ciento son transacciones fraudulentas. Al crear un modelo de minería de datos suponiendo la probabilidad de dos resultados probables, el conjunto de casos debe tener una relación de resultados raros a resultados comunes a aproximadamente del 10 al 40 por ciento, con del 20 al 30 por ciento es considerado como ideal. Esta relación puede lograrse por medio del oversampling, proporcionando un mejor enfoque de muestra estadística en el resultado.

La dificultad con este conjunto de casos es que un caso no fraudulento, en esencia, representa 11,111 en el conjunto de datos original. Evaluar un modelo de minería de datos para usar este oversampled significa tomar esta relación dentro de una cuenta al momento de computar, por ejemplo, la cantidad que provee el

modelo de minería de datos al evaluar las transacciones fraudulentas tiende a ser elevada.

2.6 Selección de casos

Al preparar un conjunto de casos, se debe seleccionar que los datos no sean ambiguos, en la medida de lo posible para obtener el resultado que pueda ser modelado. La ambigüedad del conjunto de datos escogido debe ser directamente proporcional al ancho del enfoque según el escenario que se desea representar. Por ejemplo, si se está intentando agrupar los productos que han fallado para poder descubrir posibles patrones de falla, seria apropiado escoger todos los productos que han fallado dentro del conjunto. Por el contrario, si se está tratando de predecir la falla de un producto para productos específicos debido a ciertas condiciones ambientales, se debería escoger sólo esos casos donde el producto específico directamente ha fallado a causa de condiciones ambientales, no todos los productos que han tenido fallas.

Esto puede parecer que se están agregando predisposiciones al conjunto de datos, pero una de las razones primordiales para tener grandes variaciones entre los resultados que se predicen y los reales al trabajar con los modelos de minería de datos se debe al hecho que los patrones almacenados en el modelo de minería de datos no es pertinente para la predicción del escenario deseado, y los patrones irrelevantes son introducidos en parte por casos ambiguos.

Una de las dificultades encontradas en la selección de casos es la definición de un escenario para obtener el resultado deseado. Por ejemplo, un escenario común supone agrupar casos según un conjunto de atributos conocidos para descubrir patrones ocultos. El algoritmo de clustering es usado sólo en este caso para poder encontrar atributos ocultos; el agrupamiento de casos basado en atributos

expuestos puede ser utilizado para revelar un atributo oculto, que a su vez puede ser la llave del comportamiento de un clustering.

Antes de seleccionar casos, hay que estar seguro de entender escenarios utilizados para crear un modelo de minería de datos y la información producida por el modelo de minería de datos creado.

El conjunto de casos no es la única fuente de patrones de información almacenados para el modelo de minería de datos. La evaluación del modelo de minería de datos, como un paso del proceso de minería de datos, puede permitirle refinar esta información almacenada con el uso de conjuntos de casos adicionales. El modelo de minería de datos, por refinación, puede no reconocer patrones irrelevantes para mejorar su exactitud de predicción. Pero, el modelo de minería de datos usa el conjunto de casos como su primer paso para poder obtener información de los datos, así su modelo se beneficiará por la selección cuidadosa del conjunto de casos.

2.7 Construcción del modelo de minería de datos

La construcción de un modelo de minería de datos consiste en la selección de un algoritmo de minería de datos que se ajuste a las metas que se desean obtener al evaluar el conjunto de casos. Este, a su vez, genera un conjunto de valores que reflejan unas o más vistas estadísticas con el comportamiento del conjunto de casos. Esta vista estadística se utiliza posteriormente para proporcionar posibles patrones en conjuntos de casos similares con resultados desconocidos.

Esto puede sonar simple, pero la construcción del modelo de minería de datos es mucho más complejo. El enfoque que se utiliza puede decidir la diferencia entre

un modelo de minería de datos exacto pero inútil y un modelo de minería de datos exacto y muy útil

La persona experta en el campo, que proporciona la guía en los datos que se están modelando, debe ser capaz de proporcionar suficiente información para poder tomar decisiones en una minería de datos exacta. La aproximación, a su vez, es esencial para decidir el algoritmo y casos que van a ser modelados.

Se debe observar, el proceso de construcción del modelo de minería de datos, como un proceso de exploración y descubrimiento. No existe ninguna fórmula para construir un modelo de minería de datos; experimentación y evaluación son pasos claves en el proceso de construcción, y en el proceso de minería de datos para un escenario específico se deben examinar cuidadosamente varias iteraciones antes de la construcción de un modelo de minería de datos efectivo.

Después que los datos se han seleccionado, la minería de datos se divide en las siguientes tareas:

2.7.1 Clasificación

La clasificación es el proceso de usar los atributos de un caso para asignarlo a una clase predefinida. Por ejemplo, los clientes pueden ser clasificados en varios niveles de riesgo para las aplicaciones de préstamos hipotecarios. La clasificación tiene un mejor resultado cuando un conjunto finito de clases puede definirse como clases de alto riesgo, medio riesgo o bajo riesgo.

2.7.2 Estimación

Mientras que la clasificación se utiliza para responder preguntas de un conjunto finito de clases, la estimación es usada para responder datos ficticios dentro de un conjunto de respuestas. Por ejemplo, usando información de censos para predecir los ingresos de las familias. Técnicas de clasificación y estimación son a menudo combinadas para un modelo de minería de datos.

2.7.3 Asociación

La asociación es el proceso de determinar la afinidad de casos dentro de un conjunto de casos, basado en la similitud de atributos. Simplemente se pone una, asociación cuando se determina que los casos pertenecen a un conjunto de casos. La asociación puede ser utilizada para determinar qué productos deben agruparse en un almacén, o que servicios son más utilizados para empacar.

2.7.4 Agrupación

La agrupación es el proceso de encontrar grupos en casos esparcidos, dividiéndolos en conjuntos más sencillos, distintos conjuntos de casos en varios subconjuntos se basan en la similitud de atributos. La agrupación es similar a la clasificación, excepto que la agrupación no requiere de un conjunto finito de las clases predefinidas; la agrupación simplemente agrupa los datos según las reglas y patrones inherentes en los datos que se basan en la similitud de sus atributos.

2.8 Modelo dirigido de Minería de datos

Minería de datos dirigida, es el uso de técnicas de clasificación y estimación para derivar un modelo de datos con resultados conocidos, que entonces se utiliza para llenar un escenario específico. El modelo se compara entonces contra los datos de un resultado desconocido para determinar la probabilidad de tales datos para satisfacer el mismo escenario. Por ejemplo, una ilustración común de minería de datos dirigida, es la tendencia de usuarios para cambiar o suprimir las cuentas.

Hablando en términos generales, los modelos de minería de datos manejan el proceso en minería de datos dirigida ejemplar. Clasificación y estimación son típicamente categorizadas como técnicas de minería de datos dirigidas.

Esta aproximación se emplea mejor en un escenario claro y puede ser empleado contra un grupo grande de datos históricos conocidos para construir un modelo de datos predictivo. Si se tiene una buena idea de escenario para ser modelado, y se poseen datos sólidos que ilustran tal escenario, pero no está seguro sobre el propio resultado o las relaciones que llevan a este resultado entonces no se tiene el modelo correcto. La minería de datos dirigida es tratada como una "caja negra", en que el usuario se preocupa cada vez menos sobre el modelo y más sobre los resultados que se pueden obtener mirando los datos a través del modelo.

2.8.1 Datos dirigidos de Minería de datos

Los datos dirigidos de la minería de datos se utilizan para descubrir las relaciones entre atributos de datos desconocidos, con o sin los datos conocidos con qué comparar el resultado. Puede o no puede existir un escenario específico. Agrupación y asociación, por ejemplo, son técnicas primariamente de la minería

de datos dirigida. En la minería de datos dirigida, los propios datos manejan el proceso de minería de datos.

La aproximación se emplea de mejor manera en situaciones en que se necesita descubrir reglas y patrones en datos desconocidos. Se puede llegar a descubrir atributos significativos y patrones en un diverso conjunto de datos sin usar datos de prueba o sin tener un escenario predefinido. Los datos dirigidos en la minería de datos son tratados como una operación de "caja blanca", en la que el usuario se interesa sobre ambos procesos utilizados por el algoritmo de la minería de datos para crear el modelo y los resultados generados por las vistas a través del modelo.

¿Cuál de los dos es mejor?

Hacerse esta pregunta es como preguntar si un martillo es mejor que una almagana; la respuesta depende del trabajo. La minería de datos depende de ambas técnicas de datos dirigidos y modelos dirigidos de datos para ser verdaderamente efectivo, en dependencia de qué preguntas se deben realizar y que datos deben ser analizados. Por ejemplo, la técnica de datos dirigidos puede ser utilizada en transacciones fraudulentas con tarjeta de crédito para aislar grupos de transacciones similares. La agrupación utiliza una aproximación de auto-comparación para encontrar grupos significativos de elementos de datos. Los atributos de cada elemento de datos son igualados a través de los atributos de todos los otros elementos de datos en el mismo conjunto, y son agrupados con otros registros que son bastante similares al elemento de datos del ejemplo.

Después que se han descubierto, estos grupos individuales de datos pueden ser modelados utilizando la técnica de datos dirigidos para construir un modelo de minería de datos sobre transacciones fraudulentas con tarjeta de crédito que se ajusta a cierto conjunto de atributos. El modelo entonces puede estar usado como

parte de un proceso de estimación, también como modelo dirigido, para predecir la posibilidad de fraude en las transacciones con tarjeta de crédito.

Varias tareas no son completamente cerradas tanto en los modelos de datos no dirigidos como en los datos dirigidos. Por ejemplo, un modelo de minería de datos de árbol de decisión puede ser usado para los modelos de datos dirigidos, para predecir datos desconocidos de los datos conocidos, o datos dirigidos, para hallar nuevos modelos referentes a un atributo de datos específico.

Los datos dirigidos y los modelos dirigidos de la minería de datos pueden ser empleados separadamente o en conjunto, esto varía dependiendo de las necesidades del negocio. No existe ninguna fórmula establecida para la minería de datos; cada conjunto de datos tiene sus propios modelos y reglamentos.

3. ANÁLISIS DE ALGORITMOS DE MINERIA DE DATOS

Con una cantidad enorme de los datos almacenados en bases de datos y data warehouse, es cada vez más importante desarrollar herramientas poderosas para el análisis de tales datos y obtener conocimiento a partir de allí. La minería de datos es un proceso que infiere conocimiento de una cantidad grande de datos. La minería de datos tiene tres componentes principales la agrupación o clasificación, los reglamentos de asociación y el análisis de sucesión.

Por definición simple, en la clasificación o agrupación se analiza un conjunto de datos y se genera un conjunto de reglas de agrupación que pueden ser utilizadas para clasificar datos futuros. Por ejemplo, uno puede clasificar enfermedades y obtener los síntomas que describen cada clase o subclase. Esto tiene mucho en común con el modo que trabajo tradicional en estadísticas y máquinas de aprendizaje. Sin embargo, existen puntos importantes que salen a relucir debido al tamaño de los datos.

Un problema importante en la minería de datos es la clasificación de las reglas de aprendizaje, las cuales tratan de encontrar más reglas que van dividiendo los datos en las clases predefinidas. En el dominio de la minería de datos donde millones de los registros y un gran número de atributos son complicados, el tiempo de ejecución de algoritmos puede ser excesivo, particularmente en aplicaciones interactivas.

Una regla de asociación, es una regla que implica ciertas relaciones de asociación entre un conjunto de objetos en una base de datos. En este proceso se descubre, un conjunto de reglas de asociación a múltiples niveles de abstracción. del conjunto más relevante de datos en una base de datos. Por ejemplo, se puede

descubrir un conjunto de síntomas que a menudo ocurren junto con ciertos tipos de enfermedades y estudiar las razones de esos síntomas. Después de encontrar una asociación interesante dentro las bases de datos se pueden revelar ciertos patrones útiles para soporte de decisión, comercialización selectiva, pronóstico financiero, diagnóstico médico, y muchas otras aplicaciones, han atraído una gran cantidad de atención en investigaciones recientes de minería de datos.

3.1 Clasificación de los algoritmos

En la clasificación de datos, se desarrolla una descripción o modelo para cada clase en una base de datos, basado en las características presentes en un conjunto de datos de prueba. Existen muchos métodos de clasificación de datos, incluyendo los métodos de árboles de decisión, métodos estadísticos, redes neuronales, conjuntos ásperos, base de datos orientadas a objetos, etc.

3.1.1 Métodos de clasificación de datos

- **Algoritmos estadísticos:** Sistemas de análisis estadístico tales como SAS y SPSS, han sido usados por analizadores para detectar patrones inusuales y explicarlos, utilizando modelos estadísticos tales como modelos lineales. Tales sistemas tienen su lugar y continuarán siendo usados.
- **Redes neuronales:** Las redes neuronales artificiales imitan la capacidad del cerebro humano para encontrar patrones y por lo tanto ciertos investigadores han sugerido aplicar los algoritmos de redes neuronales para el mapeo de patrones. Las redes neuronales se han aplicado con muy buen resultado en aplicaciones que requieren de clasificaciones.

- **Algoritmos genéticos:** Las técnicas de optimización de algoritmos utilizan procesos como combinaciones genéticas, mutación, y la selección natural en un diseño basado en los conceptos de evolución natural.
- **Método del vecino más próximo:** Esta es una técnica que clasifica cada registro, en un conjunto de datos basado en una combinación de las clases en el k(s) registro más similar para a ello en un conjunto de datos históricos.
- **Regla de inducción:** Es la extracción de reglas útiles if-then de datos basados en significación estadística.
- **Visualización de datos:** Es la interpretación de complejas relaciones visuales en datos multidimensionales.

3.1.2 Abstracción de datos

Muchos algoritmos sugieren abstraer datos de prueba antes de clasificarlos en varias clases. Existen varias alternativas para hacer abstracción antes de clasificación: un conjunto de datos puede generalizarse para una abstracción hasta el mínimo nivel de generalización, una abstracción de nivel intermedio, o un nivel de abstracción más alto. Un nivel de abstracción demasiado bajo puede dar por resultado las clases muy esparcidas, árboles de clasificación espesos, y la dificultad de la interpretación semántica concisa; mientras que un nivel demasiado alto puede dar por resultado la pérdida de la exactitud en la clasificación. La abstracción de datos ha sido implementada en el sistema DB-miner

3.1.3 Aprendizaje de la regla de clasificación

El aprendizaje de la regla de clasificación supone encontrar reglas o árboles de decisión que particionan los datos en clases predefinidas. Para cualquier problema

en el aprendizaje de la regla de clasificación, el conjunto de árboles de decisión es demasiado grande para realizar una búsqueda detalladamente.

La mayoría de los algoritmos basados en la inducción utilizan el método de Hunt como algoritmo básico. Aquí está una descripción recursiva del método de Hunt para construir un árbol de decisión de un conjunto T de casos en las clases denotadas como {C1, C2, & ,Ck }.

Caso 1: T contiene uno o más casos, todo pertenece a una clase simple Cj: El árbol de decisión para T es una hoja que identifica la clase Cj.

Caso 2: T no contiene ninguno caso: El árbol de decisión para el T es una hoja, pero la clase para ser asociada con la hoja debe ser determinada de la información aparte de T.

Caso 3: T contiene los casos que pertenecen a una mezcla de clases: una prueba es escogida, basado en un atributo sencillo, que tienen unos o más resultados mutuamente exclusivos {O1, O2 ,.. ,On }. T es dividido en los subconjuntos T1, T2, ...,Tn, donde T contiene todos los casos en el T que tenga el resultado Oi de la prueba escogida. El árbol de decisión para el T consiste de un nodo de decisión que identifica la prueba, y una rama para cada posible resultado. La misma mecánica de construcción del árbol es aplicada en forma recursiva a cada subconjunto de casos.

3.1.3.1 Algoritmo de ID3

El algoritmo de ID3 (Quinlan86) es un árbol de decisión que construye el algoritmo que determina la clasificación de objetos probando los valores de sus propiedades. Ello construye el árbol de arriba hacia abajo, empezando de un

conjunto de objetos y una especificación de propiedades. A cada nodo del árbol, una propiedad es ensayada y los resultados se dividen en un conjunto de objeto. Este proceso se hace en forma recursiva hasta que el conjunto en un subárbol dado es homogéneo con respecto a los criterios de clasificación, en otros términos ello contiene los objetos que pertenecen a la misma categoría. Esto se convierte en entonces un nodo de hoja. A cada nodo, la propiedad para probar es escogida basado en los criterios teóricos de la información que tratan de maximizar la información y minimiza la entropía. En términos más simples, esa propiedad se prueba que candidato es el que se debe dividir y poner en los subconjuntos más homogéneos.

3.1.3.2 Algoritmo C4.5

Este algoritmo era propuesto por Quinlan (1993). El algoritmo C4.5 genera un árbol de decisión de clasificación para el conjunto de datos dado por un particionamiento recursivo de los datos. La decisión crece utilizando la estrategia de profundidad. El algoritmo considera todas las posibles pruebas que pueden partir el conjunto de datos y escoge una prueba que da los mejores resultados de información. Para cada atributo discreto, se considera una prueba con resultados no menor del número de valores distintos del atributo considerado. Para cada atributo continuo, pruebas binarias suponiendo cada valor distinto del atributo. El conjunto de datos que pertenece a un nodo se clasifica por los valores de los atributos continuos y para cada valor distinto se realizan los cálculos en una revisión de los datos almacenados. Este proceso es repetido para cada uno de los atributos continuos.

3.1.3.3 Algoritmo de SLIQ

SLIQ (Aprendizaje supervisado en búsqueda) por sus siglas en inglés fue desarrollado por el equipo IBM Quest, es un árbol de decisión clasificador diseñado para clasificar grandes cantidades de datos. Usa una técnica de pre-clasificación en la fase de crecimiento de árbol. Esto ayuda a evitar que la clasificación sea compleja en cada nodo.

SLIQ guarda una lista de clasificación separada para cada atributo continuo y una lista separada llamada lista de clase. Una entrada en la lista de clase corresponde a un dato específico, éste posee una etiqueta y un nombre de la clase del nodo a que pertenece en el árbol de decisión. Una entrada en la lista de atributo clasificada tiene el valor de un atributo y el índice de datos en la lista de la clase. SLIQ crea el árbol de decisión en manera de primero a lo ancho. Para cada atributo, se examina la lista de clasificación y se calculan los valores de entropía para cada valor distinto de todos los nodos en la frontera del árbol de decisión simultáneamente. Después que los valores de entropía se han calculado para cada atributo, un atributo es escogido para dividir cada nodo en la frontera actual, y se expanden para tener una nueva frontera. Luego se vuelve a examinar nuevamente la lista de atributos clasificados para actualizar la lista con los nuevos nodos.

Mientras SLIQ maneja los datos grabados en disco ya que es demasiado grande para realizarlo en memoria, ello todavía requiere que cierta información que se mantenga residente en memoria la cual crece directamente proporcional al número de registros ingresados, poniendo un límite en el tamaño de datos de prueba. El equipo de Quest recientemente ha diseñado un nuevo algoritmo de clasificación basado del árbol de decisión llamado SPRINT que resuelve todas las restricciones de memoria.

3.1.4 Algoritmos paralelos

La mayoría de los algoritmos existentes, usan funciones heurísticas locales para manejar la complejidad computacional. Los rangos de complejidad computacional de estos algoritmos van de un AN(logN) a un $AN(logN)^2$ siendo N el número de items y A el número de atributos. Estos algoritmos son bastante rápidos para los campos de aplicación donde n son relativamente pequeños. Sin embargo, en el campo de la minería de datos donde millones de registros y un gran número de atributos son bastante complejos, el tiempo de ejecución de estos algoritmos puede convertirse demasiado alto, particularmente en aplicaciones interactivas. Los algoritmos paralelos han sido sugeridos por grupos de desarrolladores de algoritmos de minería de datos.

3.1.4.1 Idea básica

Inicialmente N items de datos son distribuidos de forma aleatoria a P procesadores tal que cada procesador tiene N=P datos. En este punto, todos los procesadores cooperan para expandir el nodo raíz de un árbol de decisión. Para esto, los procesadores necesitan decidir que atributo usar para generar nodos hijos de raíz. Esto puede ser hecho en tres pasos. En el primer paso, cada procesador reúne la información de distribución de clase de los datos locales. En el segundo paso, los procesadores cambian la información de distribución de clase local usando reducción global. Finalmente, cada procesador puede computar simultáneamente las ganancias de entropía de los atributos y encuentra el mejor atributo para dividir el nodo raíz.

Existen dos tipos de propuestas en el progreso. La propuesta de construcción de árbol sincrónico, el conjunto entero de procesadores expande sincrónicamente un nodo del árbol de decisión a la vez. La propuesta de construcción de árbol

particionado, cada nuevo nodo generado es expandido por un subconjunto de procesadores que ayudó la expansión del nodo matriz.

3.1.4.2 Propuesta de construcción de árbol sincrónico

En esta propuesta, todos los procesadores construyen un árbol de decisión sincrónicamente enviando y reciben información de una clase de distribución de datos locales.

Los pasos para esta propuesta son:

1. Escoger un nodo para expandirse según una estrategia de expansión de árbol de decisión (por ejemplo. por profundidad, a lo ancho o el primero mejor).
2. Para cada atributo de datos, reunir información de la clase de distribución de los datos locales al nodo actual.
3. Cambiar la información de la clase de distribución con todos otros procesadores y agregar la información de la clase de distribución para conseguir una distribución completa de todos los atributos.
4. Calcular las ganancias de entropía de cada atributo y seleccionar el mejor atributo para la expansión del nodo hijo.
5. Basado en los valores de atributo, se crean nodos hijos y se reparten los datos según los valores del atributo seleccionado.
6. Se repiten los pasos del 1 al 5 hasta que ya no hayan más nodos disponibles por la expansión.

La ventaja de esta propuesta es que no requiere ningún movimiento de los datos de prueba. Sin embargo, este algoritmo padece de altos costos de comunicación y carga desproporcionada.

La carga desproporcionada puede reducirse si todos los nodos en la frontera están expandidos simultáneamente, en una pasada de todos los datos para cada procesador se utiliza para computar la información de la clase de distribución para todos los nodos sobre la frontera. Este mejoramiento también reduce el volumen de comunicaciones y reduce el inicio de sobrecarga de mensajes, pero no reduce el volumen completo de comunicaciones. Ahora la única fuente de las desproporciones de carga es cuando ciertos nodos de hoja se convierten en nodos terminales. Esta desproporción de carga se puede minimizar además si el conjunto de datos es distribuido aleatoriamente.

3.1.4.3 Propuesta de construcción de árbol particionado

En esta propuesta, cada nodo n de la frontera del árbol de decisión se maneja por un subconjunto distinto de los procesadores P(n). Una vez el nodo N es expandido en nodos hijos, n1 , n2,.....n_k , el grupo del procesador P(n) también se particiona en k partes, P1 , P2,..... ,P_k, tal que el Pi maneja el nodo ni. Todos los items de datos son entremezclados de manera que los procesadores en el grupo Pi tienen los items de datos que pertenecen a la hoja ni.

A continuación se muestran los pasos para esta propuesta:

(A) Si el número de los nodos es menor que p(n),

1. Asignar un subconjunto de procesador a cada nodo de la hoja tal ese número de procesadores asignados a un nodo de hoja es proporcional al número de items de datos contenidos en el nodo.
2. Mezclar los datos de manera que cada subconjunto de procesadores contenga datos que pertenezcan a los nodos de hoja de la que es responsable.

3. Los subconjuntos de procesador son asignados a nodos diferentes de subárboles de los nodos responsables independientemente, siguiendo los pasos en forma recursiva.

(b) De otra manera,

1. Particionar los nodos de hoja en los grupos p(n) tal que cada agrupo tenga igual número de items de datos. Asignar cada procesador a un grupo de nodos.
2. Mezclar el conjunto de datos de manera que cada procesador tenga items de los datos que pertenezcan a los nodos de hoja de la que es responsable.
3. Ahora la expansión de los subárboles a un grupo de nodos es completamente independientemente a cada proceso.

Al fin, el árbol de decisión entero es construido combinando subárboles de cada procesador.

La ventaja de esta propuesta es que solo una vez, un procesador llega a ser responsable para un nodo, éste puede desarrollar un subárbol del árbol de decisión independientemente sin ninguna comunicación con los de arriba. Existen varias desventajas de esta propuesta. La primera desventaja es que requiere de movimiento de datos después de cada expansión un nodo hasta que un procesador se vuelven responsable para un subárbol entero. Los costos de comunicación son particularmente caros en la expansión de la parte superior del árbol de decisión. La segunda desventaja se debe a balance de carga.

3.2 Algoritmos de reglas de asociación

Una regla de asociación es una regla que implica ciertas relaciones de asociación entre un conjunto de objetos en una base de datos. De un conjunto de transacciones, donde cada transacción es un conjunto de literales (llamados items), una regla de asociación es una expresión de la forma X Y, donde X y Y son conjuntos de datos. El significado intuitivo de tal regla es que las transacciones de la base de datos que contiene la X tiende a contener Y. Un ejemplo de una regla de asociación es: "el 30% de transacciones que contienen cerveza también contenga pañales; 2% de todas las transacciones contenga ambos de estos items". Aquí 30% es llamado la confianza de la regla, y 2% el apoyo de la regla. El problema es encontrar todas las reglas de asociación que satisfaga un mínimo de usuarios específicos.

3.2.1 Algoritmo Apriori

Un algoritmo de regla de asociación Apriori se ha desarrollado para reglas de minería de datos para grandes transacciones sobre bases de datos por el equipo de IBM Quest.

Este algoritmo divide el problema de las reglas de asociación en dos partes:

1. Primero se debe encontrar todas las combinaciones de items que tienen soporte de transacción. Esas combinaciones se denominan frecuencia de conjunto de items.
2. Utilizar las frecuencias de los conjuntos de items para generar las reglas que se desean. La idea general es que si, por ejemplo, ABCD y AB son las frecuencias conjuntos de items, entonces podemos determinar si la regla AB CD se mantiene al computar el radio R =

soporte (ABCD) /soporte (AB). La regla se mantiene sólo si R >= mínimo de confianza. La regla tendrá el mínimo de soporte porque ABCD tiene mas frecuencia. El algoritmo de Apriori usado en la búsqueda para encontrar la frecuencia de todos los conjuntos de items se describe a continuación:

```
procedure AprioriAlg()
begin
        L1 := {frequent 1-itemsets};
        for ( k := 2; Lk-1 0; k++ ) do {
           Ck= apriori-gen(Lk-1) ; // nuevas candidatas
        for todas las transacciones t en el  dataset do {
          for todas  las candidatas c Ck contenidas en t do
                  c:count++
          }
          Lk = { c Ck | c:count >= min-support}
        }
        Answer := k Lk
end
```

3.2.2 Algoritmo distribuido/paralelo

Las bases de datos o los data warehouse pueden almacenar una cantidad enorme de datos que pueden ser minados. Las reglas de asociación en tales bases de datos pueden requerir poder de procesamiento substancial. una posible solución a este problema puede ser un sistema distribuido. Además, muchas bases de datos grandes son distribuidas lo que hace más factible el uso de algoritmos distribuidos.

Los costos principales de las reglas de asociación es la manipulación de los conjuntos de registros demasiado grandes en la base de datos. Un problema es que se puede manipular los conjuntos fácilmente si es localmente, pero un conjunto de registros localmente grande no puede ser un conjunto a nivel mundial. Dado esto resulta muy caro emitir la manipulación entera de otros sitios, una opción es emitir todas las cuentas de todos los conjuntos de registros, sin importar si son grandes o pequeños, a otros sitios. Sin embargo, una base de datos puede contener combinaciones enormes de conjuntos de registros, y ello supondrá pasar un número enorme de mensajes.

Un algoritmo distribuido de minería de datos es el FDM (reglas de asociación minera rápidamente distribuidos) por sus siglas en ingles, tienen las características siguientes:

1. La generación de los conjuntos candidatos es en el mismo espíritu de Apriori. Sin embargo, ciertas relaciones entre conjuntos localmente grandes y mundialmente grandes se exploran para generar un conjunto más pequeño del candidato propuesto para cada iteración y así reducir el número de mensajes para que pueda ser pasado.
2. Después de que el conjunto candidato se haya generado, dos técnicas de depuración, depuración local y depuración global, son desarrolladas para poder depurar ciertos conjuntos de cada grupo individual.
3. A fin de determinar si un conjunto candidato es grande, este algoritmo requiere solo (n) mensajes enviados para el soporte de intercambio, donde la n es el número de sitios en la red. Esto es mucho menos que una adaptación recta de Apriori, que requiera n^2 mensajes enviados.

3.3. Análisis secuencial

3.3.1 Patrones secuenciales

Los datos de entrada son un conjunto de las sucesiones, llamadas sucesiones de datos. Cada sucesión de datos es una lista ordenada de transacciones o conjunto de datos, donde cada transacción es un conjunto de registros. Típicamente existe un tiempo asociado con cada transacción. Un patrón secuencial también consiste de una lista de conjuntos de registros. El problema es encontrar todos los patrones secuenciales con mínimo de soporte de un usuario especializado, donde el soporte de un patrón secuencial es el porcentaje de sucesiones de datos que contiene el modelo.

Un ejemplo de tal modelo es que si los clientes típicamente rentan la película "Star Wars", entonces también rentan "Empire Strikes Back", y también "Return of the Jedi". Estos alquileres no necesariamente tienen que ser secuenciales. Los elementos de un patrón secuencial no por fuerza tienen que ser artículos simples. sábanas, cobertores y fundas de almohadas, por lo regular lo acompaña una cubrecama, y también una cama. Éste es un ejemplo de un patrón secuencial en que los elementos son conjuntos de registros. Este problema se motiva inicialmente por aplicaciones en la industria de venta al menudeo, incluyendo envío de paquetes y satisfacción de cliente. Pero los resultados se aplican mucho en el campo científico y de negocio. Por ejemplo, en el campo médico, una sucesión de datos puede corresponder a los síntomas o enfermedades de un paciente, se puede diagnosticar que los síntomas expuestos son la causa una enfermedad durante una visita al médico. Los patrones pueden descubrir que usando estos datos en la investigación de enfermedades pueden ayudar a identificar los síntomas que preceden ciertas enfermedades.

3.3.2 Algoritmos para encontrar patrones secuenciales

Varios grupos que se encuentran trabajando en este campo sugieren algoritmos de patrones consecutivos para la minería de datos. Los que se listan a continuación son algoritmos propuestos por equipo Quest de IBM.

3.3.2.1 Algoritmo

El problema de patrones secuenciales en la minería de datos se describe en las fases siguientes:

- Fase de ordenamiento. Este paso implícitamente convierte la base de datos de original en una base de datos de sucesiones.

- Fase de item. En esta fase encontramos el conjunto de todos los items L. Se encuentran también de forma simultánea el conjunto de toda las sucesiones grandes,

- Fase de transformación. Se necesita determinar repetidamente si en un conjunto dado de sucesiones grandes existe una sucesión de clientes. Para hacer esta prueba rápidamente, transformamos cada sucesión de cliente en una representación alternativa. En una sucesión de cliente transformada, cada transacción es reemplazada por el conjunto de todos los items contenidos en esa transacción. Si una sucesión de cliente no contiene ningún item, esta sucesión es desechada de la base de datos transformada. Sin embargo, todavía contribuye en el conteo total de clientes. una sucesión de cliente se representa ahora por una lista de conjuntos de los items.

- Fase de sucesión. Se utiliza el conjunto de items para encontrar las sucesiones deseadas.

- Fase máxima. Encuentra las sucesiones máximas entre el conjunto de sucesiones grandes. En ciertos algoritmos esta fase es combinada con la fase de sucesión para reducir el tiempo al contar las sucesiones no máximas.

La estructura general de los algoritmos para la fase de sucesión es que ellos utilizan múltiples pasos sobre los datos. En cada paso, se empieza con un conjunto de sucesiones grandes. Se utiliza una pequeña parte para generar nuevas sucesiones potencialmente grandes, estas se llaman sucesiones candidatas. Se encuentra el soporte para estas sucesiones candidatas durante una revisión sobre los datos. Al final la revisión, se determina que las sucesiones candidatas son en realidad grandes. Estas sucesiones grandes se convierten en la semilla para el próximo paso.

Existe dos familias de algoritmos los de cuenta-todo y los de cuenta-ciertos. Los algoritmos de cuenta-todo son los que cuentan todas las sucesiones grandes, incluyendo sucesiones no máximas. Las sucesiones no máximas se deben dividir en la fase máxima. AprioriAll es un algoritmo de cuenta-todo, basado en el algoritmo de Apriori para encontrar items grandes. Apriori es un algoritmo de cuenta-ciertos. La intuición detrás de estos algoritmos es la que si nos interesamos sólo en sucesiones máximas, podemos evitar contar sucesiones que son contenidas en una sucesión más larga si primero se cuentan las sucesiones. Sin embargo, se tiene que tener cuidado para no contar una gran cantidad de sucesiones más largas que no tienen apoyo mínimo. De otra manera, el tiempo ahorrado en no contar las sucesiones contenidas en una sucesión puede se menos del tiempo gastado contando sucesiones sin apoyo mínimo que nunca se han contado porque sus subsecuencias no eran grandes.

3.3.2.2 Algoritmo apriori/All

L_1 = large 1-sequences; // resultado de la fase item
for (k = 2; L_{k-1} 0; k++) **do**
begin

C_k = nuevas candidatas generados L_{k-1}
foreach secuencia de clients *c* en la base de datos **do**
incrementa el contador de todos los candidates en C_k que son contenidos en *c*.
L_k = canditadas en C_k con soporte mínimo.

end
resultado = Maxima Sequencia en $_k$ L_k ;

En cada paso de este algoritmo, se usan las sucesiones grandes del paso previo para generar la secuencia candidata y entonces se mide el soporte realizando una revisión sobre la base de datos. Al final del paso, el soporte de las candidatas se utiliza para determinar las sucesiones grandes. En el primer paso, la salida de la fase de items se utiliza para inicializar el conjunto primer conjunto grande. Los candidatos se ordenan para encontrar rápidamente todos los candidatos contenidos en una sucesión de cliente.

3.3.2.2.1 Generación de candidatos apriori

La función de generación apriori toma el argumento L_{k-1}, del conjunto de secuencias (k-1). Trabaja de la siguiente manera, el primero une L_{k-1} con L_{k-1}

Insert into C_k
Select p.litemset1 ,..., p.litemsetk-1, q.litemsetk-1
from L_{k-1} p, 1 q de L_k
where p.litemset1 = q.litemset1 ,. . .,
p.litemsetk-2 = q.litemsetk-2 ;

Después se borran todas las secuencias c C_k tal que algunas subsecuencias (k-1) de c no están en Lk-1.

Ejemplo

Si se considera una base de datos con siguientes sucesiones de clientes:

1 5 2 3 4

1 3 4 3 5

1} 2} 3} 4}

1} 3} 5}

4} 5}

Las sucesiones de cliente se encuentran ya en forma transformada donde cada transacción se ha reemplazado por el conjunto de los items contenidos en la transacción y los Litems que han reemplazados por números enteros. El soporte mínimo se ha especificado para ser 40%. La primera pasada sobre base de datos se hace en la fase de Llitems, Las sucesiones grandes junto con su soporte a final de la segunda, tercera, y cuarta pasada Ningún candidato es generado para la quinta pasada. Las sucesiones grandes máximas pueden ser las tres sucesiones siguientes: 1 2 3 4, 1 3 5 y 4 5.

3.3.2.3 Algoritmo aprioriSome

En este algoritmo se cuentan sólo las sucesiones de ciertas longitudes. Por ejemplo, nosotros podemos contar sucesiones de longitud 1,2,4 y 6 y en la fase forward y en la fase backward se cuentan sucesiones de longitud 3 y 5. La función

next toma como el parámetro la longitud de sucesiones incluidas en el último paso y retorna la longitud de sucesiones para ser incluidas en próximo paso. Así, esta función determina exactamente que sucesiones han sido contadas, y balancea el |tradeoff| entre el tiempo gastado al contar sucesiones no máximas y las extensiones de cálculo de sucesiones candidatas pequeñas. Un extremo es next(k) = k + 1 (k es la longitud para los candidatos fueron contados de último), cuando todas las sucesiones no máximas se cuentan, pero ningunas extensiones de sucesiones candidatas pequeñas se cuentan. En este caso, AprioriSome se convierte en AprioriAll. El otro extremo es una función como next(k) = 100 * k, cuando casi ninguna sucesión no máxima es contada, pero gran cantidad de extensiones de candidatas pequeños se cuentan.

// Forward Phase
L_1 = large 1-sequences; // resultado de la fase item
$C_1 = L_1$;
last = 1; // el ultimo contado C_{last}
for (k = 2; C_{k-1} 0 **and** L_{last} 0; k++) **do**
begin
if (L_{k-1} known) **then**
  C_k= nuevas candidatas generas desde L_{k-1} ;
else
  C_k= nuevas candidatas generas desde C_{k-1} ;
  if (k == next(last)) **then begin**
    foreach secuencia de cliente *c* en la base de datos **do**
      incrementa el contador de todas las candidatas en C_k que son contenidas en *c*.
      L_k = candidatas en C_k con soporte mínimo
      last = k;
  end
end

// Backward Phase
for (k-- ; k >=1; k==) **do**
  if (L_k no se encuentra en forward phase) **then begin**
    elimina todas las secuencias en C_k contenidas en algunas L_i , i>k;
  foreach secuencia de clientes en D_T **do**
    incrementa el contador de las candidatas en C_k que son contenidas en *c*.
    L_k = candidatas en C_k con soporte mínimo
end
else // L_k ya conocidos
  elimina todas las secuencias en L_k contenidas en algunas L_i , i > k.
Answer = $_k$ L_k ;

hit_k denota el radio del número de sucesiones grandes k hasta el número de sucesiones candidatas k (de L_k a C_k). La función next que se ha utilizado se muestra a continuación:

```
function next(k: integer)
begin
        if (hit_k< 0.666) return k + 1;
        elsif (hit_k < 0.75 ) return k + 2;
        elsif (hit_k < 0.80 ) return k + 3;
        elsif (hit_k < 0.85 ) return k + 4;
        else return k + 5;
end
```

La intuición detrás de la heurística es el porcentaje de candidatos contínuos en la pasada actual que tengan el soporte mínimo en aumento, el tiempo utilizado contando extensiones de candidatos pequeños cuando saltamos una longitud es más corto.

Se utiliza la función de generación de sucesiones a priori dada con aterioridad para generar nuevas sucesiones candidatas. Sin embargo, en el k-ésimo paso, posiblemente todavía no se tienen grandes conjuntos de sucesiones L_{k-1} disponibles como no se ha contado (k-1) sucesiones candidatas. En ese caso, se utiliza el conjunto de secuencias C_{k-1} para generar C_k. La corrección se mantiene por C_{k-1} y L_{k-1}

En la fase backward, se cuentan sucesiones para las longitudes que se pasan por alto durante la fase forward, después de esto se borran todas las secuencias contenidas en algunas sucesiones grandes. Estas sucesiones más pequeñas no pueden ser en la respuesta porque lo que interesa son las sucesiones máximas. También se borran las sucesiones grandes encontradas en la fase forward que no son máximas.

Ejemplo:

Encontramos la primera sucesión grande (L_1) en la fase de item:

Tabla II. Algoritmo aprioriSome, sucesión L_1

1-Sucesión	Soporte
1	4
2	2
3	4
4	4
5	4

Para simplicidad del ejemplo se toma, f (k) = 2k. En la segunda pasada se cuenta C_2 para conseguir L_2.

Tabla III. Algoritmo aprioriSome, sucesión L_2

2-Sucesión	Soporte
1 2	2
1 3	4
1 4	3
1 5	3
2 3	2
2 4	2
3 4	3
3 5	2
4 5	2

Después de la tercera pasada, la generación de sucesiones apriori es llamada con L_2 como el argumento para conseguir C_3. No contamos C_3 , y por lo tanto no generamos L_3. Después la generación de sucesiones apriori es llamada con C_3 para conseguir C_4. Después de contar C_4 para conseguir L_4 se prueba generar C_5 la cual retorna vacío.

Tabla IV. Algoritmo a prioriSome, sucesión L_3

3-Sucesión	Soporte
1 2 3	2
1 2 4	2
1 3 4	3
1 3 5	2
2 3 4	2

Tabla V. Algoritmo a prioriSome, sucesión L_4

4-Sucesión	Soporte
1 2 3 4	2

Empezamos entonces la fase backward. No se pudo borrar nada de L_4 desde que no existen sucesiones grandes. Se pasó el conteo de soporte para las sucesiones C_3 en la fase forward. Después de borrar esas sucesiones en C_3 que son subsecuencias de las sucesiones en L_4, y subsecuencias de 1 2 3 4, se dejan con las sucesiones 1 3 5 y 3 4 5. Esta puede ser contada para conseguir 1 3 5 como un máxima tercera sucesión. Después, todas las sucesiones en L_2 excepto 4 y 5 se borran desde que son contenidas en cierta sucesión más larga. Por la misma razón, todas las sucesiones en L_1 se borran también.

3.3.2.4 Ejecución relativa de los dos algoritmos

Como se esperaba, los tiempos de ejecución de todos los algoritmos se incrementan conforme el soporte va disminuyendo debido a un aumento en el número de sucesiones grandes. La generación apriori no cuenta cualquier sucesión candidato que contenga alguna subsecuencia que no sea grande. La ventaja principal de AprioriSome sobre AprioriAll es que evita el conteo de muchas sucesiones no máximas. Sin embargo, esta ventaja es reducida debido a dos razones. En primer lugar, las candidatas C_k en AprioriAll se generan usando L_{k-1}, mientras que AprioriSome a veces usa C_{k-1} para este propósito. Desde C_{k-1} hasta L_{k-1}, el número de candidatas generadas usando AprioriSome puede ser más grande. En segundo lugar, aunque AprioriSome se salta el cálculo de candidatas de ciertas longitudes, no obstante quedan residentes en memoria. Si la memoria se llegara a llenar, AprioriSome es forzado a contar el último conjunto de

candidatas generadas aún si la heurística sugiere saltar más conjuntos de candidatas.

Este efecto disminuye la distancia de los saltos entre dos conjuntos de candidatas que en realidad son contados, y AprioriSome empieza a comportarse más como AprioriAll. Para soporte bajo, existen más sucesiones grandes, y por lo tanto más sucesiones que no son máximas, y AprioriSome en este aspecto lo hace mejor.

CONCLUSIONES

1. Se determinó que la minería de datos es el conjunto de herramientas y técnicas de análisis de datos que permiten crear escenarios, de los cuales se puede obtener información útil para la toma de decisiones a nivel gerencial.

2. Las técnicas que utiliza la minería de datos para la exploración consisten el la identificación de patrones.

3. El proceso de la minería de datos genera conocimiento por medio de la depuración, enriquecimiento y transformación de datos que sirve para la creación de un modelo en el que se evalúa un conjunto de casos.

RECOMENDACIONES

1. Promover el conocimiento de la minería de datos, ya que puede ser utilizada para minimizar costos o para incrementar las ganancias de un negocio.

2. La mejor manera de sacarle provecho a la minería de datos es utilizándola en conjunto con un data warehouse, obteniendo ventajas de las técnicas de depuración de datos del data warehouse.

3. Para obtener un mejor resultado conviene hacer una selección del algoritmo a utilizar de acuerdo al caso de estudio.

BIBLIOGRAFÍA

1. Hilda Ruth Flores Muñoz Diseño de un plan de implementación de data mining como soporte a la toma de decisiones con base en el registro académico de la Universidad Rafael Landívar. Junio 2002.

2. Two Crows Corporation
   Introduction to Data Mining and Knowledge Discovery

3. Universidad Tecnológica Nacional
   Facultad Regional San Nicolás
   Secretaría de Ciencia y Tecnología
   Grupo Ingeniería del Conocimiento
   REDES NEURONALES ARTIFICIALES

REFERENCIAS ELECTRÓNICAS

4. Jiawei Han and Micheline Kamber
   Data Mining: Concepts and Techniques,
   The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor
   Morgan Kaufmann Publishers, August 2000.
   http://www.cs.sfu.ca/~han/dmbook

5. Karuna Pande Joshi
   Analysis of Data Mining Algorithms
   http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm

6. Data Mining Techniques
http://www.statsoftinc.com/textbook/stdatmin.html

7. Database Systems Research Laboratory
Simon Fraser University, B.C., Canada V5A 1S6
DMQL: A Data Mining Query Language for Relational Databases
http://www.cs.ualberta.ca/~zaiane/postscript/dmql96.pdf

8. Fourth International Conference on Knowledge Discovery & Data Mining
John F. Elder IV and Dean W. Abbott
A Comparison of Leading Data Mining Tools
http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf

9. DBMiner
A data mining tool for large relational databases
http://db.cs.sfu.ca/sections/projects/dbminer.html

10. Computer Science Department
North Carolina State University
Real Time Data Mining-based Intrusion Detection
http://www1.cs.columbia.edu/ids/concept/dmids-discex01.html

11. Data Mining: What is Data Mining?
http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

12. Kurt Thearling, Ph.D.
Director, Advanced Data Mining
http://thearling.com