



San Marcos

MIEMBRO DE LA RED
ILUMNO

PRE PROCESAMIENTO DE DATOS



San Marcos

MIEMBRO DE LA RED
ILUMNO

PRE PROCESAMIENTO DE DATOS

En el preprocesamiento de los datos se corrigen datos erróneos o incompletos, se elimina el posible ruido que contengan los datos, etc.

En el postprocesamiento, a partir de los resultados de la minería de datos, se pueden preparar patrones y sacar conclusiones de los que se pueda extraer conocimiento, que es el objetivo final del proceso de KDD.

El proceso más importante es el de **Minería de Datos o DM (Data Mining)**. Una definición formal de la DM sería: la minería de datos es el proceso automático para el descubrimiento de información útil en grandes cantidades de datos. Este proceso es un campo multidisciplinario, en el que se pretende predecir resultados y/o descubrir relaciones entre los diferentes datos. Las diferentes tareas que puede realizar la DM son:

- **Clasificación:** mediante la clasificación se busca encontrar un modelo que pueda predecir el comportamiento de una variable a partir de sus características.
- **Análisis de asociaciones:** estas técnicas pretenden sacar patrones de las relaciones que hay entre diferentes rasgos de los datos.
- **Detección de anomalías:** el objetivo de la detección de anomalías es encontrar aquellos elementos o características que son significativamente diferentes del resto de los datos. (Ramírez y Ferri, 2004, p. 128)





LIMPIEZA DE DATOS

Este proceso consiste en llenar los valores faltantes dentro de la base de datos y la información recolectada, suavizando los datos erróneos, identificar o remover los datos

inconsistentes. La importancia de la limpieza de los datos es reconocida en diversa literatura como una de las tareas más importantes y exhaustiva del preprocesamiento.

ESTE PROCESO CONSISTE EN LLENAR LOS VALORES FALTANTES DENTRO DE LA BASE DE DATOS Y LA INFORMACIÓN RECOLECTADA, SUAVIZANDO LOS DATOS ERRÓNEOS, IDENTIFICAR O REMOVER LOS DATOS INCONSISTENTES.

Los datos, si bien corresponden a la misma información recolectada en la investigación del mercado, presentan

formatos diferentes en los diferentes períodos de tiempo de la investigación, debido a modificaciones del instrumento de recolección (formulario), de manera que deberán ser sometidos a un proceso de integración y unificación de conceptos.

Dentro del reconocimiento y limpieza de datos se tiene que en este paso tiene como objetivo reducir el ruido y las inconsistencias. Para ello, se seleccionará un resumen (muestra) de los datos, pudiendo de esta forma interpretar la validez de algún valor para algún atributo y mejorar la calidad de los datos.

En el presente trabajo resulta interesante estudiar la cantidad de datos nulos y los outliers (datos erróneos), con el fin de reducir las posibilidades de error cuando sean procesados por el algoritmo de minería. Se aplicarán y analizarán las técnicas que facilitan el reconocimiento y limpieza de los datos que provee la herramienta seleccionada (Weka), siendo algunas de estas las que se explican brevemente a continuación. (*Knowledge Discovery and Data Mining*, 1996, p. 123).

- **Tablas de resumen de atributos:** en Weka denominada Relación actual, es una relación (tabla) que contiene todos los datos necesarios, considerados relevantes para el estudio.
- **Resúmenes de estadísticas:** la herramienta proporciona un resumen de las estadísticas para el atributo que se está evaluando: porcentaje de instancias con valores nulos, cantidad de diferentes valores para el atributo, cantidad de instancias que tienen un valor único para ese atributo (diferente de las demás instancias).
- **Diagramas:** se emplea un diagrama de barras segmentadas para poder visualizar cada valor distinto del atributo y la cantidad de instancias que presentan este valor.



TÉCNICAS DE SELECCIÓN DE ATRIBUTOS

La técnica usada para realizar estas hazañas en *Data Mining* se llama modelado. Modelado es el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de *Data Mining*. Lo que ocurre en las computadoras no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de

Data Mining en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

La primera cosa que puede probar, para ver si es un buen modelo, es pedirle que aplique el modelo a su base de clientes. Con *Data Mining*, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de *Data Mining*. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

Dentro de la clasificación, quizá la tarea más extendida, tenemos varios algoritmos que se pueden utilizar:

1 ÁRBOLES DE DECISIÓN

2 BASADOS EN REGLAS

3 REDES NEURONALES

4 BASADAS EN CLUSTERS

5 REDES BAYESIANAS

UNA ARQUITECTURA PARA DATA MINING

Para aplicar mejor estas técnicas avanzadas, estas deben estar totalmente integradas con el *Data Warehouse* así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de *Data Mining* actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde *Data Mining*. El *Data Warehouse* analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, otros.

DISCRETIZACIÓN DE ATRIBUTOS

Es un paso más en el preprocesamiento de datos. Los atributos continuos son previamente discretizados antes de la tarea de clasificación. Son parte de las arquitecturas de información DSS (*Decision Support System*) y sistemas de información diseñados especialmente para la realización de consultas y reportes de datos históricos, estos son comunes en toda la organización, la información puede estar sumariada y/o detallada. Los datos que alimentan los sistemas de información del sistema de datos de cada empresa.

Data Warehouse los proporcionan los sistemas operacionales, también denominados sistemas de ejecución de procesos de negocios.

El *Data Warehouse* es un repositorio de datos de muy fácil acceso, alimentado por numerosas fuentes (llamadas grupos de información) sobre temas específicos del negocio, es un almacén o bodega de datos con información depurada que los ejecutivos extraen y consultan. El *Data Warehouse* casi siempre es un sistema de datos guardados físicamente en estructuras separadas, los datos son transformados a partir de los datos proporcionados por los sistemas operacionales. El potencial de la explotación de los datos (*Data Mining*) puede realizarse si de manera apropiada los datos son recolectados y guardados en un almacén o depósito de datos (*Data Warehouse*). Un almacén de datos o *Data Warehouse* es un sistema de gestión de bases de datos relacional diseñado específicamente para ofrecer las necesidades de los sistemas de análisis de la información que facilita la integración y homogeneización de los datos de toda la empresa. En otras palabras, el almacén de datos provee datos que ya han sido transformados y sumariados, por lo tanto crea el entorno apropiado para un uso más eficiente de las herramientas DSS y EIS.

La estructura básica de la arquitectura *Data Warehouse* (DW) incluye:

- **Datos operacionales:** un origen de datos para el componente de almacenamiento físico DW.
- **Extracción de datos:** selección sistemática de datos operacionales usados para poblar el componente de almacenamiento físico DW.
- **Transformación de datos:** procesos para sumarizar y realizar otros cambios en los datos operacionales para reunir los objetivos de orientación a temas e integración principalmente.
- **Carga de datos:** inserción sistemática de datos en el componente de almacenamiento físico DW.
- **Datawarehouse:** almacenamiento físico de datos de la arquitectura DW.
- **Herramientas de acceso al componente de almacenamiento físico DW:** herramientas que proveen acceso a los datos.





COMPONENTES PRINCIPALES

Hay cuatro tipos de componentes que normalmente se involucran en la minería de datos:

- **Clasificación** - la tarea de generalizar una estructura familiar para utilizarla en los nuevos datos.
- **Agrupamiento** - la tarea de encontrar grupos y estructuras en los datos que son de alguna manera u otra lo mismo, sin necesidad de utilizar las estructuras observadas en los datos.
- **Aprendizaje de reglas de asociación** - busca relaciones entre las variables.
- **Regresión** - su objetivo es encontrar una función que modele los datos con el menor error.

La forma en la cual se estructure el almacenamiento de datos *Data Mining* dentro de los componentes genera una clasificación respecto a la forma de implementar una arquitectura de la base de datos. La estructura adoptada para el almacén de datos se debe realizar de la manera que mejor satisfaga las necesidades empresariales, siendo entonces dicha elección factor clave en la efectividad del proceso. Las más básicas son:

- **EL DW central:** es una implementación de un solo nivel con un solo almacén para soportar los requerimientos de información de toda la empresa.
- **El DW distribuido:** es también una estructura de un nivel, pero que particiona el almacén para distribuirlo a nivel departamental.
- **El DW de dos niveles:** combina ideas de los dos anteriores, siendo sus distribuciones tanto el almacén empresarial como los departamentales. (*Knowledge Discovery and Data Mining, 1996, p. 124*).

CARACTERÍSTICAS

Existen generalmente cuatro características que describen un almacén de datos:

1. Orientado al sujeto:

Los datos se organizan de acuerdo al sujeto en vez de la aplicación, por ejemplo, una compañía de seguros usando un almacén de datos podría organizar sus datos por cliente, premios, y reclamaciones, en lugar de por diferentes productos (automóviles, vida, etc.). Los datos organizados por sujetos contienen solo la información necesaria para los procesos de soporte para la toma de decisiones.

2. Integrados:

Cuando los datos residen en muchas aplicaciones separados por los distintos entornos operacionales, la decodificación de los datos es a menudo inconsistente. Por ejemplo, en una aplicación, la palabra gender podría codificarse como "m" y "f" en otra como "0" y "1". Cuando los datos fluyen de un entorno operacional a un entorno de almacén de datos o de data warehouse, ellos asumen una codificación consistente, por ejemplo gender siempre se transformaría a "m" y "f". (Benzecri, 2015, p. 221)

3. Variación-temporal:

El almacén de datos contiene un lugar para guardar datos con una antigüedad de 5 a diez años, o incluso más antiguos, para poder ser usados en comparaciones, tendencias y previsiones. Estos datos no se modificarán.

4. No son inestables:

Los datos no serán modificados o cambiados de ninguna manera una vez ellos han sido introducidos en el almacén de datos, solamente podrán ser cargados, leídos y/o accedidos.

Si la organización tiene muchos datos de aplicaciones tradicionales y está buscando una solución para transferir grandes volúmenes de datos de un Mainframe, se necesita una solución de bodega de datos de fuerza industrial para hacer transferencia bruta de datos diferentes de fuentes en Mainframes a bodegas de datos en DB2 o en Unix.

Distribución de datos es el proceso de mover los datos extraídos y trasladarlos a la bodega de datos o a diferentes bases de datos en cualquier plataforma en cualquier sitio. Una herramienta de distribución define base de datos objetivo, información de conversión y entrada/salida de datos. Una vez creadas estas definiciones, pueden ser salvadas para ser reutilizadas, editadas o ejecutadas posteriormente. (*Knowledge Discovery and Data Mining*, 1996, p. 234).



MANEJO DE LOS DATOS

El repositorio sirve como un sitio para almacenar los datos de los activos de información de una organización. Abarca todos los datos de la organización, sin importar cuál es la fuente original y facilita el entendimiento de toda la empresa y controla la existencia de los recursos de datos existentes.

El repositorio sirve como una guía para definir un ambiente de migración de datos y contiene:

- El mapeo entre las fuentes y la bodega de datos objetivo
- Requerimientos de traslado de la información
- Reglas de negocio
- Pistas de auditoría
- Otros criterios de selección para construir la bodega de datos

De acuerdo con Frappaolo (2006), existen facilidades para lograr que los usuarios no

técnicos puedan acceder la información a nivel de metadatos con herramientas intuitivas, fáciles de usar y de hacer análisis del tipo *Drill - Down* de tópicos generales a categorías más específicas de metadatos predeterminados. Este tipo de herramientas proveen al usuario final de una cabina virtual de datos que les permite a través del mouse ir a cajones gráficos, a los folders dentro de los cajones y páginas de apuntes rotuladas.

ABARCA TODOS LOS DATOS DE LA ORGANIZACIÓN, SIN IMPORTAR CUAL ES LA FUENTE ORIGINAL Y FACILITA EL ENTENDIMIENTO DE TODA LA EMPRESA Y CONTROLA LA EXISTENCIA DE LOS RECURSOS DE DATOS EXISTENTES

ACCESO Y ANÁLISIS DE DATOS

Una vez que la bodega de datos se ha llenado de información, los usuarios finales pueden acceder y analizar los datos. Para satisfacer las necesidades de usuarios finales en cualquier plataforma se provee de algunas herramientas especializadas para hacer reportes y queries, para desarrolladores de aplicaciones de oficina y usuarios que necesitan revisar datos sumariados de la bodega así como crecientes niveles de detalle (*Knowledge Discovery and Data Mining*, 1996, p. 232).

MANEJO DE SISTEMAS

La base de datos de la bodega debe ser frecuentemente mantenida y manejada por DBA's para reducir el impacto en el desempeño del sistema y recursos. Para ser eficiente y productivo, el proceso de bodega de datos debe ser automatizado dentro de un ambiente de producción. Las herramientas necesarias para su mantenimiento se clasifican en:

- 1 HERRAMIENTAS DE MANEJOS DE BASES DE DATOS**
- 2 SISTEMAS PARA MANEJO DE PROCESOS DE JOBS**
- 3 RESOLUCIÓN DE PROBLEMAS (*HELP DESK*)**
- 4 SEGURIDAD**
- 5 DISTRIBUCIÓN (*KNOWLEDGE DISCOVERY AND DATA MINING*, 1996 P. 234)**

CONSTRUCCIÓN DE LA BODEGA DE DATOS

Para construir una bodega de datos es necesario saber algunos temas generales, que se agrupan en los siguientes tópicos:

1. AMBIENTE ACTUAL:

Es necesario entender el negocio y el ambiente tecnológico actual de la organización para construir una bodega que sirva. Cualquier solución propuesta de bodega de datos debe estar muy orientada por las necesidades del negocio y debe ser compatible con la arquitectura técnica existente y planeada de la compañía.

2. AMBIENTE DE NEGOCIOS:

Es indispensable tener el conocimiento exacto sobre el tipo de negocios de la Organización y el soporte que representa la información dentro de todo su proceso de toma de decisiones.

3. AMBIENTE TÉCNICO:

Se debe tener un claro concepto desde una perspectiva técnica de los sistemas de información de la organización. En este análisis se debe tener claridad del ambiente técnico actual y futuro a nivel de detalle. Se debe incluir tanto el aspecto de ambiente *hardware*: *mainframes*, servidores, redes, así como aplicativos y herramientas.

4. ESPECTATIVAS DE LOS USUARIOS:

Esto es determinante en el éxito de un proyecto de bodega de datos puesto que este no es un proyecto tecnológico, es una forma de vida de las organizaciones y como tal, tiene que contar con el apoyo de todos los usuarios y su convencimiento sobre su bondad. (*Knowledge Discovery and Data Mining*, 1996, p. 234).



5. OTROS COMPONENTES:

RapidMiner, antes llamado YALE (Sin embargo, otro ambiente de aprendizaje), es un ambiente de experimentos en aprendizaje automático y minería de datos que se utiliza para tareas de minería de datos tanto en investigación como en el mundo real. Permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML y se hacen con la interfaz gráfica de usuario de *RapidMiner*. *RapidMiner* ofrece más de 500 operadores para todos los principales procedimientos de máquina de

aprendizaje, y también combina esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka. Está disponible como una herramienta stand-alone para el análisis de datos y como motor para minería de datos que puede integrarse en tus propios productos. (Maoe, 2016, párr. 4)

EN LUGAR DE CREAR EL "GRAN REPOSITORIO" DE LA EMPRESA, SE HA RECOMENDADO EL DESARROLLO DE PROYECTOS MÁS PEQUEÑOS, CREANDO ASÍ CUBOS DE INFORMACIÓN QUE RESPONDAN A LAS NECESIDADES DE UN ÁREA ESPECÍFICA.

aprendizaje, y también combina esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka. Está disponible como una herramienta stand-alone para el análisis de datos y como motor para minería de datos que puede integrarse en tus propios productos. (Maoe, 2016, párr. 4)

DATAMART.

Un *Datamart* puede verse como una bodega entre un gran almacén de datos (*Datawarehouse*) que almacena data para un propósito específico. El concepto *Datamart* es una extensión natural de *Datawarehouse*. Cuando se habla de *Datamart*, se está refiriendo, en pocas palabras, a un *Datawarehouse* en pequeño y concentrado en una sola unidad del negocio, en lugar de en toda una corporación. (Hernández, Ramírez y Perri, 2004, p. 235)

Surge a partir de la necesidad de abordar los proyectos de *Datawarehouse* por áreas temáticas de análisis (esto es por la gran complejidad que alcanzan a tener los proyectos de *Datawarehouse*). En lugar de crear el "gran repositorio" de la empresa, se ha recomendado el desarrollo de proyectos más pequeños, creando así cubos de información que respondan a las necesidades de un área específica. Su función principal es apoyar a otros sistemas para la toma de decisiones, además de disponer de la información solicitada en el momento preciso, permitiendo así un mejor control de la información que se está abarcando.





San Marcos

MIEMBRO DE LA RED
ILUMNO

Entre los diferentes tipos de *Datamarts* prevalecen los *Datamarts* móviles. Al hablar de *Datamarts* móviles se trata de bodegas personales que pueden cargarse en una laptop y llevarse a donde quiera que sea.

Para el proceso de construcción de *Datamarts* existen dos enfoques: el primero es

HAY DIFERENTES ESTILOS DE USUARIOS FINALES CADA UNO CON SU PROPIO NIVEL DE CONOCIMIENTO Y NECESIDADES, PARA ASÍ PROVEER DE APROPIADOS MECANISMOS DE ACCESO PARA CADA CLASE DE USUARIOS.

construir primero un núcleo de la bodega de datos y luego hacer varios *Datamarts* sobre dicha bodega. La segunda es construir primero un *Datamart* e ir expandiendo poco a poco la bodega de datos y añadiendo nuevos *Datamarts* conforme sea necesario hacerlo.

Los *Datamarts* deben consistir en datos extraídos del corazón de la bodega de datos y reorganizados y/o reformateados para hacer más fácil su

uso para diferentes propósitos. Pero dado que esos propósitos específicos pueden cambiar en el tiempo, los *Datamarts* deben ser concebidos con estructuras de datos temporales. Cuando los usuarios no ven más los datos como están presentados por un *Datamart* en particular, este *Datamart* debe ser removido. Y mientras los usuarios desarrollan nuevas formas de hacer búsquedas y mirar los datos, deben ser creados nuevos *Datamarts* para hacer sus búsquedas más simples y con un mejor desempeño.

Los *Datamart* pueden incluir una gran variedad de estilos de tablas. Algunas pueden ser simplemente un subconjunto de datos en la bodega de datos, que contienen solamente datos para una particular zona geográfica, un período específico de tiempo, una unidad de negocios. Otros *Datamarts* pueden ser el resultado de reunir información proveniente de diferentes tablas del corazón de la bodega de datos en una tabla *Datamart* desnormalizada. O tal vez los *Datamarts* serán construidos para contener elementos de datos calculados y derivados que no están explícitamente almacenados en el corazón de la bodega de datos.

También es posible mencionar que el uso de estructuras de datos multidimensionales debería estar reservado para *Datamarts*. Esto corresponde a datos que están en el corazón de la bodega de datos deberían almacenarse en forma relacional y luego ser extraídos en un *Datamart* multidimensional si es requerido.

Hay diferentes estilos de usuarios finales cada uno con su propio nivel de conocimiento y necesidades, para así proveer de apropiados mecanismos de acceso para cada clase de usuarios.



DIFERENCIAS CON BASES DE DATOS TRADICIONALES

La primera diferencia es que los primeros sistemas de bases de datos tenían un propósito transaccional, en un *Datamart* se manejan grandes cantidades de datos, y se enfocan a contestar preguntas del tipo "Que pasaría si..." Por lo que requieren mayor flexibilidad. Una base de datos tradicional se optimiza para transacciones y un *Datamart* se optimiza para queries con grandes volúmenes de datos.

Orange es una suite de software para minería de base de datos y aprendizaje automático basado en componentes que cuenta con un fácil y potente, rápido y versátil front-end de programación visual para el análisis exploratorio de datos y visualización, y librerías para Python y secuencias de comando. Contiene un completo juego de componentes para preprocesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo, y técnicas de exploración. Está escrito en C++ y Python, y su interfaz gráfica de usuario se basa en la plataforma cruzada del framework Qt. (Maoe, 2016, párr. 4)



REFERENCIAS BIBLIOGRÁFICAS

- Alegret Rodríguez, M., Herrera, M., Grau Abalo, R. (2008). Las técnicas de estadística espacial en la investigación salubrista. *Rev Cubana Sal Públ* [revista en la Internet]. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400003&lng=es
- Barcellos, C y Buzai, G. (2006). La dimensión espacial de las desigualdades sociales en salud: aspectos de su evolución conceptual y metodológica. México: Universidad Nacional de Luján: Anuario de la División Geografía.
- Benzecri, J. (2015). Gestión del conocimiento y minería de datos. PHC, París
- Berry, M. y Linoff, G. (2014). *Data Mining Techniques for Marketing Sales and Customer Support*. USA: John Wiley & Sons.
- Chatfield, C. y Collins, A.J. (1999). *Introduction to multivariate analysis*. London: Chapman and Hall.
- Kamber M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann;
- Hernández C (2004). USA: Centricity Solution for Marketin Automation.
- Hoaglin D.C., Mosteller F., Tukey J.W. (2005). N.Y: *Exploring Data tables. Trends and Shapes*. Wiley.
- Ester, M., Kriegel, H., y Sander, J (1999). . Knowledge discovery in spatial databases. K1-99. *Advanc Artif Intellig*.
- Jambu, M. (2000). *Classification Automatique pour l'Analyse des données*. París: Dunod.
- Johnson Dallas, E. (2013). *Métodos multivariados aplicados al análisis de datos*. México: Thomson editores.
- Johnson, R.A. y Wichern Dean, W. (2010). *Applied Multivariare Statistical Analysis*. 3rd De. USA: Prentice Hall Inc.

- Lebart, L., Morineau, A. y Tabard, N. (2000). *Téchniques de la description statistique*. París: Dunod.
- Lebart, L., Morineau, A. y Piron, M. (2005). *Statistique exploratoire multidimensionnelle*. París: Dunod.
- Macoe, J. (2016). 5 de los mejores software de minería de datos de Código Libre y Abierto. Recuperado de:
[Http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/](http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/)
- Méndez, D. (2011). *Estrategia para la Gestión Empresarial*. McGraw Hill. México:
- Pla, L. (2006). *Análisis Multivariado: Método de Componentes Principales*. Monografía N° 27 Serie de matemática. USA: O.E.A.Washington.
- Tecnologías-información.com. (2015). *Minería de Datos*. Recuperado de: <http://www.tecnologias-informacion.com/mineria-de-datos.html>
- Vinnakota, S. y Lam, N. (2006). Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. *Internat J Heal Geogr*.
- Zhao, F., Zhu, R., Zhang, L., Zhang, Z, Li, Y. y He. M. (2011). Application of satscan in detection of schistosomiasis clusters in marshland and lake areas.

