Universidad Tecnológica Nacional, Facultad Regional Rosario
Departamento de Ingeniería Química
Grupo de Investigación Aplicada a la ingeniería Química (GIAIQ)

Informática Aplicada a la Ingeniería de Procesos I (Orientación I)

Profesores:

Carlos Alberto Ruiz
Marta Susana Basualdo

Tema:

Data Mining

Autor:

Sonia Pighin

16 de Abril de 2001

CONTENIDOS

1. ¿Qué es Data Mining?

Data Mining es la extracción de información en grandes *Bases de Datos*. Su nombre deriva de la analogía existente entre buscar dicha información valiosa y minar una montaña para encontrar un yacimiento de metales preciosos, ya que ambos procesos requieren examinar una inmensa cantidad de material o investigar inteligentemente hasta concretar la búsqueda.

Es un conjunto de tecnologías que ayuda a las empresas a enfocar sus objetivos sobre la información más importante de sus *Fuentes de Datos*.

Las herramientas de *Data Mining* pueden responder preguntas que generalmente demandan demasiado tiempo, encontrando información que ni un profesional experto podría hallar porque se encuentra fuera de sus expectativas.

Data Mining es en realidad, un proceso iterativo de descubrimiento de patrones y tendencias dentro de los datos, a través de métodos automáticos, manuales o más generalmente semiautomáticos, y que no serían necesariamente revelados por otros métodos tradicionales de análisis.

Las herramientas de *Data Mining* exploran las *Bases de Datos* en busca de patrones ocultos, permitiendo a partir de éstos predecir las futuras tendencias y comportamientos de información nueva.

Como es de esperar, los patrones descubiertos deben ser significativos en el hecho que conduzcan a alguna ventaja y ésta generalmente es económica.

2. ¿Por qué se hace necesario el *Data Mining*?

- Porque existen extensos volúmenes de datos almacenados en *Fuentes de Información*, los cuales se acumulan bajo la creencia que alguien, en algún momento los utilizará. Sin embargo, crece progresivamente la diferencia entre *Generación* de datos y *Entendimiento* de éstos: como el volumen de datos aumenta, el número de personas que entienden estos datos desafortunadamente disminuye.

- Porque la información oculta en los datos es útil y generalmente no se encuentra en forma explícita para tomar ventaja de ésta.

- Porque en algunos casos, los datos no se pueden analizar por métodos estadísticos estándar, porque pueden existir valores perdidos o bien, los datos pueden estar en forma cualitativa y no cuantitativa.

- Porque en ciertas situaciones, el acceso a los datos no es sencillo.

- Las técnicas de *Data Mining* también se hacen necesarias por el desarrollo actual de *Almacenes de Datos (Data Warehouse)* a gran escala, que son los sistemas utilizados para el almacenamiento y distribución de cantidades masivas de datos.

3. ¿Qué no es *Data Mining*?

El proceso de *Data Mining* tienen como objetivo fundamental descubrir patrones y tendencias en complejas *Fuentes de Datos*. Una vez, que se identifica un patrón particular, el proceso de descubrimiento finaliza y este patrón se convierte en un patrón conocido.

Por lo tanto, *Data Mining* no comprende aquellas aproximaciones analíticas que buscan set de datos a partir de patrones conocidos, así mismo, las técnicas que requieren *Implementación de Reglas, Casos de Entrenamiento Preestablecidos o Aprendizaje Automático Supervisado* son útiles pero no constituyen el proceso de *Data Mining*.

3.1 Aprendizaje Automático vs. *Data Mining*

El *Aprendizaje* puede definirse como:

"un ser aprende, cuando cambia sus comportamientos en la forma que logre la mejor performance en el futuro"

Este concepto une al *Aprendizaje* más con *Performance* que con *Conocimiento*, es decir que uno puede evaluar lo aprendido al observar el comportamiento actual y compararlo con el comportamiento anterior.

Por otro lado, el *Aprendizaje* es muy diferente del *Entrenamiento*, porque el *Aprendizaje* implica pensar, implica tener propósitos, implica tener intenciones. El *Aprendizaje* sin objetivos es *Entrenamiento* propiamente dicho.

Por último, el *Aprendizaje* puede resumirse como:

"Adquisición del Conocimiento y habilidad de usarlo ventajosamente".

Data Mining abarca, más bien, un sentido práctico y no teórico combinando la intervención humana con técnicas de *Aprendizaje Automático*. Es una herramienta que ayuda a interpretar los datos, permite encontrar patrones y hacer predicciones a partir de éstos.

3.2 Análisis vs. Monitoreo

(i) En el análisis, se recogen los datos y éstos se mantienen constantes en el tiempo.

(ii) El análisis, generalmente, no se ejecuta en *Tiempo Real* y los patrones se descubren *Off-line*.

(iii) Como los datos no varían con el tiempo durante el análisis, la toma de decisiones puede no ser inmediata.

(iv) El proceso de *Data Mining*, se orienta al *Análisis* de los datos, porque éste se mantiene dentro del campo de los descubrimientos.

(i) En el monitoreo, en cambio, los datos no se mantienen constantes, sino más bien son comparados continuamente con un set de condiciones o límites.

(ii) Sucede en *Tiempo Real* abarcando operaciones automáticas *On-line*.

(iii) Los datos se procesan en forma permanente, siendo la toma de decisiones aproximadamente inmediata.

(iv) En el monitoreo no se hallan nuevos patrones, éstos ya han sido identificados por dichos análisis previos, y el monitoreo se ocupa, en realidad, de detectar posibles violaciones a tales patrones.

Cabe aclarar, que algunas herramientas de *Data Mining* permiten integrar datos históricos con entradas de datos en *Tiempo Real*, dando como resultado la posibilidad de ejecutar *Data Mining* en *Tiempo Real.*

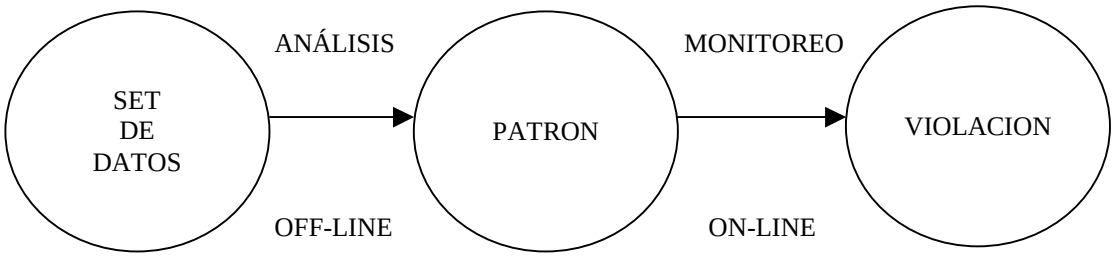

Figura 1. Análisis vs. Monitoreo

4. ¿Qué debe considerarse previamente para aplicar *Data Mining*?

- No perder dinero: Que el costo de implementación sea menor que las mejoras obtenidas, produciéndose así un retorno de la inversión.

- Obtención rápida de resultados: Generalmente, se espera obtener resultados dentro de un período de tiempo razonable. Si éste se hace muy extenso debería retornarse al inicio y realizar los cambios que se consideren necesarios.

- Acceso a los datos: No es imprescindible un acceso *On-line* a las *Fuentes de Datos*, ya que *Data Mining* no se realiza en *Tiempo Real*, pero sí se hace necesario el acceso a toda la información para la ejecución del análisis.

- Empleo del sistema: Generalmente, nunca se utiliza una simple aplicación, sino más bien una combinación de técnicas y metodologías.

- Toma de decisiones: Como regla general, nunca una herramienta sólo proporcionará la solución buscada, sino simplemente ayuda a encontrarla, como tal lo indica la palabra herramienta. Encontrar la solución al problema propuesto y la total responsabilidad de la toma de decisiones se deposita sobre el o los profesionales que realizan el análisis.

5. ¿Dónde Aplicar *Data Mining*?

Actividades Gubernamentales
Bancos y Finanzas
Computación
Industria Automotriz
Manufactura
Química y Farmacéutica
Salud
Seguridad / Investigación
Servicios Financieros
Servicios Públicos
Telecomunicaciones / Media
Ventas por menor (Retail) / Distribución

5.1 ¿Qué Industrias utilizan *Data Mining*?

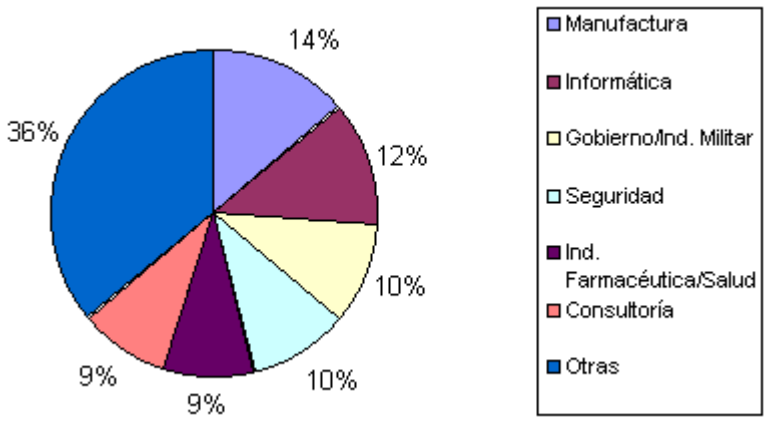

Figura 2. Campos de Aplicación.

6. ¿Cómo es la entrada al proceso de *Data Mining*?

Los datos como tales no puede ser manejados directamente por *Data Mining*, sino es necesario modelarlos llevándolos a un formato tal que sí pueda ser empleado, y el desarrollo de dicho modelo es decisivo, ya que determina los tipos de resultados que se pueden obtener.

Este modelado de datos, asume una estructura orientada a objetos, donde la información está representada por objetos, sus atributos descriptivos y las relaciones entre las clases de objetos.

6.1 Clases de Objetos

- Las clases de objetos se consideran como categorías conceptuales. Por ejemplo: personas, lugares, direcciones, y más.

- La decisión de designar a un campo de variables como una clase de objeto es completamente arbitraria, pero críticamente importante.

- Para escoger cuales variables serán clases de objetos, generalmente se prefiere entidades tangibles, pero pueden ser, en cambio, abstractas e incluir estados, fechas, valores, actividades y más.

Ejemplo 1:

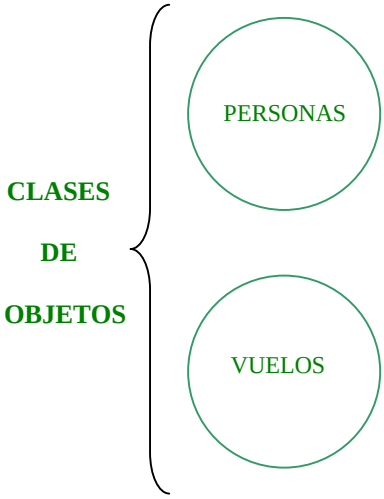

Figura 3. Clases de Objetos

6.2 Objetos

- Los objetos dentro de una clase son las distintas entidades que ésta contiene.

Ejemplo 2:

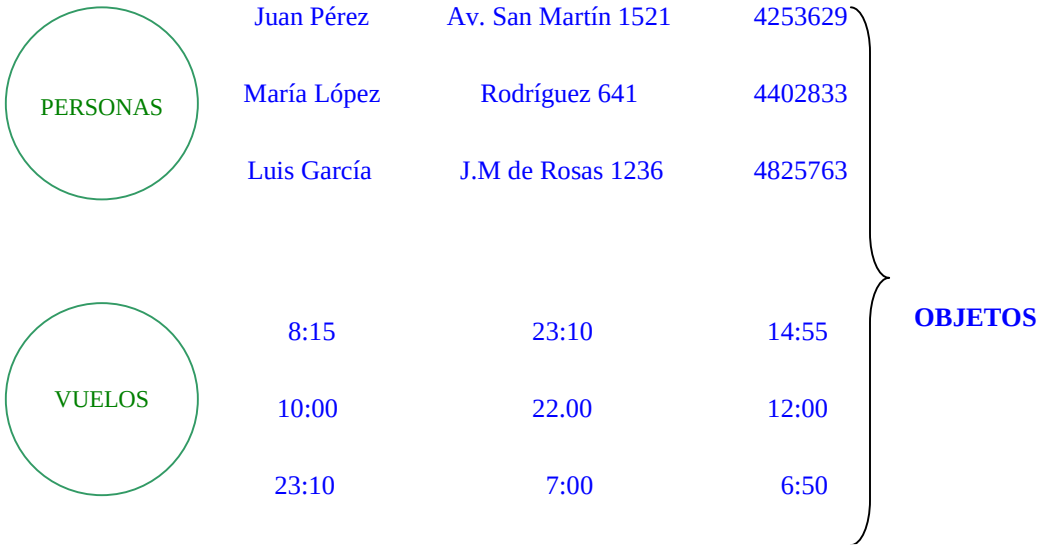

Figura 4. Objetos.

6.3 Atributos

- Los atributos definen los comportamientos de una clase de objeto ya determinada como tal.

- También ayudan a diferenciar los objetos individuales dentro de una clase.

- El valor de un atributo es una medida de la cantidad que dicho atributo se refiere a un objeto.

- Para caracterizar un objeto cualquiera, cada atributo deber tener un único valor. Si en el set de datos hay más de un valor, no se puede determinar cual es el correcto. Esto generalmente ocurre cuando dichos valores se toman en distintos intervalos de tiempo.

- Para estos atributos que cambian con el tiempo, debe plantearse un modelo diferente y esta situación se conoce como *"análisis basado en estados" (state-based).*

Ejemplo 3:

Figura 5. Atributos

- Existen casos donde ciertas variables pueden cumplir distintas funciones dentro del modelo, por ejemplo, un campo de datos se puede elegir como una clase de objeto o un atributo de una clase de objeto o ambos según el perfil del análisis.

Ejemplo 4:

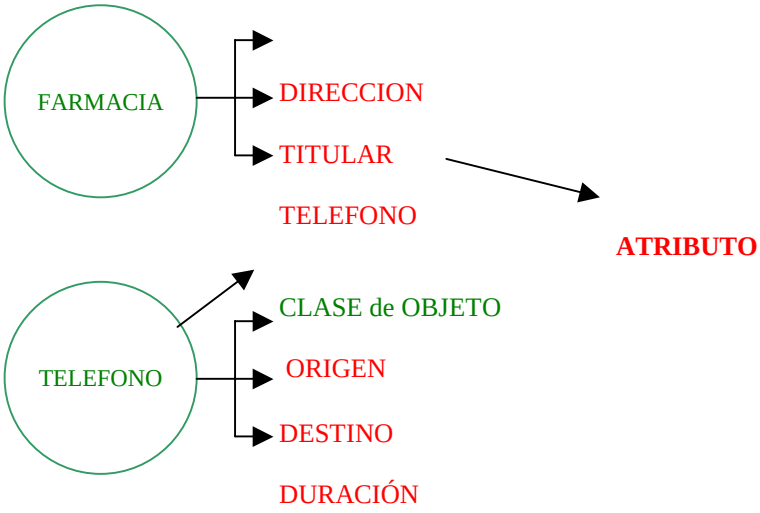

Figura 6. Bifuncionalidad de variables.

6.4 Relaciones

- Las relaciones enlazan dos clases de objetos según alguna característica en común.
- Se pueden suponer como clases de objeto, pudiéndole asignárseles atributos tales como fecha, actividades, estados, etc.
- El valor de cada atributo es único para cada conexión particular, cumpliéndose la regla de un valor por atributo.
- Se pueden establecer muchas relaciones entre par de objetos, pero son completamente independientes entre sí.

Ejemplo 5:

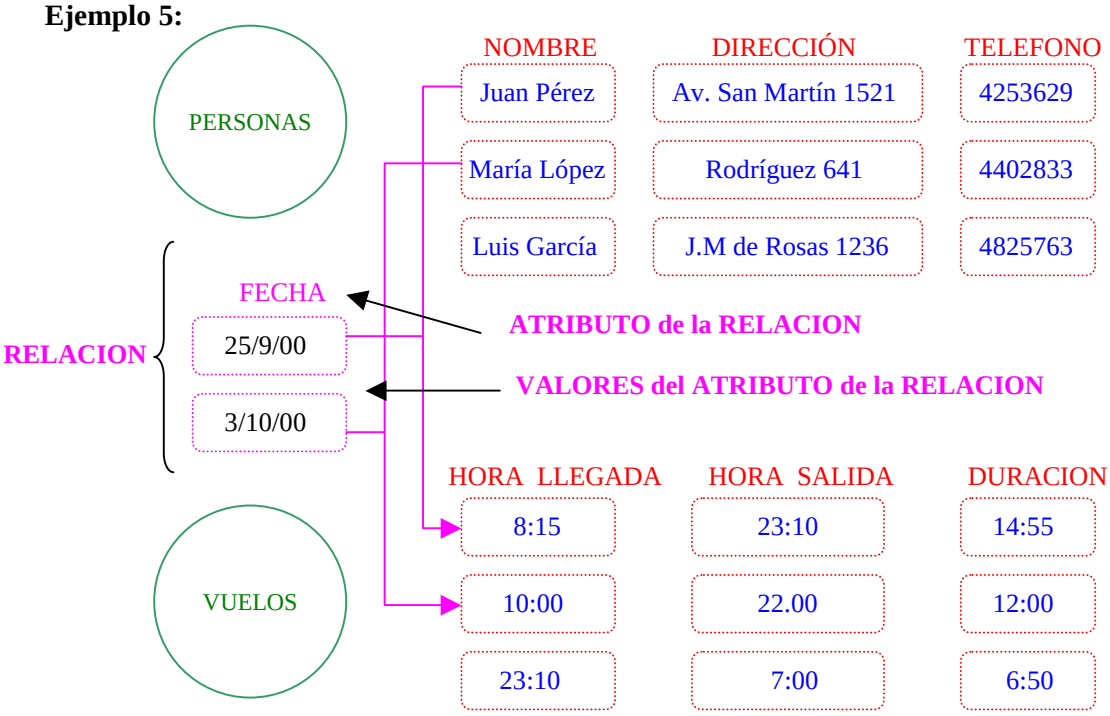

Figura 7. Relaciones.

Cuando se construye el modelo, se debe incluir la información que es relevante para el análisis, es decir, decidir qué parte de los datos incluir y cuales omitir, así, las clases de objetos, atributos y relaciones elegidos dependen de los objetivos del problema.

Sin embargo, hay que tener presente que durante el proceso de modelado, se puede retroceder y reestructurar el modelo para incluir componentes diferentes del set de datos original.

6.5 ¿Cómo pueden ser los Atributos?

- Nominales: son llamados Categóricos, Enumerados o Discretos. Enumerados se refiera a ponerlos en correspondencia con los números naturales, es decir, que implica un orden. Discretos se refiere a la posibilidad de discretizar una cantidad numérica continua. Además se incluyen los atributos *"booleen"*, que son aquellos que solo pueden tener dos valores, por ejemplo: Verdadero / Falso; Si / No; o 0 / 1.

- Ordinales: son llamados Numéricos o Continuos, pero sin la implicación matemática de continuidad porque pueden tomar valores enteros o reales.

7. ¿Qué es Abstracción de Datos?

La *Abstracción de Datos* es una técnica que permite simplificar o resumir la información, para llevar a cabo una evaluación inicial, más bien generalizada.

Se basa en la combinación de objetos con rasgos similares dentro de un único objeto.

Esta técnica permite procesar mayor cantidad de datos, pero tiene como desventaja que el nivel de detalle decrece y pueden no contemplarse patrones importantes.

En la *Abstracción*, el análisis se ejecuta sobre los datos resumidos, pero debe disponerse siempre de la información original.

Este método puede aplicarse no solo a datos numéricos, sino también a datos en forma cualitativa.

Ejemplo 6: En el caso de atributos con valores numéricos específicos, se pueden separar en distintos intervalos.

Valor Original		Intervalo
10.011	———	10
20.301	———	20
43.5	———	40
15.10	———	10
24.54	———	20
39.4	———	30
65.10	———	60
50.1	———	50

El tamaño de los intervalos, depende de los datos, si son muy pequeños fracasa el objetivo de resumirlos y si son muy grandes se corre el riesgo de perder información importante.

7.1 ¿Qué es Metadata?

Metadata puede definirse como *"datos en datos"*, así como *"datos acerca de datos"*. Esto puede considerarse como una técnica opuesta a la anterior, es decir, a la abstracción, porque se añade información para refinar los datos y aumenta el nivel de detalle, con el objetivo de exponer patrones y tendencias.

Ejemplo 7: El atributo fecha puede descomponerse de la siguiente forma:

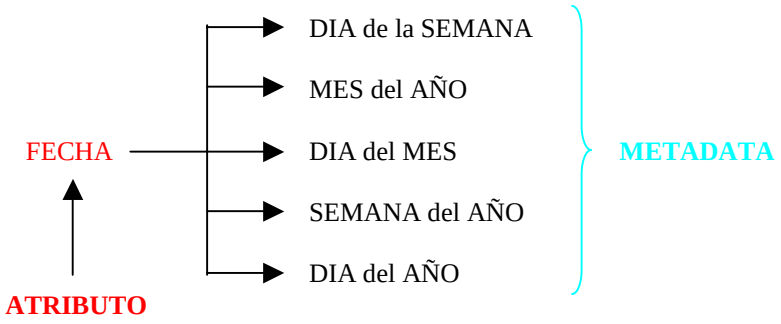

Figura 8. Metadata.

Y se puede aplicar en:

- Análisis de un set de datos de ventas basado en el mes donde pueden observarse las tendencias de las ventas entre los distintos meses del año.
- En procesos industriales donde se puede comparar rendimientos mensuales o anuales y determinar las causas de las diferencias encontradas para la futura toma de decisiones.

8. ¿Cómo pueden ser los Datos?

Los datos a modelar en un análisis pueden ser:

- Datos Descriptivos: caracterizan objetos discretos mediante atributos.
- Datos de Transacción: proporcionan información sobre el tiempo y lugar de eventos. En general, poseen una componente fecha-tiempo que permite diferenciar las distintas transacciones entre sí.

8.1 ¿Qué tipos de Modelos usar?

- Si los datos son descriptivos, se aplica un modelo descriptivo.
- Si los datos son de transacción, se utiliza un modelo descriptivo o de transacción.

Tener presente que los tipos de resultados obtenidos dependen del tipo de modelo elegido.

8.1.1 ¿Qué es un Modelo Descriptivo?

- Los modelos descriptivos exponen las relaciones entre los objetos presentes en el set de datos.
- Además, permiten mostrar redes y frecuencia de conexiones.
- Pero fracasan al intentar reflejar comportamientos o eventos.
- En un modelo descriptivo, cada enlace entre dos clases de objetos asociadas, posee un único valor o set de condiciones para describir tal relación a lo largo de todo el modelo.

Ejemplo 8: Cada individuo de la clase de objeto *"persona"* puede estar conectado con la clase de objeto *"dirección"* por un set de condiciones determinadas, que sea su domicilio particular. A su vez, cada individuo puede tener un enlace con un auto específico de la clase de objeto *"vehículo"* por una razón completamente diferente, es decir, por un set de condiciones distinto, que sea su propiedad particular.

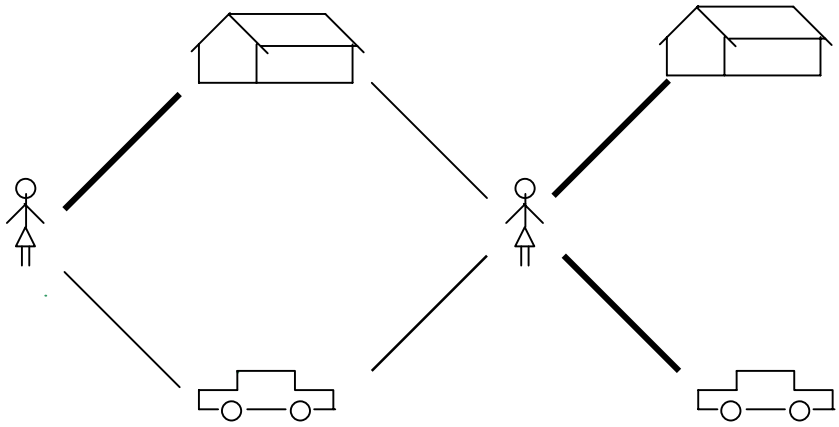

Figura 9. Modelo Descriptivo

8.1.2 ¿Qué es un Modelo de Transacción?

- En este tipo de modelo, los enlaces representan eventos propiamente dichos, donde a cualquier relación se le aplica todas las condiciones asociadas al evento que representa.

- Como los eventos se distinguen entre sí, especialmente por la fecha y tiempo de ocurrencia, todos los enlaces poseen, por lo tanto información diferente.

- Este modelo tiende, entonces, a reflejar comportamientos, pero tiene como desventaja el mayor costo computacional, porque manipula gran cantidad de objetos y enlaces.

Ejemplo 9: Una persona compra una misma mercadería repetidas veces (por ejemplo café), y cada transacción es única, ya que el hecho ocurre en un tiempo y fecha diferente.

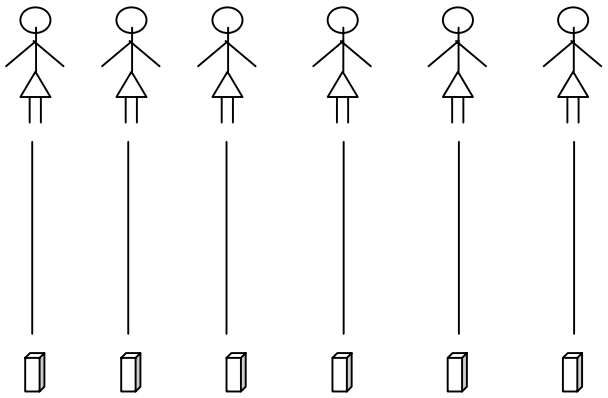

Figura 10. Modelo de Transacción.

9. ¿Cómo es el Proceso de *Data Mining*?

El proceso de *Data Mining* consta de las siguientes etapas:

- Definición del problema real.

- Acceso a los datos necesarios de las fuentes originales.

- Integración de múltiples fuentes de datos y creación de un formato que pueda representar toda la información consistentemente.

- Análisis de los datos, ya sea mediante métodos de visualización o bien métodos analíticos no visuales.

  En esta etapa, generalmente debe volverse atrás a las fuentes de datos originales y tomar nuevos casos, pudiendo ocurrir esto varias veces antes de completar el análisis.

- Presentación de resultados.

10. ¿Cómo Definir el Problema Real?

En algunas situaciones, existen objetivos determinados desde el comienzo y esto permite conocer previamente que puede ser importante descubrir y establecer límites al problema.

Los casos donde la investigación es más generalizada, y sin objetivos específicos desde el inicio, no constituyen un inconveniente, sino, todo lo contrario, porque *Data Mining* es un proceso iterativo donde la búsqueda se hace cada vez más refinada.

Estos tipos de pruebas se conocen como *Análisis Exploratorios*, en los cuales se espera simplemente hallar un resultado interesante y es donde se refleja la mayor utilidad de *Data Mining*.

Para estos casos, existen guías propuestas que ayudan a enfocar el objetivo del problema, siendo generalmente, una serie de preguntas sobre este en particular.

10.1 Análisis Reactivo vs. Análisis Proactivo.

El análisis se puede ejecutar en un *Modo Reactivo*, *Proactivo* o una combinación de ambas formas.

En los *Análisis Reactivos*, existe un objetivo fijo a perseguir desde el inicio, y es posible generar hipótesis con anticipación.

El análisis se orienta a una entidad, sus conductas y relaciones con otros objetos.

Para llevarlo a cabo, se utiliza toda la información disponible con respecto a dicho sujeto elegido como objetivo.

Se pueden detectar otras entidades conectadas a la original, las cuales se transforman en el siguiente nivel de investigación.

Permite procesar gran cantidad de información con rápidas respuestas.

En los *Análisis Proactivos*, en cambio, el punto inicial no se conoce, ni se puede definir previamente.

El análisis está enfocado en modelar los datos para descubrir patrones y tendencias no conocidos con anterioridad.

Los objetos importantes se presentan como estructuras aisladas o bien suceden con alta frecuencia, pudiendo detectarlos con facilidad.

En estos análisis, se realizan *"cortes proactivos"* de los datos, es decir, extracciones de muestras de una fuente de datos.

Estos métodos de análisis pueden combinarse dando lugar a un proceso iterativo entre ambos modos. Generalmente, comienza con la *Forma Proactiva*, a fin de, individualizar los

objetos de interés, una vez que éstos están determinados se pasa al *Modo Reactivo*, con el objetivo de analizar toda la información disponible y adicional en relación con dichos objetos.

Ejemplo 10: Búsqueda en Internet.

Palabra Clave: Data Mining

Respuesta:

1. **Data Mining** –A Practical Overview.
2. PSA Peugeot Citroën implement **Data Mining** tools.
3. Shell Canadá Ltd. Implementet **Data Mining** tools of Bussiness Objects.
4. **Data Mining** and Bussiness.
5. Hosokama Micron will implement **Data Mining** tools with XpertRule Miner.
6. Companies offering Data Mining Solutions.
7. Insurance and **Data Mining**.
8. **Data Mining** in Practice – Case Studies and Leassons Learned.
9. How Leading Companies are Competing using **Data Mining** for Customer discovery.
10. Utilizing **Data Mining** to Imcrease Sales Efficiency with Business Telecom Customers.

CORTE PROACTIVO

Elección: **Data Mining** – A Practical Overview

Respuesta:

Data Mining – A Practical Overview

Index

The evolution of Data Mining
Data Mining Tools
Types of Models and Industry Examples
.
.

ANALISIS REACTIVO

11. ¿Cómo Acceder a la Información?

11.1 Acceso a los Datos.

En esta etapa del proceso de *Data Mining*, el objetivo es la extracción de los datos de las fuentes originales, y para ello, la condición fundamental es la disponibilidad de los mismos.

No siempre el acceso es libre, en ciertas ocasiones, está limitado por la existencia de protección de propiedad literaria, o políticas de seguridad o bien se necesitan procedimientos especiales o protocolos para el acceso.

11.2 Extracción y Tests

Para llevar a cabo la extracción se utiliza como método la generación de una serie de "*tests"* que se aplican a las *Fuentes de Datos*, permitiendo el acceso buscado.

Al confeccionar *"tests"*, debe considerarse que la acumulación de gran cantidad de valores tiene un alto costo computacional, para ello existen técnicas de filtrado y segmentación que permiten refinar la extracción.

Durante el proceso de extracción debe asegurarse siempre la *Integridad* y *Consistencia* de los datos.

El análisis comienza de *Forma Proactiva*, donde se realiza la extracción de un subset de datos que es una muestra representativa del set de datos inicial.

Este proceso de extracción es un *"corte proactivo"* de la información.

CORTE PROACTIVO

Aunque si bien, este modo de análisis no se caracteriza por definir los objetivos previamente, es necesario especificar alguna estructura para poder determinar que datos se incluyen en el subset.

Una vez identificado algún objeto de interés en el "*corte proactivo"*, el análisis pasa al *Modo Reactivo*, a fin de disponer de toda la información adicional con respecto a dicho objeto y utilizando un lazo de realimentación de forma de refinar la investigación sobre los datos.

MODO REACTIVO

11.3 Transferencia de Datos

Cuando se habla de extracción, debe tenerse en cuenta el lugar de origen y de destino de los datos, es decir, la transferencia de éstos desde la plataforma principal al ambiente de *Data Mining*.

Esta transferencia no necesariamente se realiza en *Tiempo Real.*

Los datos se pueden convertir en *"flat-fileformats"*, es decir, cualquier tipo de información usada por *Data Mining* se establece dentro de depósitos temporarios donde se puede procesar por software de *Data Mining*. Estos depósitos temporarios pueden ser:

Bases de Datos.

Procesadores de Textos

Spreadsheet

12. ¿Cómo Integrar los Datos?

Los datos que se usan en un análisis no necesariamente provienen de las mismas fuentes, generalmente, las fuentes y tipos de información pueden ser ilimitados y por lo tanto, luego de acceder a los datos es necesario integrarlos.

La idea de integración se conoce como *"Almacén de Datos" (Data Warehousing)* y la tendencia hacia éste es un reconocimiento de que la información fragmentada puede tener un gran valor cuando se la reune e integra, por eso, comúnmente se dice que el *Almacén de Datos* es un precursor de *Data Mining*.

Es preferible siempre extraer los datos de *Bases de Datos* establecidas, pero durante el análisis pueden aparecer formatos no estándares, tales como:

- Texto libre:

  - Encuentran su mayor utilidad con set de datos pequeños.
  - Estos textos se pueden condensar y resumir.
  - Pero, el texto es un formato difícil de integrar con otros tipos de datos.

- Tablas:

  - Proveen mejores mecanismos para presentar y analizar la información.
  - Permiten ordenar los tipos de datos similares en determinadas zonas para su rápida detección.
  - Poseen la capacidad de combinar información de varias fuentes.
  - Pero, no transfieren grandes cantidades de datos.

- Otros formatos:

  - Incluye gráficas, imágenes, fotos, videos, sonidos.

12.1 Normalización de Datos.

Para realizar la integración de los datos de varias fuentes, es necesario que dichos datos estén normalizados:

- Todas las unidades de media deben llevarse a una misma escala.
- Debe asegurarse una terminología consistente.
- Los tipos de datos similares convienen representarse juntos.
- Es beneficioso aplicar técnicas de reducción de datos, eliminando información duplicada.

12.2 Limpieza de los Datos.

En general, cualquier base de datos contiene datos inconsistentes, incompletos o erróneos, los cuales pueden ocurrir por varias razones, tales como, caracteres transformados y/o mal deletreados en la entrada de datos, datos perdidos, formatos incompatibles, datos ingresados incorrectamente en pantallas de entrada, etc.

12.2.1 Valores Perdidos.

Se asume comúnmente, que un valor perdido es un valor no conocido.

Pero pueden existir muchas causas para que esto ocurra, por lo tanto, la responsabilidad de decidir si el valor perdido es significativo o no, si es posible, cae sobre alguien bien familiarizado con la información.

12.2.2 Valores Erróneos.

- Valores incorrectos pueden ocurrir cuando éstos cambian de forma insignificante, ya que cualquier perturbación pequeña en el deletreado de un valor da como resultado múltiples representaciones de la misma entrada.

Ejemplo 11:

Sr. Juan López	López, A. Juan	Representan la misma persona
Juan A. López	Sr. Juan A. López	
López, Juan	Dr. Juan López	

- Pueden existir errores no en le deletreado, sino cuando se crean diferentes valores para un mismo objeto.

Ejemplo 12:

Coca Cola
Coca
Cola

- Se hace importante la limpieza de los datos de manera de asegurar la *Consistencia* y *No Ambigüedad* de los mismos.

- Pueden aplicarse rutinas de limpieza, tal como para el Ejemplo 11: eliminar los caracteres aislados (inicial del segundo nombre), eliminar sufijos y prefijos (Sr., Dr.):

Ejemplo 13:

Sr. Juan López	López, A. Juan
Juan A. López	Sr. Juan A. López
López, Juan	Dr. Juan López

Luego pueden aplicarse métodos más sofisticados como algoritmos o rutinas automáticas donde el objetivo es siempre reducir los sets de datos eliminando duplicaciones.

12.2.3 Información en Formato Tipo Texto.

Algunas *Fuentes de Datos* comprenden solo texto con un arreglo consistente de la información, pero en la mayor parte de los casos, dicho formato consistente no existe.

En tales situaciones, pueden emplearse distintas aproximaciones, por ejemplo, algunas toman el contenido de un documento y lo separan en un set de estructuras con indicadores de referencia. Al solicitar una palabra o clave determinada, se pueden encontrar de forma automática todas las ocurrencias de éstos términos dentro de los documentos, produciendo un set de palabras que se clasificarán según diversos factores, tal como, frecuencia de ocurrencia.

A veces, los resultados de dicha búsqueda son muy extensos y se pueden resumir sobre la base de un valor o vector que se crea en función de los contenidos verdaderos del documento.

Así, cualquier otro documento que posea valores similares, se supone que incluye contenidos semejantes.

13. ¿Cómo llevar a cabo el Análisis de los datos?

13.1 Métodos de Visualización vs. Métodos Analíticos.

En el uso de *Métodos de Visualización* no necesariamente se debe tener un objetivo de búsqueda desde el inicio del análisis, el cual toma entonces un *Modo Exploratorio*.

Los *Métodos de Visualización* permiten descubrir tendencias y patrones que no sería detectados por análisis no visuales.

Los patrones se detectan en función de violaciones de límites, frecuencia de ocurrencia.

La representación gráfica de los datos soporta inspeccionar grandes cantidades de información al mismo tiempo.

Los *Métodos Analíticos* no visuales requieren saber previamente qué se espera encontrar, permitiendo formular hipótesis previas.

Los resultados obtenidos son más bien, grupos de tendencias y deferencias generales.

En éstos análisis, se emplean aproximaciones incluyendo *Pruebas Estadísticas, Arboles de Decisión, Asociación de Reglas, Redes Neuronales y Algoritmos Genéticos.*

En la ejecución de *Data Mining*, es muy útil combinar los *Métodos Analíticos* y de *Visualización*.

14. Métodos de Visualización

14.1 Representación Visual vs. Representación Tabular

En la información verbal se procesa un ítem en cada momento y para hallar relaciones de interés, se debe examinar cada ítem, calcular la frecuencia de ocurrencia de enlaces, pudiendo omitir relaciones fundamentales.

El desarrollo visual, en cambio, sucede de forma inmediata y en paralelo. Las relaciones importantes se descubren automáticamente, señalando los objetos más activos directamente.

La representación visual en forma de diagrama de red además de indicar los objetos más activos, expone asociaciones con otras entidades en el set de datos permitiendo exhibir patrones de forma inmediata.

Ejemplo 14: Se analiza datos acerca de reuniones de trabajo dentro de una organización.

PERSONA	PERSONA	FECHA	HORA
Juan	María	13-10-00	9:30
Susana	Juan	24-10-00	15:15
Norma	María	15-10-00	17:35
Alejandra	Lucas	14-10-00	10:00
María	Alejandra	28-10-00	12:10
Cecilia	Juan	03-11-00	12:25
Carlos	María	29-09-00	8:15
Susana	Carlos	02-10-00	14:00
Alejandra	Susana	30-09-00	15:35
María	Cecilia	15-10-00	18:45
Cecilia	Lucas	21-10-00	11:20
Norma	Carlos	27-09-00	13:05
María	Susana	01-11-00	16:40
Cecilia	Alejandra	05-10-00	19:25

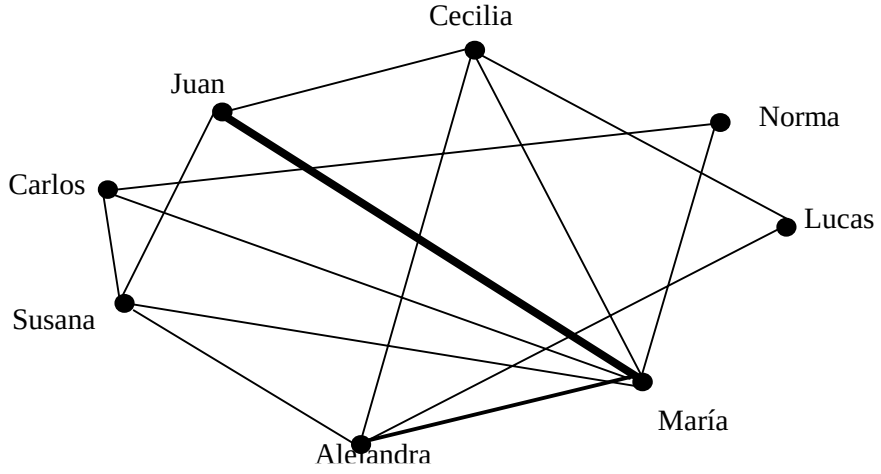

Figura 11. Representación Visual.

La mayor utilidad del formato visual se refleja cuanto más grande es la cantidad de datos analizados.

14.2 ¿Cómo Colocar los Datos dentro de una Representación Visual?

La estructura de la representación visual para una aplicación en particular está determinada por los tipos de datos y el modelo elegido.

Los objetos dentro de la representación tienen determinadas características según los valores de los atributos, tales como, color, forma, tamaño, estilo, rotación, intensidad, textura, brillo.

Dichas características pueden utilizarse para destacar valores específicos.

Los datos dentro de dicha representación se ubican en función de los valores de los atributos y/o de los enlaces entre clases de objetos, y para ello existen diferentes formatos:

- Agrupaciones *(Clustering)*

Los objetos se sitúan según valores, más generalmente descriptivos, es decir, atributos.

En este caso, se divide un conjunto de datos en grupos mutuamente excluyentes, de forma tal, que cada miembro de un grupo se encuentre lo más cerca posible a otro, y los distintos grupos estén lo más alejados posibles uno de los otros.

Puede medirse la distancia entre los grupos diferentes en función de todas las variables disponibles.

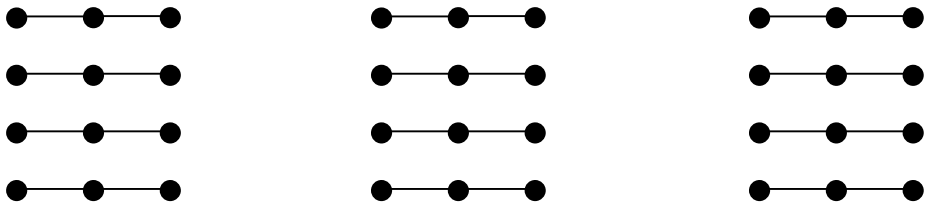

Figura 12. Agrupaciones.

- Estructuras Jerárquicas o Arboles

En esta situación, la posición de los objetos se establece según sus relaciones con otros objetos, pero en algunas situaciones la ubicación se encuentra en función se valores de los atributos.

La parte superior del árbol se conoce como *"nodo raíz"* y su selección determina la construcción de la estructura jerárquica.

Si este se escoge incorrectamente, la estructura se hace difícil de interpretar.

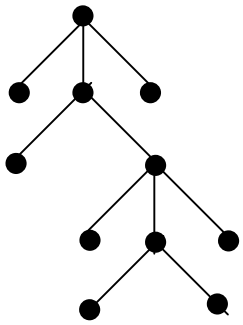

Figura 13. Estrucuturas jerárquicas.

- Redes:

Los objetos se ubican dentro de grupos discretos de datos.

La red de objetos dinámica alcanza el estado estacionario cuando la posición de los objetos según sus asociaciones es la óptima.

Tales redes pueden separarse formando grupos de redes pequeñas.

Para que la red sea fácil de comprender, es preferible que los objetos en el set de datos no estén altamente interconectados.

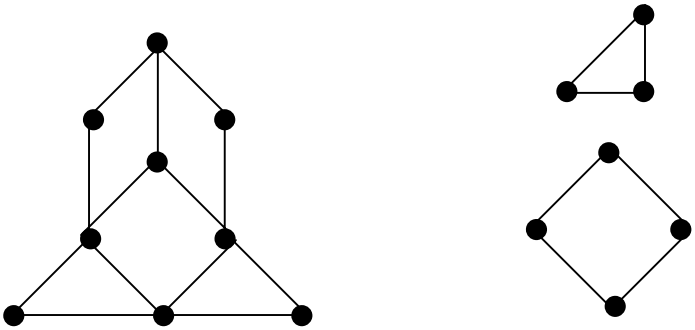

Figura 14. Redes.

- Panoramas o Posiciones Geográficas

Se considera como una forma de agrupación, pero donde se trazan los valores de los atributos en una estructura espacial o grilla, permitiendo ubicar los objetos dentro del panorama.

Este método se emplea, más generalmente, para representar información dada en coordenadas espaciales, pero la ubicación de los objetos en el panorama no necesariamente debe estar basada en valores de atributos que posean tales coordenadas espaciales, pero sí que pueden analizarse en este formato de forma simple.

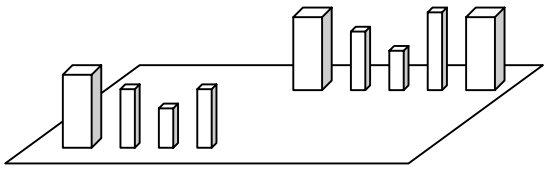

Figura 15. Panoramas

14.3 El Análisis Propiamente Dicho

Existen distintos tipos de aproximaciones o técnicas para llevar a cabo el análisis, tales como:

14.3.1 Análisis de Características Estructurales

El análisis en este contexto se orienta a la *Agrupación* donde las representaciones visuales pueden transmitir, por sí solas, grandes cantidades de información al ser simplemente examinadas.

Por la ubicación de los objetos en la representación, pueden detectarse fácilmente patrones importantes como también características inusuales, las cuales transfieren información acerca de:

- Datos que exceden Condiciones Límites

Mediante la representación panorámica de los datos, pueden localizarse fácilmente aquellos valores alejados de la representación principal.

Si tales valores superan condiciones límites o posibles se consideran como erróneos o malos y pueden eliminarse.

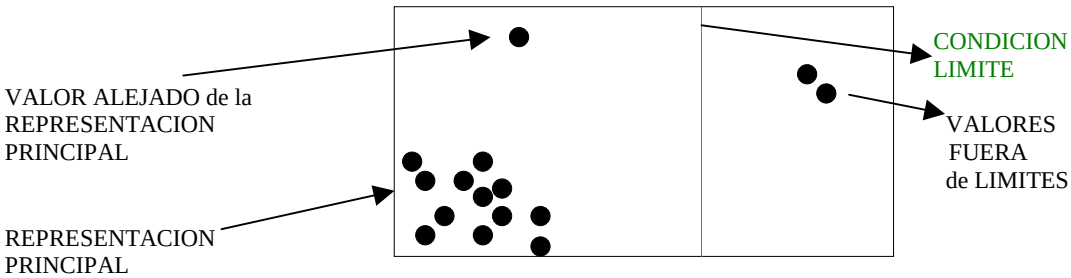

Figura 16. Datos que exceden Condiciones Límites.

- Datos Perdidos

Utilizando la alternativa de *Agrupación* de datos, la representación permite detectar inmediatamente aquellos datos perdidos, lo cual sería tedioso de concretar utilizando un formato tabular.

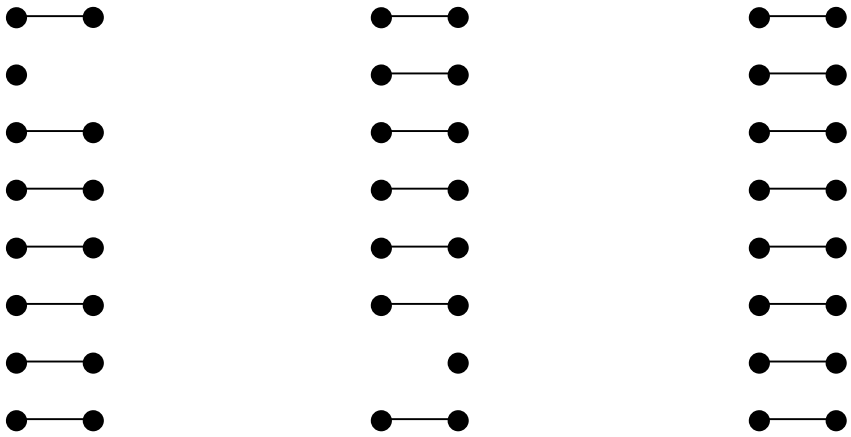

Figura 17. Datos Perdidos.

- Patrones Anómalos

En algunas situaciones, los sets de datos se representan como eventos que suceden en un orden particular, y si dicho orden se altera, se produce una anomalía que puede señalar patrones importantes.

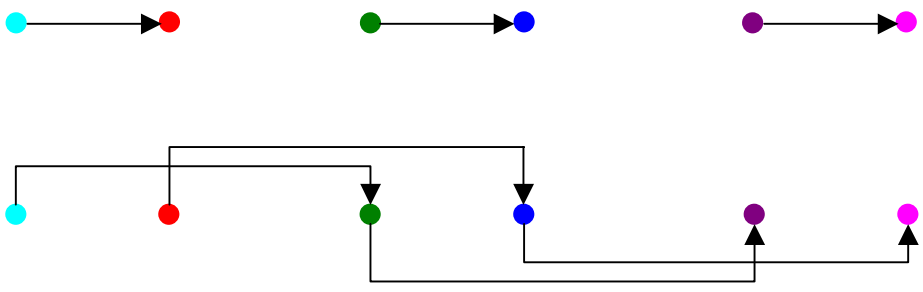

Figura 18. Patrones Anómalos.

14.3.2 Análisis de Redes

El análisis, sin la incorporación de los enlaces entre objetos tiende a ser una simple *Agrupación*.

Al incluir dichas conexiones aparecen otros tipos de patrones que proporcionan información adicional y esto es lo que ocurre al llevar a cabo un *Análisis de Redes*.

Algunas capacidades de las redes son:

- Interconectividad

Mediante el empleo de herramientas de análisis de enlaces, se pueden determinar aquellos objetos altamente "*interconectados*", como también relaciones inusuales que pueden indicar potenciales patrones.

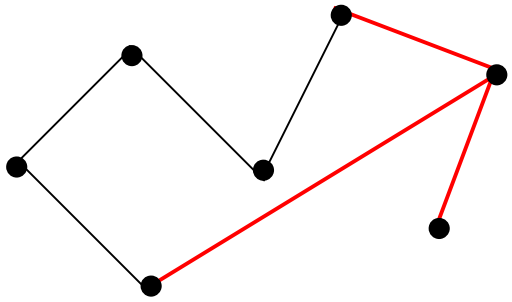

Figura 19. Interconectividad.

- Puntos de Articulación

La detección de *Puntos de Articulación* en el set de datos es una aproximación muy útil en los *Métodos de Visualización*. Tales *Puntos de Articulación* son objetos que enlazan dos o más subredes, e indican la importancia de dichos objetos.

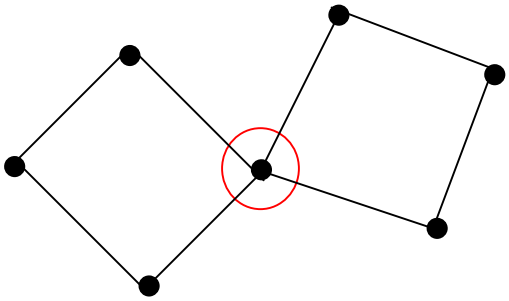

Figura 20. Puntos de Articulación.

- Redes Discretas

En algunas situaciones, el objetivo es identificar subredes dentro de una estructura de red grande y compleja y analizarlas independientemente.

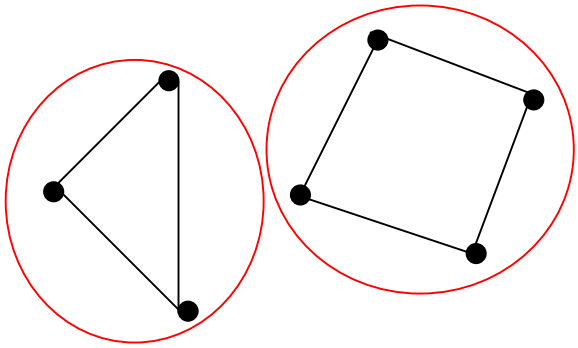

Figura 21. Redes Discretas.

- Conexiones Perdidas

Las *Conexiones Perdidas* son una forma especial de las *Redes Discretas*, donde la subred esta formada por un único objeto.

Encontrar objetos separados de la estructura principal puede indicar datos *Inconsistentes* o *Incompletos*, aunque se hallan aplicado técnicas de filtrado.

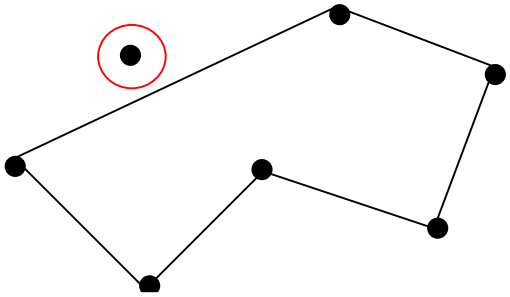

Figura 22. Conexiones Perdidas

- Enlaces Fuertes y Débiles

En ciertas situaciones, es útil examinar la frecuencia de enlaces entre objetos, que está determinada por el número de observaciones en el set de datos original y la cual puede visualizarse fácilmente en función de la intensidad relativa de las relaciones dentro de la red.

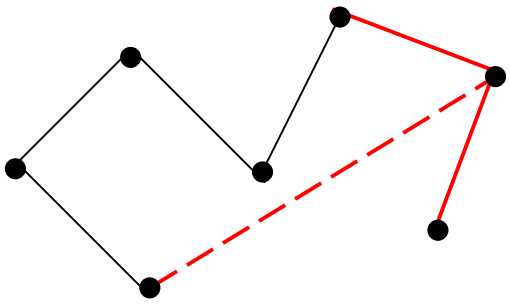

Figura 23. Enlaces Fuertes y Débiles.

- Frecuencias en Abanico

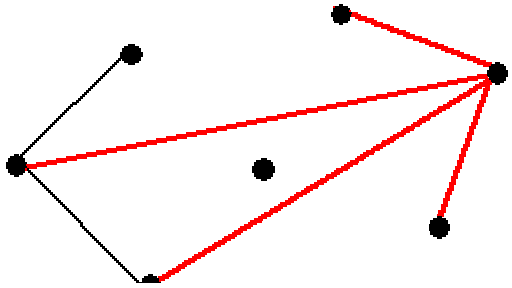

Figura 24. Frecuencias en Abanico.

En otros casos pueden diferenciarse aquellos objetos conectados con una gran cantidad de otros, formándose un gran *Efecto Abanico*, que puede indicar comportamientos no usuales.

Estos objetos también pueden considerarse como *Puntos de Articulación*.

- Análisis de Caminos

A veces, la finalidad del análisis es establecer si ciertos objetos pueden conectarse a través de una serie o secuencia de enlaces, es decir, encontrar si existen al menos uno o más *Caminos* que asocien dos objetos definidos.

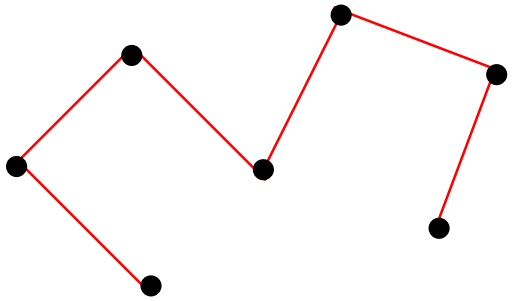

Figura 25. Análisis de Caminos.

- Enlaces en Común

En otras ocasiones, se buscan *Enlaces en Común* dentro de una red, es decir, cuando dos o más objetos comparten un tercer objeto.

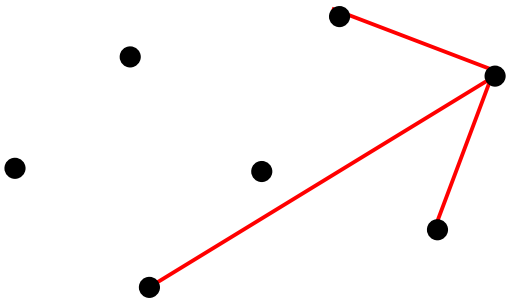

Figura 26. Enlaces en Común.

Estas técnicas se pueden aplicar a través de un simple set de datos, obteniendo distintos tipos de patrones, pero además utilizando una combinación de éstas es posible detectar patrones más complejos.

14.3.3 Análisis de Patrones de Conexión

Existen algoritmos que exponen patrones de conexión, los cuales se detectan por la *Agrupación de Objetos o "nodos"* que comparten más enlaces entre sí que con otros objetos fuera del grupo.

Estos análisis son más significativos en contextos donde existen relaciones entre entidades, tal como la estructura de una organización o bien entre personas.

- Grupos

Estos algoritmos identifican aquellos *"grupos"* de objetos íntimamente unidos según sus relaciones con otros objetos.

No todos los datos se encuentran dentro de estos "*grupos*", pero pueden una gran importancia en la estructura de la red:

- Enlaces

Son objetos que pueden estar altamente conectados con otros *"nodos"* en la red, ya sea con varios "*grupos*" o bien con otros "*enlaces*".

- Objetos Adjuntos

Estas entidades están dentro de la red conectadas con otros objetos, ya sea con un miembro de un *"grupo"* o con un "*enlace*", pero solo uno por vez.

- Objetos Aislados

Estas entidades no son parte de ninguna red, de ningún *"grupo*", ni tienen conexiones con otras entidades en la red.

Para comprender mejor este tipo de análisis, puede considerarse el siguiente ejemplo:

Ejemplo 15: Si se considera la estructura de una organización en cuanto a sus operaciones y sus comunicaciones, mediante la examinación de "*grupos*" se puede determinar inmediatamente subgrupos de individuos que se encuentran más estrechamente interconectados, ya sea por como

los individuos se comunican con otros dentro del "*grupo*" o quienes trabajan juntos.

Los "*enlaces*" permiten en una organización determinar como funcionan las comunicaciones y los "*objetos adjuntos"* pueden representar individuos ajenos a la organización, pero que por otra parte se encuentran conectados con algunos individuos de ésta, por ejemplo, el uso de terceros por parte de cualquier sector de la organización.

Los "*objetos aislados"* pueden representar individuos que no estaban presentes en el momento de la recolección de datos, ya sea por licencia, maternidad, franco, vacaciones, o viajes de trabajo.

14.3.4 Análisis de Patrones Temporales

Como el tiempo es una componente significativa en muchas aplicaciones, una cuarta clase de análisis abarca la detección de *Patrones Temporales*.

Estos patrones se localizan al determinar el ciclo de tiempo donde los cuales ocurren.

En ciertos casos se definen "*ciclos de conveniencia"* en intervalos más simples de interpretar y que generalmente abarcan representaciones temporales *Abstractas* (minutos, horas, días, semanas, años), pero no siempre los patrones de interés se encuentran en tales intervalos.

Los *Patrones Temporales* pueden ser *Absolutos* o *Contiguos*.

Los *Patrones Absolutos* exponen la cantidad de tiempo que toma cualquier evento o un set de eventos.

Los *Patrones Contiguos* se concentran sólo en el orden de ocurrencia de tales eventos.

Un *Evento Cíclico* es una conexión dada entre dos objetos que sucede con cierta frecuencia y lo que se persigue en estos análisis es determinar patrones a eventos cíclicos.

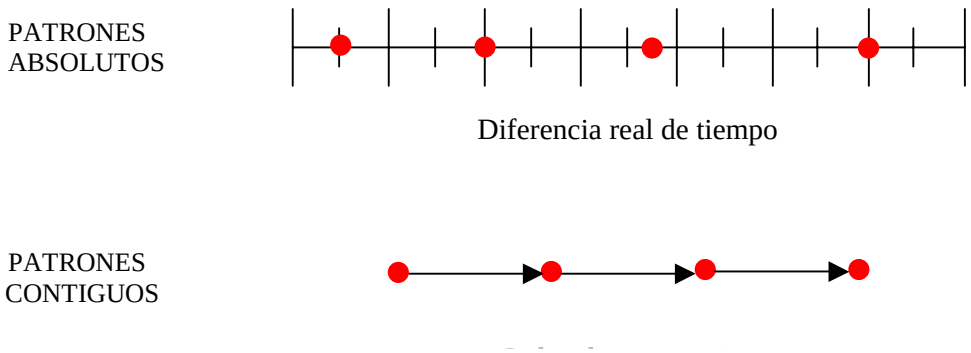

Figura 27. Patrones Absolutos y Contiguos.

- Eventos Cíclicos Absolutos

En el inicio de la búsqueda automática de un patrón se debe definir un *"evento disparador"*, pudiendo ser la ocurrencia de un objeto determinado con un valor específico del atributo, o bien, con más de un atributo para una o varias clases de objetos, siendo la búsqueda, entonces, mucho más enfocada.

Se pueden definir los límites y el tamaño del ciclo de tiempo para la búsqueda. Este ciclo se descompone en una serie de intervalos discretos de tiempo de igual longitud, y en cada uno de ellos se establece la ocurrencia o no del "*evento disparador*".

Esto se repite con lapsos de tiempo más pequeños hasta alcanzar los límites prefijados. Si el número de intervalos en que ocurre el "*evento disparador*" excede un valor previamente definido, se obtendrá un patrón y luego puede correrse nuevamente con un "*evento disparador*" más refinado.

Ejemplo 16: Bajo un modelo de evento cíclico absoluto, puede determinarse un patrón que ocurre cada mes.

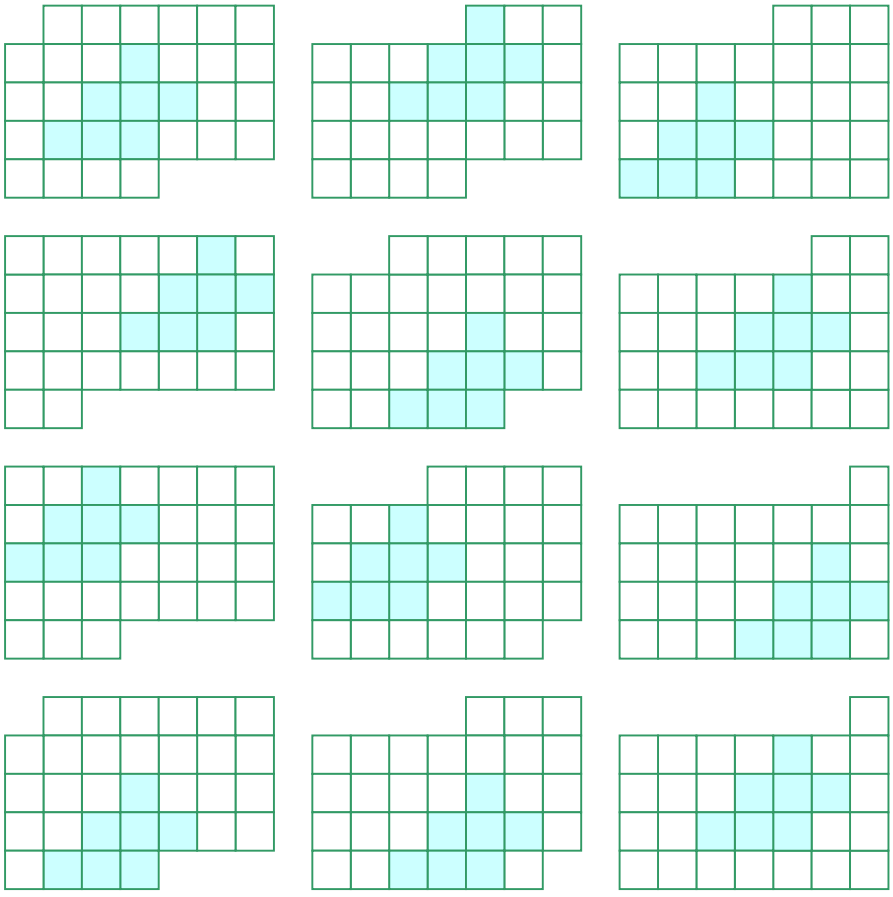

Figura 28. Modelo de Evento Cíclico Absoluto.

- Eventos Cíclicos Contiguos

En esta situación se busca la ocurrencia de dos o más eventos consecutivos a través de una serie de intervalos de tiempo.

Se define el "*evento disparador*", como también los límites de los intervalos de tiempo, sus incrementos y decrementos.

Por cada detección del "*evento disparador inicial*", se reúnen todos los datos dentro de los límites de un intervalo subsiguiente, ya sea, el siguiente al disparador, el anterior o ambos.

Luego de pasar por todos los datos se determina un patrón por la repetición de un evento junto con el "*evento disparador inicial*".

Ejemplo 17: Se puede detectar un patrón, por la repetición del evento disparador seguido de otros dos eventos, manteniendo siempre el mismo orden.

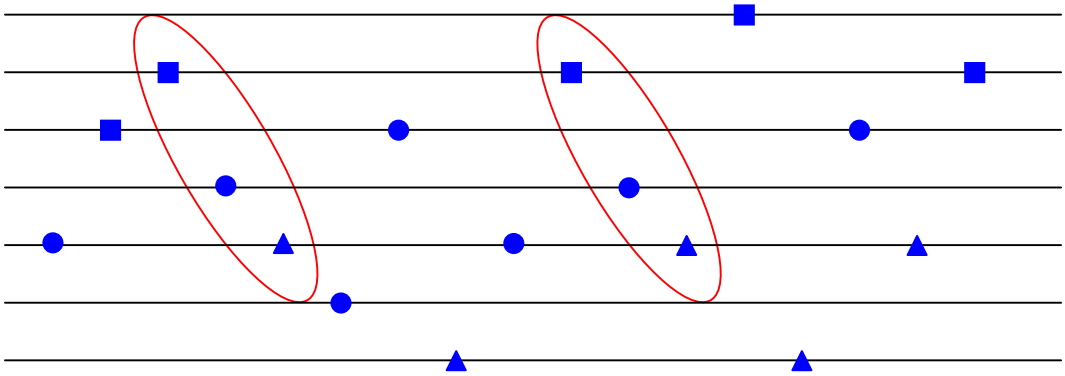

Figura 29. Modelo de Evento Cíclico Contiguo.

Cuando los algoritmos de detección de *Patrones Temporales* corrieron sobre todo el set de datos, se origina un set de patrones seguramente en forma tabular, que es difícil de interpretar.

Para salvar esta situación se usan *metodologías de Visualización*, por ejemplo, puede emplearse una grilla tridimensional que permite incrementar o disminuir el tamaño de los intervalos para que emerjan patrones.

15. Métodos Analíticos No Visuales

15.1 Métodos Estadísticos

Para llevar a cabo el análisis de los datos se puede utilizar la *Estadística Descriptiva* o *Inferencial*.

La *Estadística* requiere datos numéricos y abarca cómputos matemáticos.

La *Estadística Descriptiva* incluye medidas como promedio, media, moda, desviación estándar, rango y además existen pruebas estadísticas como las posteriores.

15.1.1 Análisis de Grupos (Cluster Analisys)

En este tipo de análisis se crean distintos segmentos o *Grupos Estadísticos* a partir de un set de datos, con el objetivo de establecer la existencia de diferencias entre estos grupos predefinidos.

El método trabaja en forma de pruebas de hipótesis, las cuales predicen ciertas diferencias entre grupos contra la hipótesis nula que no supone diferencia alguna.

Lo más importante en este análisis no es que exista cualquier diferencia, sino que la misma sea lo suficientemente significativa como para señalar una tendencia interesante.

15.1.2 Análisis Predictivos: Regresión

En este caso el objetivo es realizar *Predicciones* acerca de valores de una o varias variables y para ello se puede ejecutar:

- Regresión Lineal

Es una técnica estadística utilizada para encontrar la mejor *Relación Lineal* entre una variable seleccionada dependiente (y) y las variables independientes (x).

Este modelo requiere que todas las variables sean lineales, pero que no exista interacción alguna entre las variables independientes.

Si existen variables no lineales, se necesita una transformación lineal.

También tiene como requisito que todas las variables sean continuas (ej. Tiempo) y no categóricas (ej. Si/No, Verdadero/Falso, Femenino/Masculino).

El resultado de este modelo de *Regresión Lineal* es la ecuación matemática de la línea que mejor ajusta al set de datos, la cual puede utilizarse para llevar a cabo la *Predicción*.

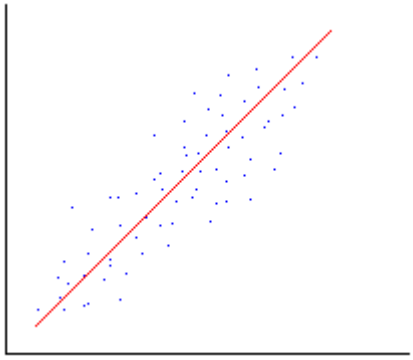

Figura 30. Regresión Lineal.

- Regresión Logística

Es una *Regresión Lineal* que predice las proporciones de una *Variable Categórica* seleccionada.

Es necesario que todas las variables sean lineales, pero que las variables independientes no interactúen entre sí.

También, en el caso de variables no lineales, se precisa una transformación lineal.
Este método es más complejo en términos del tiempo necesario, la preparación de los datos y la necesidad de usuarios experimentados con este tipo de variables categóricas.

15.2 Arboles de Decisión

Son estructuras en forma de árbol que representan un conjunto de decisiones.

Tales decisiones generan reglas para la clasificación de un conjunto de datos.

Los *Arboles de Decisión* se utilizan, generalmente, cuando el objetivo es realizar una clasificación o una predicción categórica y no tanto para ejecutar predicciones de variables cuantitativas, por lo tanto, requiere que todas las variables independientes sean categóricas.

Afortunadamente, es posible convertir variables continuas en categóricas a través de procesos de segmentación.

15.2.1 Uso y Construcción de Arboles de Decisión

Los "*nodos"* en un árbol de decisión abarcan la prueba de un atributo particular, comparando el valor del atributo con una constante, comparando dos atributos entre sí o usando una función de uno o más atributos.

Las "*hojas"* de los nodos dan una clasificación o "*clase*" que se aplica a todos los objetos que lleguen a la misma.

Para clasificar un objeto no conocido se encamina el árbol hacia abajo según los valores de los atributos probados en los nodos sucesivos y cuando se llega a una hoja, el objeto se clasifica según la clase de dicha hoja, en otras palabras, el árbol predice la clase para el objeto nuevo.

Para la construcción del árbol primero debe seleccionarse el atributo para determinar el *"nodo raíz"*.

Luego se realiza una "*ramificación"* por cada valor de dicho atributo.

El proceso se repite para cada rama eligiendo los "*nodos hijos*" y utilizando solo los objetos que llegan verdaderamente a ella.

Cuando todos los objetos en un nodo tienen la misma clasificación o clase, se detiene el desarrollo del árbol en dicha parte.

Para establecer el nodo raíz se utilizan aquellas variables que proporcionen un nivel de segregación o división de los datos máximo y esto se repite para le selección de los nodos hijos.

El número de ramas por nodo puede variar y el caso extremo sucede cuando un atributo tiene un valor diferente para cada objeto en el set de datos, originando un enorme "*efecto abanico*".

Ejemplo 18: El objetivo es predecir si una persona dada que es entrevistada para un puesto de trabajo, según sus condiciones, lo obtendrá o no.

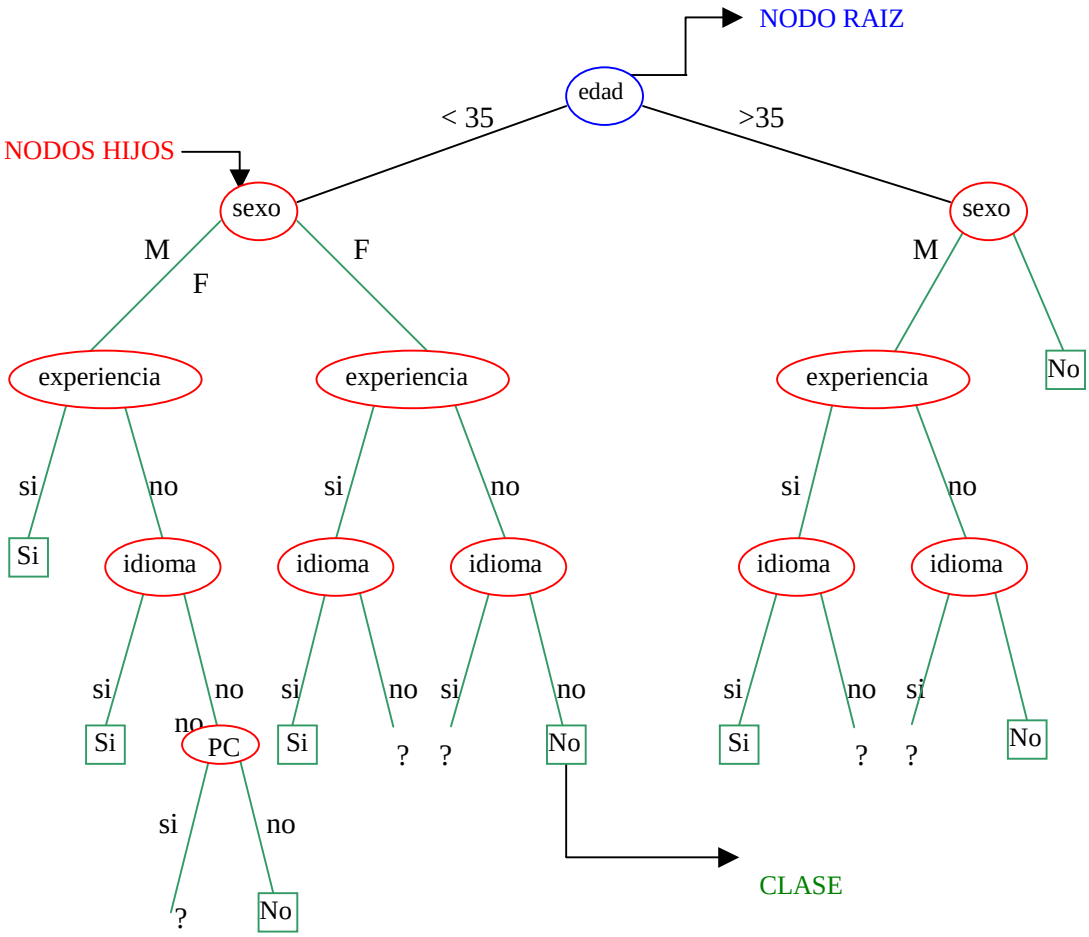

Figura 31. Arbol de Decisión.

15.2.2 Construcción de Reglas: *Clasificación*

Las decisiones de un árbol originan *Reglas* para la *Clasificación* de un conjunto de datos.

La *Clasificación* es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de forma tal que cada miembro de un grupo esté lo más cerca posible de otros y

grupos diferentes están lo más lejos posible de otros, donde la distancia se mide con respecto a las variable/s especificada/s, la/s cual/les se quiere/n predecir.

Un *Arbol de Decisión* se puede convertir en un set de reglas efectivas, pero cuya conversión no es tan trivial.

Una regla se genera por cada hoja, incluyendo una condición por cada nodo según el camino desde el nodo raíz hacia la hoja y finaliza con la clase asignada por ésta última.

Este procedimiento, por la *Redundancia* incluida en las estructuras de las reglas, *no hay Ambigüedad* alguna en la interpretación, es decir, nunca existirán dos clasificaciones diferentes para el mismo objeto y por este hecho el orden en que se ejecutan es irrelevante.

Sin embargo, las reglas tienen la intención de ser interpretadas en orden, como una *Lista de Decisión* y algunas de ellas fuera de contexto puede ser incorrecta.

Ejemplo 19: A partir del árbol de decisión del ejemplo 17 puede obtenerse el siguiente conjunto de reglas.

```
If edad < 35, sexo: M y experiencia: si
                                                    then
                                                si

If edad < 35, sexo: M experiencia: no y idioma: si
                                                    then
                                                si

If edad < 35, sexo: M experiencia: no, idioma: no y PC: no
                                                    then
                                                no

If edad < 35, sexo: F, experiencia: si y idioma: si
                                                    then
                                                si

If edad < 35, sexo: F, experiencia: no y idioma: no
                                                    then
                                                no

If edad > 35 y sexo: F
                                                    then


If edad > 35, sexo: M, experiencia: si y idioma: si
                                                    then
                                                si

If edad > 35, sexo: M, experiencia: no y idioma: no
                                                    then
                                                no


otherwise  ?
```

15.2.3 Reglas vs. Arboles

En algunos casos, un árbol corresponde exactamente al set de reglas, pero en otras situaciones las reglas son más compactas y esta diferencia crece al aumentar el tamaño del árbol.

Se pueden añadir reglas a las existentes sin perturbación alguna, en cambio, añadirlas a la estructura de un árbol requiere modificar o rehacer el árbol por completo.

Los algoritmos para la construcción de reglas, están enfocados en crear un a regla con la máxima exactitud, en cambio, los algoritmos para la construcción de árboles se concentran en examinar la separación entre clases. En cada caso se trata de encontrar un atributo para dividir, pero el criterio de selección es diferente.

Métodos específicos de árboles de decisión incluyen:

- Arboles de Clasificación y Regresión (CART, Classification and Regression Tree)

Es una técnica de *Arbol de Decisión* que se usa para la clasificación de un conjunto de datos.

Provee una serie de reglas que se pueden aplicar a un nuevo set de datos (sin clasificar), para predecir cuales registros darán un cierto resultado.

Segmenta al set de datos creando dos divisiones.

- Detección de Interacción Automática de Chi Cuadrado (CHIAD, Chi Square Automatic Interratio Detection)

Es una técnica de *Arbol de Decisión* que se usa para la clasificación de un conjunto de datos.

Provee una serie de reglas que se pueden aplicar a un nuevo set de datos (sin clasificar), para predecir cuales registros darán un cierto resultado.

Segmenta un set de datos utilizando test Chi cuadrado para crear múltiples divisiones.

15.3 Asociación de Reglas

Esta técnica se utiliza cuando el objetivo es realizar *Análisis Exploratorios*, buscando relaciones dentro del set de datos.

Las *Asociaciones* identificadas se pueden usar para predecir comportamientos.

Las *Asociaciones* permiten descubrir correlaciones y co-ocurrencia de eventos.

Las *Reglas de Asociación* pueden predecir cualquier valor del atributo, no solo la clasificación o clase específica, y además permiten predecir más de un valor del atributo al mismo tiempo.

Esto da como resultado una gran cantidad de reglas, incrementando el costo computacional, pero pueden reducirse en función de:

- Coverage: número de objetos que la asociación predice correctamente.

- Accuracy: número de objetos que la asociación predice correctamente expresado como una proporción de todos los objetos para los cuales se aplica.

Generalmente, se especifican los valores mínimos de éstos y se buscan solo aquellas reglas donde el *Coverage* y *Accuracy* sean al menos estos valores determinados previamente.

Debe tenerse en cuenta que aunque dos cosas ocurran próximas entre sí, no garantiza que la relación sea importante, por lo tanto, luego de identificar una asociación se deberá emplear otro método analítico.

15.4 Redes Neuronales

Las *Redes Neuronales* son modelos no lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

La *Red Neuronal* trabaja asignando pesos a cada variable independiente y luego determinar si existen patrones en los datos.

Cuando se halla un patrón, la *Red Neuronal* lo optimiza por reasignar nuevos pesos a las variables según el grupo de validación, con el fin de determinar si es capaz de "adivinar" correctamente como se asignaron tales pesos.

La *Red Neuronal* continúa con este proceso aprendiendo los resultados una y otra vez como el cerebro humano, y luego puede aplicar el modelo resultante a cualquier set de datos.

La *Red Neuronal* puede manejar variables continuas y categóricas simultáneamente como también datos no lineales y colineales.

La interpretación de los resultados no es simple, porque no se tiene la ecuación del modelo a diferencia de la *Regresión Lineal* y *Logística*.

15.4.1 Aprendizaje Supervisado

La *Red Neuronal* consta de "*nodos*" interconectados por enlaces de excitación e inhibición y según las entradas a la red se puede activar cualquier número de nodos al mismo tiempo.

Cuando el sistema es entrenado con un set de objetos como entradas del mismo, el aprendizaje se encuentra en un *Modo Supervisado*.

En cada prueba se presenta una entrada que activa ciertos nodos y el sistema genera una salida basada en el patrón de activación. Si la salida es incorrecta se la puede modificar por realimentación.

Cuando el sistema ha aprendido la respuesta correcta que corresponde al set de entrenamiento, la *Red Neuronal* ya se puede usar para detectar y clasificar patrones entre nuevas entradas.

Como no existen nuevos descubrimientos, estos sistemas no ejecutan *Data Mining*, ya que se determina previamente las entradas en el set de entrenamiento, las salidas permitidas y el trazado correcto entre ambas.

Como la *Red Neuronal* no descubre esto, puede definirse, en este caso como un sistema automático de clasificación de patrones.

El aprendizaje en un *Modo Supervisado* puede considerarse como una forma de *Aprendizaje Automático (Machine Learning).*

15.4.2 Aprendizaje No Supervisado

En este caso, no se requiere la definición previa de las respuestas permitidas y sus entradas.

La red forma el set de salida durante el entrenamiento, basándose en características que ella misma extrae.

Una aproximación que se utiliza es una mapa de representación que está formado por dos capas, una para la entrada y otra para la salida, y los nodos de la capa de salida pueden tener conexiones de excitación e inhibición con otros nodos vecinos.

Utiliza un proceso de activación competitiva, donde para una entrada los nodos de la capa de salida que se activan, intentan inhibir a los otros.

El nodo o set de nodos que responde más fuertemente a la entrada es el "*ganador*", pero está a su vez conectado a una red de nodos que también tienen un alto grado de respuesta a la misma entrada. Así, entradas similares activan el mismo vecindario de nodos.

Operar de esta forma le permite a la red determinar cuales características se pueden utilizar para separar las entradas en grupos o categorías, sin necesidad de especificarlas antes de tiempo.

Cuando la red trabaja en este modo, no es necesaria formular hipótesis previas acerca de lo que se espera encontrar y por lo tanto puede utilizarse en análisis exploratorios. El aprendizaje en un *Modo No Supervisado* puede considerarse dentro del campo de *Data Mining*.

15.5 Algoritmos Genéticos

Los *Algoritmos Genéticos* están enfocados a la optimización más que a la clasificación y predicción como los otros métodos.

Pueden definirse como técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural.

15.5.1 Selección

El proceso de selección empleado por estos algoritmos es análogo al proceso de selección natural que ocurre durante la evolución y está fundamentado en el principio de supervivencia, donde aquellos individuos que se encuentran mejor adaptados al medio ambiente son los que sobreviven, transfiriendo el material genético a la próxima generación.

La selección de la nueva generación se realiza al azar, donde aquellos individuos con mayores valores de resistencia pueden reproducirse, asegurando que la información se transfiera a la siguiente generación.

En general, la población completa de individuos se reemplaza en cada ciclo, pero el tamaño de ésta se mantiene constante.

15.5.2 Combinación

Durante la combinación se unen o "*aparean*" dos individuos de la población elegidos al azar.

El individuo resultante contiene replicaciones parciales de la información que poseen sus padres.

En los *Algoritmos Genéticos*, todos los individuos de una población se representan como vectores en función de los registros establecidos en el set de datos.

Cuando se produce la combinación se crea un nuevo individuo por la fusión de los subset de información contenidos dentro de los vectores de ambos padres.

La cantidad de información que se replica en el nuevo individuo se determina por el punto de quiebre, el cual puede establecerse externamente.

15.5.3 Mutación

Si la información se recibe por combinación, la transferencia de ésta desde los padres al hijo es perfecta.

Sin embargo, el proceso de evolución sucede a veces por medio de mutaciones imprevistas en la composición genética de los individuos.

Las mutaciones pueden ocurrir de forma natural, cuando se produce un error en la transmisión de la información genética, lo cual puede tener buenos o malos efectos.

Se pueden realizar mutaciones en *Algoritmos Genéticos*, donde los nuevos individuos sufren un cambio en algunas de sus partes componentes del vector.

Generalmente, se prefiere realizar mutaciones pequeñas, ya que generan grandes efectos en las nuevas generaciones.

16. Presentación de Resultados

La presentación de resultados es el último paso, pero críticamente importante, ya que si luego de la finalización del análisis no se obtienen resultados o bien no se comunican efectivamente, dicho análisis no tiene ninguna utilidad.

Es preferible presentar poca cantidad de información, pero importante, para evitar posibles distracciones del enfoque del problema.

Ejemplo 20: Presentar la siguiente red de un problema en particular provocaría seguramente una perturbación de los objetivos que el análisis persigue.

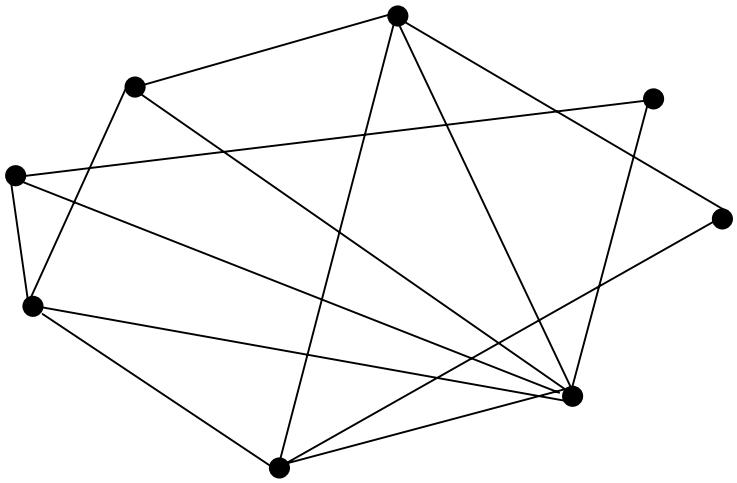

Figura 32. Red del Problema Real.

En cambio, si se presenta como resultado el siguiente gráfico, que es una simplificación de la red del problema real, será más fácil de interpretar y sin posibilidades de distracción.

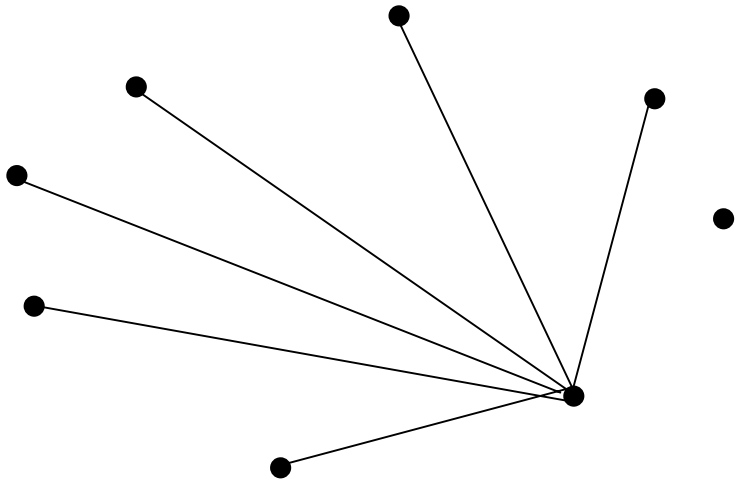

Figura 33. Red Simplificada.

Como regla general una presentación consta de:

- La descripción del problema original y los objetivos del análisis.
- El planteo de los distintos estados por el cual transita el análisis.
- Y por último la exposición de los resultados obtenidos.

También debe considerarse la posibilidad de una audiencia de presentación, la cual siempre debe estar enfocada a los intereses de las personas que conforman dicha audiencia.

Los resultados obtenidos deben estar justificados con documentación, y toda la información utilizada para llevar a cabo el análisis debe estar disponible para poder ser revisada y en un formato de fácil acceso e interpretación.

17. Anexo

Herramientas de Data Mining

Como el mundo de Data Mining afronta un cambio constante con el tiempo, es difícil mantener una variedad de productos y servicios disponibles para la industria. Sin embargo, existen varios sistemas en la actualidad que permiten resolver distintos tipos de problemas.

Una punto importante es el criterio para elegir la/s herramienta/s de Data Mining apropiadas para una aplicación dada.

Para ello, cabe aclarar que algunos sistemas se adaptan mejor y más rápido para ciertos tipos de problemas específicos y en el momento de elegir por alguno/s de ellos, debe considerarse entonces cual es el campo de aplicación y los objetivos perseguidos para cada aplicación en particular.

Tales herramientas pueden dividirse en:

- Sistemas de Análisis de Enlaces:

  NETMAP
  Analyst´s Notebook
  Imagix 4D
  Daisy
  ORION Systems
  Watson
  Crime Link

Estos sistemas están enfocados al análisis de enlace, donde se generan redes de objetos interconectados a través de sus relaciones, para descubrir patrones y tendencias en los datos.

Cuando se representan relaciones entre objetos, se obtiene una perspectiva diferente sobre como analizar los datos y los tipos de patrones que se pueden encontrar.

Tales metodologías permiten agregar más dimensiones al análisis, lo cual en otras formas de visualización no es posible.

Una desventaja de éstos sistemas es que el número de datos que se pueden representar es limitado comparado con otros métodos de visualización.

Los sistemas de análisis de redes fueron inicialmente utilizados en el mundo de las investigaciones (tales como aplicaciones legales, identificación de fraudes, lavado de dinero y otros delitos financieros), y las telecomunicaciones, pero actualmente pueden competir con otros sistemas en una gran variedad de aplicaciones comerciales.

- Sistemas de Visualización Panorámica:

MineSet 2.0
Metaphor Mixer
Visible Decisions 3D
Spotfire
Visual Insights
AVS/Express
IBM Visualization Data Explorer

Estas herramientas están enfocada a metodologías de visualización panorámica para llevar a cabo el análisis de los datos y permitir exponer patrones y tendencias interesantes.

En estos ambientes, los datos se establecen dentro de terrenos geométricamente limitados, y una característica importante es que la posición relativa de los datos dentro de tales terrenos se utiliza para representar información importante para el análisis.

En algunos de éstos sistemas es posible la integración de datos históricos con entradas de datos en tiempo real, y por lo tanto provee una forma de mantener Data Mining en tiempo real (por ejemplo, Metaphor Mixer – Visible Decisions 3D). Esto permite el reporte de los datos e incluso la toma decisiones, todo en tiempo real.

Estos sistemas encuentran el mayor dominio de aplicación en actividades financieras y bancarias. Además, para aquellos sistemas que constan con la posibilidad de poder ejecutar Data Mining en tiempo real, se amplía aún más el campo de aplicación.

- Sistemas enfocados a Análisis Cuantitativos:

Clementine
Enterprise Miner
Diamond
Gross Graphs
Grahp-FX
Temple MVV

Las aproximaciones cuantitativas se prefieren cuando se necesitan estimaciones de fiabilidad en un sentido estadístico.

Los análisis cuantitativos pueden proveer un resumen de la información acerca de diferencias de grupos o tendencias generales.

Para llevar a cabo el análisis no es necesario que los datos sean numéricos, ya que es posible utilizar técnicas de abstracción de datos.

Los diagramas cuantitativos pueden manejar extensos volúmenes de datos y permiten identificar tendencias lineales o exponenciales dentro de los sets de datos.

Estas herramientas tienen mayor aplicación en el estudio de tendencias de ventas.

La división de los sistemas en éstas tres categorías no significa que la aplicación de alguna de ellas excluye a las demás; existen algunos casos donde estas categorías pueden estar combinadas para formar aproximaciones híbridas que son muy útiles.

18. Bibliografía

- "Data Mining Solutions, Methodos and Tools for Solving Real-Work Problems" (Christopher Westphal – Teresa Blaxton)
- "Practical Machine Learning Tools and Techniques with Java Implementations" (Jan. H. Witten – Eibe Frank)
- Cátedra de Orientación I. Tema 5: Data Mining.