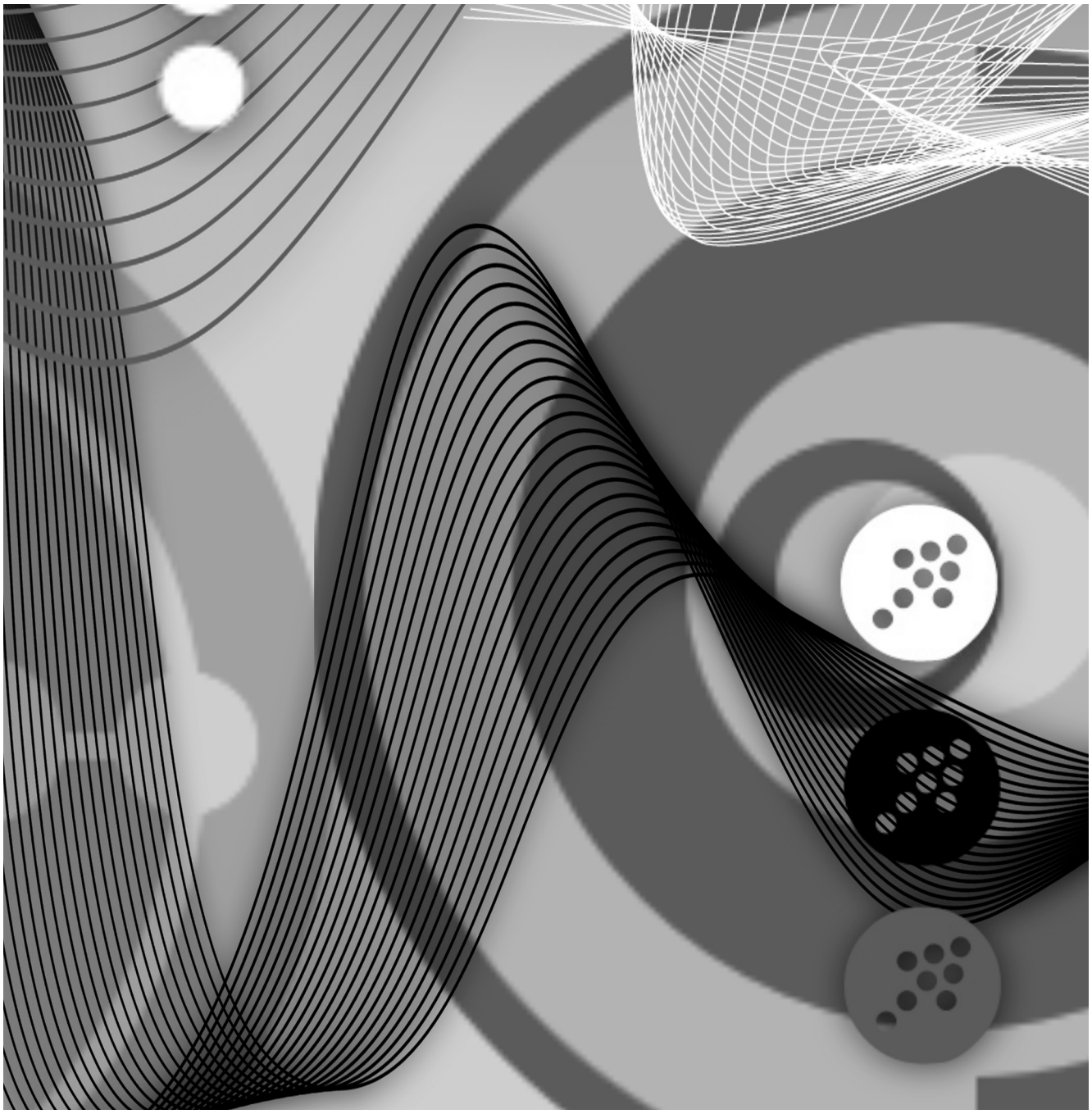
Revista Cubana de Ciencias Informáticas
ISSN: 1994-1536
rcci@uci.cu
Universidad de las Ciencias Informáticas
Cuba

Rodríguez Suárez, Yuniet; Díaz Amador, Anolandy
Herramientas de Minería de Datos
Revista Cubana de Ciencias Informáticas, vol. 3, núm. 3-4, julio-diciembre, 2009, pp. 73-80
Universidad de las Ciencias Informáticas
Ciudad de la Habana, Cuba

Disponible en: http://www.redalyc.org/articulo.oa?id=378343637009

Cómo citar el artículo
Número completo
Más información del artículo
Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto



RCCI Vol. 3, No. 3-4 JULIO-DICIEMBRE, 2009 p. 73-80

Recibido: 11/06/2009

Herramientas de Minería de Datos

Data Mining Tools

Yuniet Rodríguez Suárez[1]* y Anolandy Díaz Amador[1]

[1] Universidad de las Ciencias Informáticas, Carreterra a San Antonio de los Baños km 2 1/2. Rpato Torrens, Boyeros, La Habana, C.P.: 19370. Cuba

*Autor para correspondencia: yuniet@hab.uci.cu

Resumen

En la actual sociedad de la información, donde cada día se multiplica la cantidad de datos almacenados casi de forma exponencial, la minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización. La minería de datos se define también como el análisis y descubrimiento de conocimiento a partir de datos. La minería de datos hace uso de todas las técnicas que puedan aportar información útil, desde un sencillo análisis gráfico, pasando por métodos estadísticos más o menos complejos, complementados con métodos y algoritmos del campo de la inteligencia artificial y el aprendizaje automático que resuelven problemas típicos de agrupamiento automático, clasificación, predicción de valores, detección de patrones, asociación de atributos. En este trabajo se hace un estudio de herramientas que se utilizan en la minería de datos así como algunas de las aplicaciones y deficiencias que tiene la misma.

Palabras clave: Extraer, herramientas, minería de datos.

Abstract

In today's information society, where every day is multiplied by the amount of data stored almost exponentially, data mining is a fundamental tool to analyze and exploit them effectively to the objectives of any organization. Data mining is also defined as the analysis and knowledge discovery from data. Data mining uses all the techniques that can provide useful information, from a simple graphical analysis, statistical methods through more or less complex, complemented with methods and algorithms in the field of artificial intelligence and machine learning to problems typical automatic clustering, classification, value prediction, pattern detection, association of attributes. In this paper a study of tools used in data mining and some of the applications and has the same shortcomings.

Keywords: Extract, tools, Data Mining.

Introducción

El almacenamiento de información en formatos digitales es cada vez más barato y sencillo. Se genera gran cantidad de datos. Hay que intentar sacar partido a estos volúmenes de información para la toma de decisiones. La tecnología informática constituye la infraestructura fundamental de las grandes organizaciones y permite, hoy, registrar múltiples detalles de la vida de las empresas. Las bases de datos posibilitan almacenar cada transacción, así como otros muchos elementos que reflejan la interacción de la organización con otras organizaciones, clientes, o internamente, entre sus divisiones y empleados, etcétera. Es imprescindible convertir los grandes volúmenes de datos existentes en experiencia, conocimiento y sabiduría, formas que atesora la humanidad para que sea útil a la toma de decisiones, especialmente en las grandes organizaciones y proyectos científicos. La búsqueda de información relevante siempre es útil a la administración empresarial: el control de la producción, el análisis de los mercados, el diseño en ingeniería y la exploración científica, porque pueden ofrecer las respuestas más apropiadas a las necesidades de información. La minería de datos, es un conjunto de técnicas agrupadas con el fin de crear mecanismos adecuados de dirección, entre ellas puede citarse la estadística, el reconocimiento de patrones, la clasificación y la predicción. Para descubrir patrones de relaciones útiles en un conjunto de datos se empezaron a utilizar métodos que fueron denominados de diferente forma. El término *Data Mining*, en inglés, no era, al principio, del agrado de muchos estadísticos, porque sus investigaciones estaban dirigidas a procesar y reprocesar suficientemente los datos, hasta que confirmasen o refutasen las hipótesis planteadas. Esta tecnología ha sido de gran ayuda en áreas como la banca, telecomunicaciones, seguros y otros. En la actualidad hay un número creciente de organizaciones inmersas en proyectos de Minería de Datos o *Data Mining*. La tecnología se puede aplicar a cualquier organización que disponga de una gran cantidad de datos y que se plantee explotarlos para obtener reglas de negocio o mejorar el servicio que presta.

Desarrollo

La idea de Minería de Datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como *data fishing*, *data mining* o *data archaeology* con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh

Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de *data mining* y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios. Es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software.

Definiciones de Minería de Datos

La definición de Minería de Datos puede variar entre los diferentes investigadores ya sean estadísticos, analistas de datos u otros. A continuación se muestran algunas definiciones:

- "La minería de datos puede definirse como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, a partir de grandes volúmenes de datos" (González, 2006).
- "La minería de datos es la exploración y análisis, mediante métodos automáticos o semiautomáticos, de grandes cantidades de datos para descubrir reglas o patrones significativos" (Berry y Linoff, 1997).
- "La minería de datos es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma autorizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" (Piatetsky-Shapiro y Frawley, 1991).
- "La minería de datos es la extracción de información implícita, previamente desconocida y potencialmente útil de una base de datos" (Witten y Frank, 2000).
- "La minería de datos combina técnicas de la estadística, inteligencia artificial, bases de datos, visualización y otras áreas, para descubrir, de forma automática o semiautomática, modelos de series de datos" (Siebes, 2000)
- "La minería de datos es el análisis de habitualmente grandes, series de datos para encontrar relaciones inesperadas y resumir la información de nuevas maneras que sean entendibles y útiles por el propietario de los datos" (Thuraisingham, 1999).

En el ámbito del descubrimiento de conocimiento en bases de datos o *Knowledge Discovery in Databases* (KDD) tiene otro significado, el KDD se empezó a utilizar en 1989 (Piatetsky-Shapiro y Frawley, 1991) popularizándose por los expertos en inteligencia artificial (IA) y aprendizaje de ordenadores (*Machine Learning*), por lo que la minería de datos se define como:

- "Minería de Datos consiste en obtener modelos comprensibles o patrones de una base de datos" (Siebes, 2000).
- "Minería de Datos: búsqueda de patrones de interés mediante árboles o reglas de clasificación, técnicas de regresión, clusterizado, modelizado secuencial, dependencias, ect" (Wang, 1999).

Los investigadores la definen diferente yo coincido con todos, resumiendo la minería de datos es el análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones.

Algunas dificultades en la aplicación de Minería de Datos

Problemas a los que se enfrenta cualquier proyecto de Minería de Datos

El número de posibles relaciones es demasiado grande, y resulta prácticamente imposible validar cada una de ellas. Para resolver este problema se utilizan estrategias de búsqueda, extraídas del área de aprendizaje automático (Berry y Linoff, 1997). Las herramientas funcionan mejor fijándoles objetivos de búsqueda concretos. Si bien la minería de datos es la impresión de que se puede simplemente aplicar como herramienta a los datos, se debe tener un objetivo, o al menos una idea general de lo que busca. El coste de esta prospección de datos debe ser coherente con el beneficio esperado. Si bien las herramientas han bajado su precio, el coste en tiempo, personal y consultoría se ha incrementado, llegando en algunos casos a hacer no viable el proyecto. Suele funcionar mejor en problemas ligados a empresas de éxito que en otros casos, debido a la gran dependencia que estas herramientas tienen respecto a todos los estamentos de la empresa, desde mantenimiento a compras. Es necesario trabajar en estrecha colaboración con expertos en el negocio para definir modelos. A veces la información esta corrompida, tiene ruido o simplemente le faltan partes. Para esto se aplican

técnicas estadísticas que ayudan a estimar la confiabilidad de las relaciones halladas.

Aplicaciones de la Minería de Datos

Las técnicas de minería de datos se están utilizando desde hace varios años para la obtención de patrones en los datos y para la extracción de información valiosa en el campo de la Ingeniería del Software. Entre estas aplicaciones podemos citar:

- La utilización de árboles de decisión en la construcción de modelos de clasificación de diferentes características del desarrollo de software.
- Aspectos climatológicos: predicción de tormentas, etc.
- Medicina: encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.
- Mercadotécnica: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos.
- Inversión en casas de bolsa y banca: análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- Análisis de canastas de mercado para mejorar la organización de tiendas, segmentación de mercado (clustering).
- Determinación de niveles de audiencia de programas televisivos.
- Industria y manufactura: diagnóstico de fallas.

Algoritmos y técnicas de Minería de Datos

La minería de datos es un proceso de extracción de información y búsqueda de patrones de comportamiento que a simple vista se ocultan entre grandes cantidades de información, existen varios algoritmos y técnicas que ayudan en obtener la información.

Algoritmos:

1. Supervisados o predictivos: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.
2. No supervisados o del descubrimiento del conocimiento: con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

Existen varias técnicas de recopilación de datos que muchas de estas son utilizadas por las herramientas que usan minería de datos:

Almacenamiento de datos (*Data Warehousing*): El almacenamiento de datos se define como un proceso de organización de grandes cantidades de datos de diversos tipos guardados en la organización con el objetivo de facilitar la re-cuperación de la misma con fines analíticos. El al-macenamiento de datos tiene un gran importancia en el proceso de minería de datos pues en cierta medida, permite la recuperación o al menos la referencia a determinados conjuntos de datos de importancia para un proceso de toma de decisión dado. En la actualidad existe gran variedad de sistemas comerciales para el almacenamiento de datos entre los que se destacan Oracle, Sybase, MS SQL Server, entre otros.

Análisis exploratorio de datos (*Exploratory Data Analysis (EDA)*): Las técnicas de análisis exploratorio de datos juegan un papel muy im-portante en la minería de datos. Las mismas tienen como objetivo determinar las relaciones entre las variables cuando no hay o no está totalmente definida la naturaleza de estas relaciones. Las técnicas exploratorias tienen un fuerte componente computacional abarcando desde los métodos estadísticos simples a los más avanzados como las técnicas de exploración de multivariables diseñadas para identificar patrones en conjunto de datos multivariables.

Entre las técnicas estadísticas sencillas se incluyen el estudio de distribuciones de las variables, estudio de correlaciones entre matrices, tablas de contingencias, entre otros. Por su parte, entre las técnicas más complejas se incluyen el Análisis de Factores, el Análisis de Grupos, el Escalado Multidimensional, etcétera.

Redes neuronales (Neural Networks): Las redes neuronales son técnicas analíticas que permiten modelar el proceso de aprendizaje de una forma similar al funcionamiento del cerebro humano, básicamente, la capacidad de aprender a partir de nuevas experiencias. Estas técnicas tuvieron un desarrollo impresionante en la última década, con aplicaciones tanto a la medida como generales (comúnmente llamados Shell) y tienen como objetivo fundamental sustituir la función de un experto humano.

Una de las principales características de las redes neuronales, es que son capaces de trabajar con datos incompletos e incluso paradójicos, que

dependiendo del problema puede resultar una ventaja o un inconveniente. Además esta técnica posee dos formas de aprendizaje: supervisado y no supervisado.

• Análisis Preliminar de datos usando Query tools: es el primer paso de un proyecto de Minería de Datos, se aplica una consulta SQL al conjunto de datos, para rescatar algunos aspectos visibles antes de aplicar las técnicas.
• Técnicas de Visualización: son aptas para ubicar patrones en un conjunto de datos, puede usarse al comienzo de un proceso de Minería de Datos para determinar la calidad de los datos.
• Reglas de Asociación: establecen asociaciones en base a los perfiles de los clientes sobre los cuales se realiza la Minería de Datos.
• Algoritmos Genéticos: son técnicas de optimización que usan procesos tales como combinaciones genéticas y mutaciones, proporcionan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como es el caso de las redes neuronales. Además los algoritmos genéticos son inspirados en el principio de la supervivencia de los más aptos.
• Redes Bayesianas: buscan determinar relaciones causales que expliquen un fenómeno según los datos contenidos en una base de datos. Se han usado principalmente para realizar predicciones.
• Árbol de Decisión: son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos. Los árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación.
• Clustering (Agrupamiento): Agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de minería de datos, dando como resultado un sistema híbrido.
• Segmentación. Consiste en la división de la totalidad de los datos, según determinados criterios. Ejemplo: Dividir los clientes en función de su antigüedad.
• Clasificación. Consiste en definir una serie de clases, donde poder agrupar los diferentes clientes. Ejemplo: definida unas variables de entrada se produce una determinada salida que clasifica al cliente en un grupo o en otro. Por ejemplo, si la edad está entre 20 y 40, está casado y tiene cuenta de ahorro, entonces contrata hipoteca en un 78% de posibilidades.
• Predicción. Consiste en intentar conocer resultados futuros a partir de modelizar los da-tos actuales. Ejemplo: Creamos un modelo de variables para saber si el cliente compra o no compra. Aplicamos el modelo a un futuro cliente, y ya podemos predecir si comprará o no.

Herramientas de Minería de Datos

Las herramientas de minería de datos empleadas en el proceso de extracción de conocimiento se pueden clasificar en dos grandes grupos:

• Técnicas de verificación (en las que el sistema se limita a comprobar hipótesis suministrada por el usuario).
• Método de descubrimiento (en los que se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción.

Exsten algunas herramientas diseñadas para extraer conocimientos desde bases de datos que contienen grandes cantidades de información. Las más populares de estas herramientas son *SPSS Clementine*, *Oracle Data Min*er y *Weka*.

Clmentine de SPSS: Clementine se centra en la integración de *data mining* con otros procesos y sistemas de negocio que ayuden a entregar inteligencia predictiva en un tiempo eficiente durante las operaciones de negocio diarias. La funcionalidad abierta de data mining en bases de datos que posee Clementine permite que muchos de los procesos de data mining se realicen en entornos que mejoran tanto el rendimiento como el despliegue de los resultados de *data mining*. La última versión de Clementine extiende la funcionalidad de data mining al incluir un conjunto de reglas de scoring y modelos de árboles de decisión y carga de resultados de data mining en la base de datos. Sistema integrado de minería de datos que permite encontrar patrones en la información para facilitar la toma de decisiones a los usuarios. Utilizando Clementine se podrá:

• Acceder, preparar e integrar fácilmente datos numéricos, de texto, datos provenientes de páginas Web y de encuestas.
• Construir y validar modelos rápidamente, utilizando las técnicas estadísticas y de aprendizaje automático disponibles más avanzadas.
• Implantar eficientemente los modelos predic-

tivos, en tiempo real o según una programación establecida.
- para las personas que toman decisiones y hacen recomendaciones, y para los sistemas que los utilizan.
- Obtener rápidamente un mejor Retorno de la Inversión y mejores tiempos de respuesta aprovechando las características de rendimiento y escalabilidad.
- Transmitir de forma segura los datos confidenciales a las aplicaciones de data mining en los casos donde la seguridad es crítica.

Esta herramienta permite seleccionar campos o filtrar los datos, permite mostrar propiedades de los datos, encontrar relaciones, ambiente integrado de minería de datos para usuarios finales y desarrolladores. Algoritmos múltiples de minería de datos y herramientas de visualización. Su compañía es SPSS/*Integral Solutions Limited* (ISL). Funciona sobre todas las plataformas hardware y sistemas operativos, incluyendo Unix, VMS y Windows NT. Las organizaciones utilizan el conocimiento extraído con Clementine para:

- retener a los clientes rentables,
- identificar oportunidades de venta cruzada,
- detectar fraudes,
- reducir riesgos y mejorar la prestación de servicios a la administración,
- alcanzar un mayor nivel de conocimiento de sus clientes online, y por lo tanto, mejorar el diseño de sus sitios web.

YALE: Es una herramienta creada en la universidad de Dortmund bastante flexible para el descubrimiento del conocimiento y la minería de datos. Puesto que YALE está escrito enteramente en Java, funciona en las plataformas o sistemas operativos más conocidos. Es un software de código abierto GNU y con licencia GPL. Recientemente fue lanzada la última versión, la cual incluye características como las de implicar nuevos formatos de entrada de datos con operadores para Microsoft Excel y SPSS. Desde la perspectiva de la visualización YALE ofrece representaciones de datos en dispersión en 2D y 3D; representaciones de datos en formato SOM (Self Organizing Map); coordenadas paralelas y grandes posibilidades de transformar las vi-sualizaciones de los datos.

WEKA: Es de libre distribución (licencia GPL) y destacada por la cantidad de algoritmos que presenta así como por la eficiencia de los mismos, por los generadores de reglas, esta desarrollada por miembros de la Universidad de Waikato, ella proporciona gran cantidad de herramientas para la realización de tareas propias de minería de datos, la visualizacion y permite la programación en JAVA de algoritmos más sofisticados para analisis de datos y modelado predicativo, unidos a una interfaz grafica de ususario para acceder facilmente a sus funcionalidades. En ella se implementan las técnicas de clasificación, asociación, agrupamiento, y predicción existentes en la actualidad. Su sistema operativo es multiplataforma. Los puntos fuertes de Weka son:

- Está disponible libremente bajo la licencia pública general de GNU.
- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para reprocesamiento de datos y modelado.
- Es fácil de utilizar por un principiante gracias a su interfaz grafica de usuario.

Weka soporta varias tareas estándar de minería de datos, especialmente, reprocesamiento de datos, clustering, clasificación, regresión, visualización, y selección. Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un fichero plano (*flan file*) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (normalmente numéricos o nominales, aunque también se soportan otros tipos). Weka también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (J*ava Database Connectivity*) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con Weka.

RAMSES: (Sistema de Gestión de Selección y Evaluación de Análisis de Riesgo - *Risk Analysis Management Selection & Evaluation System*): es un programa de gestión de riesgos integrado en el sistema de proceso de datos de Bureau Veritas. El programa recopila todos los datos correspondientes a las operaciones de comercio internacional y está interconectado con la aplicación de minería de datos Angoss® Knowledge Studio. Este software es aceptado como uno de los líderes del mercado en minería de datos y cumple las recomendaciones de la Convención de Kyoto de la OMA (Organización Mundial de Aduanas) de 1999 y del Marco de Normas de la OMC (Organización

Mundial del Comercio). Es utilizado por organismos gubernamentales en el mundo entero. RAMSES ofrece a las autoridades gubernamentales una forma de identificar los embarques de mayor riesgo, facilitando por otro lado la circulación y el despacho de las mercancías de menor riesgo. Interconectado con las bases de datos de Bureau Veritas, RAMSES proporciona una gestión automatizada y digna de confianza de los riesgos inherentes al comercio internacional.
Beneficios:

- Analizar todos los datos del programa de inspección de importaciones.
- Evaluar los niveles de riesgo de las diferentes expediciones de mercancías.
- Favorecer los controles mejor orientados.
- Indicar las medidas a tomar para agilizar el despacho aduanero.
- Se puede aplicar a diferentes tipos de bases de datos.
- Optimizar la asignación de recursos humanos

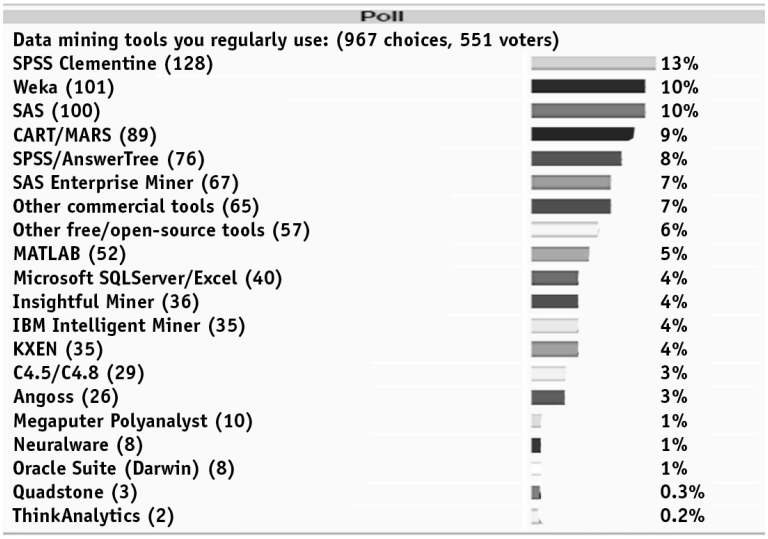
SAS Enterprise Miner: Su compañía es SAS, es una solución de minería de datos que permite incorporar patrones inteligentes a los procesos de marketing, tanto operativos como estratégicos. El software de SAS, es un sistema de entrega de información que provee acceso transparente a cualquier fuente de datos, incluyendo archivos planos, archivos jerárquicos, y los más importantes manejadores de bases de datos relacionales. También incluye su propia base de datos de información para almacenar y manejar los datos, es decir, un "data warehouse". También soporta los principales protocolos de comunicación, cubre los cinco modelos de pro-cesamiento cliente/ servidor de acuerdo a Gartner Group y cumple con las 12 reglas de OLAP. El sistema soporta un amplio rango de aplicaciones, destacándose el análisis estadístico, análisis gráfico de datos, análisis de datos guiado, mejoramiento de la calidad, diseño experimental, administración de proyectos, programación lineal y no lineal, ge-neración de reportes y gráficas, manipulación y despliegue de imágenes, sistemas de información geográfica, visualización multidimensional de datos, aplicaciones de multimedia, así como los sistemas de información ejecutiva.

PolyAnalyst de Megaputer. (Bigus, 1996): Es un sistema de minería de datos premiados de la multiestrategia para descubrir la forma exacta de relaciones funcionales ocultadas en datos. Además de des-cubrir reglas y algoritmos, PolyAnalyst les presenta explícitamente en el una forma simple y fácil de entender. En la fundación de PolyAnalyst tiene un lenguaje de programación interno universal capaz de expresar reglas y algoritmos arbitrarios.
Su compañía es Megaputer líder en negocios y software inteligentes para web. Ofrece las mejores herramientas para *data mining*, *text mining* y *web mining*. Plataformas:

- Microsoft Windows XP/NT/2000.
- Para UNIX y Linux 2001.
- Además requiere la instalación de Microsoft Excel.

Otras herramientas de libre distribución
R: herramienta excelente para el análisis de datos basada en el conocido programa estadístico S-Plus y con un manejo de las mátrices y variables equivalentes a MATLAB. Es muy útil para el análisis estadístico, transformación y manipulación de los datos. Destacar la excelente asesoría tècnica llevada a cabo principalmente por algunos de los principales profesores e investigadores en estadística del mundo.
XELOPES: Otra librería de libre distribución con cantidad de funciones para minería de datos. Permite la implementación en JAVA o C++.
SNNS: Aplicación para el desarrollo, entrenamiento y testeo de múltitud de tipos diferentes de redes neuronales. Muy útil para desarrollar clasificadores sofisticados y modelos basados en redes neuronales.
XmdvTool, Xgobi, IBM-OpenDX, Visipoint: Otras herramientas con licencia GPL que tienen diferentes funciones de visualización muy útiles para encontrar patrones ocultos en los datos.
En la Figura se puede apreciar una encuesta hecha en el conocido portal sobre Minería de Datos y gestión del conocimiento, donde se da una idea de las aplicaciones que más utilizan los profesionales y las múltiples aplicaciones que existen en el mercado. Aquí se destacan programas de familias de aplicaciones estadísticas ejemplo: SAS(SAS, SAS EnterpriseMiner) o SPSS(SPPS Clementine, SPSS AnswerTree), estas contrastan con otras desarrolladas integramente en el campo de la Minería de Datos ejemplo: CART/MARS, IBM-I-Miner, Angoss, Megaputer PolyAnalyst, KXEN estas abarcan principalmente mètodos estadísticos y de visualización combinados con algoritmos mas propios de Minería de Datos. El grado de eficiencia de cada herramineta depende de múltiples factores: tipos de algoritmos, funciones de tratamiento de la información, eficiencia de los algoritmos, generadores de informes, formas de pasar la información. Estas herramientas aportan múltiples ventajas para los campos de investi-

Figura 1. Herramientas de Minería de Datos usadas habitualmente (KDnuggets, 2002).



gación y docencia en el aprendizaje y desarrollo de la Minería de Datos, nos han demostrado que tienen grandes ventajas.

¿Por qué usar Minería de Datos?

Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Contribuye a la toma de decisiones tácticas y estratégicas.

Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de una mejor forma.

Genera modelos descriptivos: permite a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.

Genera modelos predictivos: permite que relaciones no descubiertas a través del proceso de la Minería de Datos sean expresadas como reglas de negocio.

Conclusiones

La Minería de Datos se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos y descubrir patrones que ayuden a la identificación de estructuras en los datos.

Los productos a comercializar son costosos y requieren de mucha experiencia para su utilización. Es muy fácil hallar patrones equívocos o no interesantes.

La aplicación de estas herramientas ayuda en el proceso de toma de decisión de las organizaciones.

Referencias

Berry, M.J. y G. Linoff, Data Mining Techniques For Marketing, Sales and Customer Support. 1997.

Bigus, JP. Data Mining with Neural Networks" 1996. Disponible en: http://www.megaputer.com

Delve Projects. Data for Evaluating Learning in Valid Experiments. Disponible en: http://www.cs.utoronto.ca/~delve/index.html

Hand, D., H. Mannila, and P. Smyth, Principles of Data Mining. London: The MIT Press., 2001.

Gonzalez, P.P., Desarrollo de tecnicas de mineria de datos en procesos industriales: Modelizacion en lineas de produccion de acero. Julio de 2006: Universidad de la Rioja.

KDnuggets. Data mining tools you regularly use. junio 2002 . Disponible en: http://www.kdnuggets.com/polls/2002/data_mining_tools.htm

Machine Learning Group at University of Waikato Data Mining Software in Java. Disponible en:http://www.cs.waikato.ac.nz/ml/weka/

Piatetsky-Shapiro, G. y W. J Frawley. Knowledge Discovery in Databases". AAAI/MIT Press, 1991.

Siebes, A., Data Mining and Statistics. 2000.

Thuraisingham, B. Data Mining. Technologies, Techniques, Tools and Trends CRC Press LLC, 1999.

Wang, X.Z., Data Mining and Knowledge Discovery For Process Monitoring and Control. 1999, London: Ed. Springer.

Witten, I.H. y E. Frank, Data Mining. Practical Machine Lerning Tools and Techniques with Java Implementations. 2000: San Francisco, California.