INTRODUCCIÓN AL DATA MINING

TABLA DE CONTENIDOS

HISTORIA

Desde siempre se reunieron datos para explicar fenómenos naturales, dado que su análisis proporciona instrumentos que facilitan la toma de medidas.

El data mining se apoya en distintas disciplinas, para solucionar problemas específicos de análisis como:

- Modelado de procesos biológicos: redes neuronales y algoritmos genéticos
- IA: Razonamiento basado en memoria
- Teoría de grafos: análisis de link
- Estadística
  - Muestras
  - Regresión: interpolar y extrapolar observaciones
  - Correlación.

DEFINICIÓN DE DATA MINING:

"EXPLORACIÓN Y ANÁLISIS, POR MEDIOS AUTOMÁTICOS O SEMIAUTOMÁTICOS, DE DATOS PARA DESCUBRIR PATRONES Y REGLAS"

EJEMPLOS DE USO

Qué productos ubicar en una promoción?

Por qué una persona responde a una oferta?

Qué nuevo producto querrá un cliente?

Qué particularidad tienen los clientes que cierran sus cuentas?

A quién asignarle un crédito?

AREAS DE APLICACION

- Inteligencia (FBI): Detección de patrones de conducta criminal
- Comercial:
  - Detección de los productos que garantizan la fidelidad de los buenos clientes, para mantenerlos aunque no sean rentables
  - Detección de clientes que cuestan más de lo que generan, para perderlos
  - Desarrollo de conocimiento de grupos de consumidores, de manera que cuando se capta un nuevo consumidor, por integración al grupo, se sabe qué ofrecerle
- Medicina
  - Grupos de patologías
  - Detección de riesgo de una enfermedad

CASOS

EJEMPLO 1: DATA MINING Y EL FRAUDE DE LAS TARJETAS DE CRÉDITO

El uso fraudulento de tarjetas de crédito supone un coste de miles de millones de dólares anuales para el sistema bancario y la economía mundial. Pese a las numerosas medidas ensayadas para combatirlo, la cantidad y sofisticación de este tipo de delitos aumenta cada año, superándose sistemáticamente las medidas anti-fraude. Generalmente, los bancos emisores disponen de sistemas que realizan algún tipo de comprobación de las transacciones, utilizando sencillas reglas si—entonces. El problema de estos sistemas es que, aunque intuitivamente se sepa que ciertas reglas detectan el uso irregular de una tarjeta, normalmente resulta imposible expresarlas con validez empírica. En consecuencia, el banco a menudo se enfrenta al dilema de identificar erróneamente una tarjeta como fraudulenta cuando en realidad no es el caso, lo que implica el riesgo potencial de deteriorar la relación con el cliente. El sistema desarrollado en este proyecto se basa en la hipótesis de que un usuario no autorizado utiliza una tarjeta de forma cualitativa y cuantitativamente diferente de como la ha utilizado anteriormente el usuario legítimo. Factores como la frecuencia de empleo de la tarjeta, el tipo y situación de los comercios en que suele utilizarse, hasta qué punto el usuario respeta su límite de crédito... forman en conjunto una "huella" que puede identificar de forma unívoca al usuario legítimo. Una ruptura de estos patrones puede utilizarse como indicador para detectar si otra persona está utilizando la tarjeta de forma fraudulenta. Es vital que los sistemas bancarios sean capaces de reconocer dichas violaciones de los esquemas típicos lo más pronto posible. Esencialmente, el infractor es un usuario de los servicios del banco, si bien un usuario indeseable. Aplicando técnicas de minería de datos, puede diferenciarse claramente su comportamiento del de los clientes normales. En este proyecto, el análisis mediante métodos de clustering borroso de una serie de datos sobre el titular de la tarjeta, el comercio y la transacción ha permitido:

- Definir las características que, combinadas, caracterizan los diversos tipos de fraude.
- Diferenciar el uso fraudulento del normal.

Como conclusión de los resultados obtenidos, se proponen una serie de recomendaciones y pautas de supervisión del uso de las tarjetas en tiempo real, entre ellas el análisis de comercios (para detectar aquéllos sospechosos de colaborar con los autores del fraude); la necesidad de consolidar la información sobre transacciones fraudulentas, lo que proporcionaría una visión más exacta del problema; y el análisis de vulnerabilidad, que permitiría revisar más exhaustivamente las transacciones cuando el riesgo es máximo puesto que está demostrado que las tarjetas son particularmente vulnerables en ciertas condiciones.

Bibliografía: http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/DAEDALUS-MD16-Tarjetas_Credito.pdf

EJEMPLO 2: DATA MINING Y LOS PRÉSTAMOS.

Un ejemplo popular de minería de datos está utilizando el comportamiento pasado para clasificar a los clientes. Estas tácticas han sido empleadas por las compañías financieras durante años como un medio para decidir si aprueba o no los préstamos y tarjetas de crédito. Si bien este ejemplo puede venir del sector financiero, las empresas en todas las industrias se pueden utilizar métodos similares para identificar a sus clientes más valiosos.

Bibliografía: http://www.the-modeling-agency.com/data-mining-examples.html

POR QUÉ AHORA?

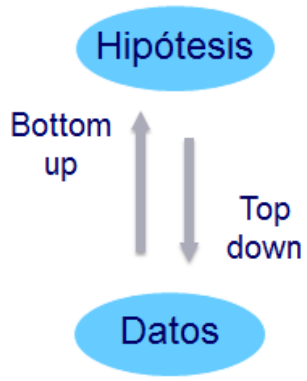
Porque existe mucha información importante, relevante y oportuna. Que por su volumen, hace imposible su análisis manual.

Alguno de los motivos por el cual surgen las herramientas de Datamining son:

- Grandes bases de datos de los sistemas transaccionales y los grandes Datawarehouse.
- Mayor poder computacional
- Mayor presión de la competencia convierte cualquier negocio en un negocio de servicio
- Existen productos comerciales que incorporan algoritmos de DM

MODOS DE APLICACIÓN

Busqueda de conocimiento (bottom-up)



Testeo de hipótesis (top-down)

TESTEO DE HIPÓTESIS (TOP-DOWN)

Descubrir conocimiento, sin asumir nada a priori (bottom-up), en forma directa (explica o categoriza un dato) o en forma indirecta (encuentra patrones sin clases predefinidas)

ETAPAS:

- Generar buenas ideas (Hipótesis Nulas):
  - Claro planteo del problema
  - Reuniones conjuntas
- Determinar los datos necesarios para el testeo
  - Armar lista completa de requerimientos para cada hipótesis
- Ubicar los datos necesarios
- Preparar los datos para el análisis
- Crear el modelo
- Aplicar el modelo a los datos
- Evaluar para confirmar o rechazar la hipótesis
  - Interpretar usando conocimiento analítico y del negocio.

BÚSQUEDA DE CONOCIMIENTO (BOTTOM-UP)

Descubrir conocimiento, sin asumir nada a priori (bottom-up), en forma directa (explica o categoriza un dato) o en forma indirecta (encuentra patrones sin clases predefinidas)

BÚSQUEDA DE CONOCIMIENTO DIRECTO

SE EXPLICA EL PASADO PARA PREDECIR EL FUTURO

EJEMPLOS DE APLICACIÓN

- Existe un campo cuyo valor predecir
- Cuál es la permanencia prevista del cliente?
- Existe un registro a clasificar
- Qué nivel de seguridad en el pago ofrece el solicitante de crédito?
- Existe una relación a explorar
- Qué ventas suben si realizo descuento en los quesos?

ETAPAS

- Identificar fuentes de datos
    - Confiables
    - Existe DW? (datos limpios, integrados, históricos, verificados)
    - Transformar los datos de los sistemas OLTP en aptos para el análisis
    - Contar con datos con el objetivo a predecir conocido (IMPORTANTE)
- Preparar los datos para análisis
    - Cuántos? Cuántas variables independientes?
    - Agregar campos derivados que hacen obvias las relaciones que el analista conoce (índices, densidad, salario promedio, etc.)
    - Dividir los datos en tres grupos, para
        - Entrenamiento (aprende)
        - Testeo (se auto ajusta)
        - Evaluación (se mide el resultado)
- Construir y entrenar el modelo
- Testear el modelo

- En busca de mayor generalidad
- Evaluar el modelo
- Encontrar el porcentaje de error
- Aplicarlo a nuevos datos

BÚSQUEDA DE CONOCIMIENTO INDIRECTO

NO EXISTE CAMPO U OBJETIVO A PREDECIR

EJEMPLOS DE APLICACIÓN

- Qué productos se venden juntos?
- Cómo segmentar a los clientes?

ETAPAS

- Identificar fuentes de datos
- Preparar los datos para análisis
- Construir y entrenar el modelo
- Testear el modelo
- Evaluar el modelo
- Aplicarlo a nuevos datos
- Identificar oportunidades para búsqueda de conocimiento directo
- Generar nueva hipótesis para testear

TAREAS DE DATAMINING

Un sistema de data mining actual realiza una o más de las siguientes tareas:

Clasificación:

Determina que un objeto, sujeto o evento pertenece a una clase predefinida.

Estimación:

Determina el valor de una variable continua y este valor se puede confirmar en el presente.

Predicción:

Es una clasificación o una estimación pero cuyos resultados se pueden confirmar en el futuro.

Grupos de afinidad o Asociación:

Determina si un evento o hecho está vinculado con la ocurrencia de otro.

Clustering:

Determina grupos homogéneos de objetos, sujetos o eventos de acuerdo a sus características.

Descripción:

Describe o explica las características de un suceso o las reglas que justifican o dan origen un determinado comportamiento.

HERRAMIENTAS

A modo de introducción nombraremos las principales herramientas que serán vistas en detalle durante la cursada:

- Cluster
- Razonamiento basado en memoria
- Canasta de mercado
- Árbol de decisión
- Redes neuronales
- Algoritmos genéticos

La elección correcta de la mejor herramienta para cada caso, conociendo qué esperar de cada una ellas es fundamental para el éxito de un proyecto de data mining. Otro factor importante de éxito, será la clara definición de una metodología para la selección y preparación de los datos, así como la evaluación del resultado devuelto por cada una de dichas herramientas.

EL CICLO VIRTUOSO

El data mining es un paso en el proceso de aplicar conocimiento que surge de entender a clientes, mercado, competidores. El foco del mismo se encuentra basado en la acción basada en el conocimiento, es decir, se busca resultado ACCIONABLE

Los algoritmos son importantes, pero deben ser aplicados en el área justa y sobre datos justos.

CICLO VIRTUOSO = técnicas + poder computacional + datos justos + área justa

ÁREA JUSTA:

- Procesos que descansan en análisis previo
  - Cuando se está planeando marketing de un producto
  - Al fijar política de precios
  - Al querer comprender el deterioro de la relación con los clientes
  - Al querer saber por qué las ventas son mejores en una sucursal que en otra
  - Al analizar si gastar + o – en soporte a clientes
- Observaciones informales (buenas preguntas sobre los datos)
- Identificación formal de áreas, si los usuarios ven la importancia y cooperan.

TOMAR ACCION:

Incorporar conocimiento al negocio, modificando los procesos propios

MEDIR:

- **Debo saber:**
  - Qué medir (EJEMPLO: Marketing por target)
    - Obtuve buenos clientes (VIP tener un perfil de cliente deseable)?
    - Obtuve clientes leales?
    - Que perfil demográfico tienen?
    - Compran productos adicionales?
    - Son rentables a largo plazo?
    - Los rentables son más leales que los otros?
    - Aumenta el ciclo de vida de los nuevos?
    - Siguen siendo leales cuando cae el incentivo?
    - Cuándo va a estar disponible los resultados
    - A quiénes transmitirles el resultado
    - El retraso en medir resultados es GRAVE.
      La medición debe estar planeada desde el principio.
    - Dificultades:
    - Problemas en transmitir/escuchar los resultados
    - Si use varias técnicas, comparar los resultados

INCONVENIENTES TÍPICOS, QUE NO PERMITEN LA IMPLEMENTACIÓN O INTERRUMPEN EL CICLO VIRTUOSO:

- Malos datos (incompletos)
- Datos que se borran
- Falta de funcionalidad de los OLTP de origen
- Resistencia de sectores a cambiar sus políticas
- Oportunidad (fuera de tiempo ya no será accionable)

EJEMPLO (CASO DE BANCO CON PÉRDIDA DE CLIENTES):

ÁREA JUSTA: Banco, con pérdida de clientes

ALTERNATIVAS:

- Traer nuevos???? Es caro.
- Aumentar tasas???? Se gana clientes volátiles
- Suspender servicios no rentables???? Si son los que usan sus clientes más fieles?
- Ofrecer más servicios???? Serán interesantes para todos?
- Entrevistar a los clientes que se fueron??? NO, no son honestos, no recuerdan todos los motivos....

SOLUCION: DM + experiencia del banco

EXPERIENCIA DEL BANCO:

- Uso de herramientas sin suficiente poder predictivo
- Clientes que hicieron reclamos de ajuste de cuenta se fueron
- Clientes que bajaron saldo en los últimos tres meses se fueron

USO DE DM:

- Ver las diferencias entre los clientes que se fueron y los que se quedaron
- Encontrar un grupo de riesgo

USO DM (TESTEO DE HIPÓTESIS):

Generar grupalmente las buenas ideas

- *Reducción de saldo y transacciones*
- *Uso creciente de cajeros*
- *Uso creciente de cajeros en una dirección*
- *Depósitos mensuales cesan*
- *Cesan transacciones*
- *Cargos que antes no se realizaban*
- *Banco rechazó una ampliación de crédito*
- *Cliente toma préstamo a menor tasa en otro banco*
- *Esperas prolongadas en atención al cliente*
- Ver si los requerimientos de datos se satisfacen con los sistemas OLTP o requieren fuentes adicionales de datos
- Ubicar datos y desistir de hipótesis para las que no hay datos
- Preparar los datos para el análisis, agregando campos que expresan relaciones
- Elaborar/testear el modelo sobre distintos segmentos creados
    - Correlaciones?
    - Redes neuronales?
    - Segmentación?
- Evaluar el modelo en cada grupo; ver si las correlaciones encontradas son significativas.
- Crear nuevas hipótesis

ACTUAR: Tomar las acciones acorde al resultado de las herramientas

MEDIR: Medir el impacto de las acciones tomadas y retroalimentar el proceso

MIDIENDO LA EFECTIVIDAD

- Cuál es mi objetivo?
- Cómo lograrlo?
- Cuál es la ventaja de lograrlo?

OBJETIVOS IDEALES

- Descubrir patrones interesantes
- Conocer más de mis clientes
- Aprender cosas útiles

OBJETIVOS MEDIBLES

- Identificar clientes dispuestos a renovar la suscripción
- Rankear clientes según la propensión a esquiar
- Listar productos que se van a ver afectados si discontinúo la venta de vinos

OBJETIVOS SEGÚN TIPO DE MODELO

MODELOS DESCRIPTIVOS

- Objetivo: Conocer, entender, explicar.
- Optar por menor precisión (70% con 5 reglas versus 75% con 48 reglas)

MODELOS PREDICTIVOS

- Objetivo: Clasificar o predecir.
- Requiere más precisión
- No descuidar la explicabilidad

OBJETIVOS PARA COMPARAR RESULTADOS

- Cuan preciso es el modelo? Qué grado de confianza le doy a las predicciones?
- Qué bien describe la realidad?
- Cuán comprensible es?

EL TODO O LAS PARTES?

- Precisión del modelo como un todo (% de registros clasificados correctamente)
- Precisión de una predicción en particular (cada hoja de un árbol tiene distinta certeza)

MIDIENDO EFECTIVIDAD DEL MODELO

- Modelo descriptivo: MDL (mínima longitud de la descripción)
- Modelo predictivo: por su comportamiento sobre una muestra
    - clasificación: % de error
    - estimación: diferencia entre valor predicho y real

CONCEPTOS

Población: Todo el grupo de estudio (Universo)

Muestra: Porcion del Universo que representa al mismo

Muestra sesgada: Muestra seleccionada con método no random

Rango de la muestra: diferencia entre el valor máximo y el minimo de la muestra

(Rango de la Muestra = R = Valor Mayor menor - valor menor)

Media (Aritmetica): Promedio = ∑x/nx

Mediana: Valor central de un serie, si es impar, existe el valor, sino lo es, se promedian los valores centrales (menos afectada por observaciones que se alejan de la media).

Moda: Valor que mas veces aparece mas veces en una serie

Distribución: Gama de valores que pueden representarse como resultado, con lo cual entendiendo la distribución , podemos predecir valores futuros (normal, chi cuadrado, poisson).

Desviación: diferencia entre un valor dado y la media

Varianza: promedio del cuadrado de las desviaciones

Desviación standard: raíz cuadrada de la varianza

Correlación: relación que existe entre el valor de una variable y otra

Regresión: predecir una variable en función del valor de otra

Ejemplos:

A modo de ejemplo y para simplificar los conceptos, aplicaremos cada uno de los mismos a los posibles resultados de un dado:

Media = (1+2+3+4+5+6)/6 = 3,5

Varianza = [(1-3,5)²+ (2-3,5)²+ (3-3,5)²+ (4-3,5)²+ (5-3,5)²+ (6-3,5)²]/6 = 2,92

Desviación estandar = $\sqrt{Varianza} = \sqrt{2,92} = 1,71$

Confianza de una relación entre dos elementos: frecuencia con que la relación es cierta en la muestra de prueba

Distancia: proximidad que permite detectar vecinos

Xi cuadrado: Prueba no paramétrica que mide la distribución observada y una teporica (bondad de ajuste). Cuanto mayor el valor de X^2, menos verosímil es que la hipótesis sea correcta. Estudia la asociación entre variables cualitativas

- con 1 variable: la bondad del ajuste de una muestra a una población
- con 2 variables:
  - la homogeneidad entre dos muestras
  - el vínculo entre dos variables de una población o contraste de independencia

Contraste de independencia: Se utiliza para analizar y contrastar las dependencias de características entre individuos o de variables entre si.

Se utiliza la tabla de contingencia y se cargan los valores

Ejemplo: Se desea conocer, si en una determinada población, existe una relación entre el sexo de sus individuos y su tendencia política)

- Se construye una tabla de contingencia de r filas y c columnas
- Se calcula la frecuencia esperada como
  - eij= ri* cj/n
- Se compara las frecuencias esperadas (eij) con las encontradas (n ij) y calcula el estadístico como SS (nij -eij)**2 /eij

COMO MEDIR LA EFECTIVIDAD DE UN MODELO

- Hay muchas metodologías
- Cómo comparo los resultados de modelos de distintos tipos?

LIFT, que deberá transformarse en ROI

HERRAMIENTAS PARA MEDIR Y COMPARAR MODELOS

Soporte: Cantidad de ocurrencias de un evento en la población

Confianza: Probabilidad de encontrar la parte derecha de la regla (consecuente) condicionada a que se encuentre la parte izquierda (antecedente)

Lift: Medida que indica la performance de una predicción o clasificación de un modelo dado. Es una función decreciente del tamaño de la muestra sesgada.

Fórmula:

$$\frac{\text{Incidencia de la clase a seleccionar en muestra sesgada}}{\text{Incidencia de la clase a seleccionar en población}}$$

Ejemplo:

Supongamos que contamos con los siguientes valores y tratamos de identificar la regla la mejor de las reglas identificadas en el set de datos:

N° Evento	Valor antecedente	Valor consecuente
1	A	0
2	A	0
3	A	1
4	A	0
5	B	1
6	B	0
7	B	1

Encontramos dos reglas candidatas:

Regla 1: A => 0

Regla 2: B => 1

Soporte

Soporte Regla 1 = cantidad de veces que ocurre “A=>0” / total de eventos = $\frac{3}{7}$

Soporte Regla 2 = cantidad de veces que ocurre “B=>1” / total de eventos = $\frac{2}{7}$

Confianza

Confianza Regla 1 = cantidad de veces que ocurre “A=>0” / cantidad de veces que ocurre “A”= 3 / 4

Confianza Regla 2 = cantidad de veces que ocurre “B=>1” / cantidad de veces que ocurre “B”= 2 / 3

Lift

Lift Regla 1 = Confianza Regla 1 “A=>0” / Soporte (antecedente) “A”= $\frac{\frac{3}{4}}{\frac{4}{7}}$

Lift Regla 2 = Confianza Regla 2 “B=>1” / Soporte (antecedente) “B”= $\frac{\frac{2}{3}}{\frac{3}{7}}$

Conclusion: Aunque notamos que la regla 1 tiene mayor confianza, el lift demuestra que deberíamos seleccionar como mejor la regla 2.

MODELO

Un modelo es una **simplificación** que imita los fenómenos del mundo real, de modo que se puedan comprender las situaciones complejas y podamos hacer predicciones.

LOS MODELOS PUEDEN ESTAR DETERMINADO POR:

Los datos del input:

- input continuo: regresión (Red Neuronal)
- variables categorizables: Árbol

o por el output:

- categorización
- variable continua
- cluster

RIESGOS EN LA CREACION DEL MODELO:

- Sobreestimar los datos
  - La predicción está influenciada por los datos del training
  - El resultado es redundante
- Subestimar los datos
  - eliminar campos predictivos
- Poca explicabilidad (cluster)

HERRAMIENTAS

Relación entre tareas del Datamining y sus distintas herramientas

	Classification	Estimation	Prediction	Affinity Grouping	Clustering	Description
Market Basket Analysis			Si	Si	Si	Si
Memory - Based Reasoning	Si		Si	Si	Si	
Genetic Algorithms	Si		Si			
Cluster Detection					Si	
Decision Trees	Si		Si		Si	Si
Neural Networks	Si	Si	Si		Si	

ARBOLES DE DECISIÓN

Los árboles de decisión son modelos que permiten clasificar, realizar predicciones y circunstancialmente estimaciones, aunque en este último caso carecen de precisión, ya que todos los registros que lleguen a una determinada hoja habrán alcanzado el mismo valor.
El objetivo de la herramienta es proveer un árbol que clasifique un set de datos en una de varias clases predefinidas. Existen distintos algoritmos para su construcción, entre los que se pueden citar al CART, CHAID, C4.5 e ID3.
El modelo se construye en base a un set de entrenamiento, un conjunto de registros que se encuentran preclasificados, y a partir del se construye el árbol, permitiendo descubrir el campo de un registro.

Cada camino desde la raíz a un nodo hoja es único, y constituye una regla utilizada para clasificar los registros.

La posibilidad de comprender el procedimiento utilizado por este método para clasificar, facilita la comprensión de las reglas generadas.
Para minimizar los errores de la herramienta se recomienda contar con sets de entrenamiento con abundantes casos clasificados.

Los árboles de decisión son ocasionalmente utilizados para producir una serie de salidas que abastezcan a otras herramientas de *data mining*.

MODALIDAD: directa

RESULTADO: reglas (que son el camino desde la raíz al nodo hoja) que clasifican o predicen

EXPRESIVIDAD: alta

METODOS:

- CART (Clasification and Regression Tree): Árboles de clasificación y regresión que provee un conjunto de reglas que se pueden aplicar a un nuevo conjunto de datos (sin clasificar) para predecir cuales registros darán un cierto resultado. Segmenta un conjunto de datos en dos divisiones y requiere menos preparación de datos que CHAID).

- CHAID (Detección de interacción automática de Chi Cuadrado, idem que pero con Chi Cuadrado para crear múltiples dimensiones y requiere mas preparación de cados que CART).

- C4.5 (Produce árboles con n variable de rama por nodo: diferente tratamiento de variables categóricas (arbustos). Posee poda pesimista puede reemplazar el sub-arbol por una de sus ramas)

MEDIDAS:

- MEDIDAS POR NODO: Número de registros entrantes

- MEDIDAS POR NODO HOJA: Porcentaje de error: (1- porcentaje de clasificaciones correctas). Peso: probabilidad de que un nuevo dato termine en ese nodo

- MEDIDAS DEL ARBOL: porcentaje de error: sumatoria del error de cada hoja multiplicado por su peso

FUNCION: (que se busca maximizar al crecer el árbol, al elegir el campo del split) diversidad (antes del split) – [diversidad (hijo derecho) + diversidad (hijo izquierdo)]

PRUNNING: técnica para eliminar ramas del árbol con bajo poder predictivo, que no hacen decrecer el porcentaje de error del árbol).

FORTALEZAS:

- Generan reglas entendibles.
- Clasifican con consumos bajos de recursos
- Manejan datos continuos y categóricos

DEBILIDADES:

- Inapropiado para estimar valores continuos
- Problemático para series temporales
- Generen errores grandes si se utilizan pocos datos de entrenamiento.

CLUSTER DETECTION

Cluster detection es una herramienta indirecta cuyo principal propósito es el de clasificar un conjunto de datos en grupos compuestos por objetos similares.

El algoritmo clasificará cada objeto dentro del grupo más representativo, acorde a las características de dicho objeto y reorganizará iterativamente la conformación de cada grupo en función de las mismas.

Este análisis será realizado identificando el centroide en cada iteración, el cual, será el objeto real más cercano a la media del *cluster*, es decir, será el objeto más representativo del mismo (mediana).

La cantidad de grupos puede ser predefinida (Algoritmo *k-means*) o puede ser la resultante del algoritmo (jerárquicos).

Algoritmos jerárquicos:

Este tipo de algoritmos puede ser TOP DOWN o BOTTOM UP, es decir que en principio cada objeto del conjunto es un *cluster* o todo el conjunto es un *cluster* y de manera iterativa se van definiendo grupos en base a la cercanía entre cada uno de los objetos y el centroide.

Algoritmo *k-means*:

En este algoritmo se predefine la cantidad de grupos en los cuales se separará el conjunto de datos. En la primera iteración los centroides serán seleccionados al azar y serán reemplazados por otros más precisos a medida que se recorre el conjunto de datos en cada iteración.

MODALIDAD: indirecta

RESULTADO: división del universo de datos en grupos

EXPRESIVIDAD: baja

METODOS: K-means, aglomeración (al incio cada elemento es un cluster). Generalmente es el primer método que se utiliza en todo proyecto de DM y luego se combina con algún otro algoritmo para profundizar el estudio (jerárquico, particional o genético).

FUNCIONES:

- DISTANCIA entre dos registros: se puede definir como distancia de puntos en el espacio o ángulo entre vectores o cantidad de campos cuyos valores coinciden
- DISTANCIA entre dos clusters: se puede definir como la de sus centros, la de sus puntos más cercanos o sus puntos más lejanos.
- VARIANZA: sumatoria de los cuadrados de la distancia al centro.
- PROBLEMAS: campos expresados en distintas unidades de medida

CRITERIO DE FORMACION DEL CLUSTER: la variable cuyo promedio de distancias es sensiblemente menor al promedio de distancia en la población total forma el criterio de agrupamiento.

FORTALEZAS:

- Buen desempeño con datos categóricos, numéricos o textuales.
- Fácil de utilizar

DEBILIDADES:

- Dificultad para elegir las medidas de distancias correcta y los pesos
- Sensibilidad a los parámetros iniciales
- Dificultad para interpretar los cluster resultantes.

CANASTA DE MERCADO

Canasta de mercado consiste en descubrir patrones a partir transacciones almacenadas que indican qué productos son comprados con qué otros productos. El resultado de la herramienta es un conjunto de implicaciones que siguen la forma X⇒Y, siendo la parte izquierda el antecedente y la derecha el consecuente, ambos ítems. Las mismas pueden interpretarse como una transacción que incluye un ítem X, entonces incluirá el ítem Y.

La herramienta surgió del análisis de las compras realizadas en puntos de ventas, siendo la principal aplicación el Análisis de Canasta de Mercado. Sin embargo, el descubrimiento de reglas de asociación no es exclusivo de éste.

Las reglas de asociación resultantes deben ser analizadas por el experto para evaluar su utilidad, pudiendo resultar en uno de los siguientes tipos:

Accionables: Son patrones comprensibles, que una vez descubiertos se pueden explicar fácilmente y utilizar para tomar decisiones.

Triviales: Son aquellos patrones que cualquier persona conocedora del área en análisis podría describir. Por ejemplo, quien adquiere una cafetera seguramente decida adquirir café.

Inexplicables: Son patrones que parecen no tener explicación y tampoco sugieren un curso de acción posible.

Existen dos formas de crear asociaciones. Una es con la generación de candidatos, utilizando en general el algoritmo APRIORI. Otra es sin generación de candidatos, habitualmente con el algoritmo *FP-TREE*.

Para evaluar la calidad de la reglas se emplean las medidas de confianza, soporte y *Lift*.

Los productos (ítems) pertenecen a una taxonomía, la cual fija una jerarquía. En ocasiones, cuando un producto aparece en pocos registros es conveniente agruparlo en niveles jerárquicos más altos. De todos modos, esto no implica que tal acción se deba llevar a cabo en todos los ítems, sino que se debería utilizar en el caso descripto o bien para reducir la cantidad de productos existentes en los registros.

Los ítems virtuales son aquellos que no aparecen en la taxonomía de productos, porque van más allá de la misma. Se recomienda solo utilizarlos cuando logren resultar en información accionable, ya que pueden ocasionar reglas triviales.

MODALIDAD: directa/indirecta

RESULTADO: reglas de asociación del tipo IF *condición* THEN *resultado "*Si A y B entonces C"

EXPRESIVIDAD: alta

PRUNNING: técnica para eliminar reglas poco útiles: se exige un soporte mínimo para cada ítem de la regla

FORTALEZAS:

- Resultado entendible
- Muy útil para DM indirecto
- Trabaja con datos de longitud variable
- Bajo nivel de cómputo.

DEBILIDADES:

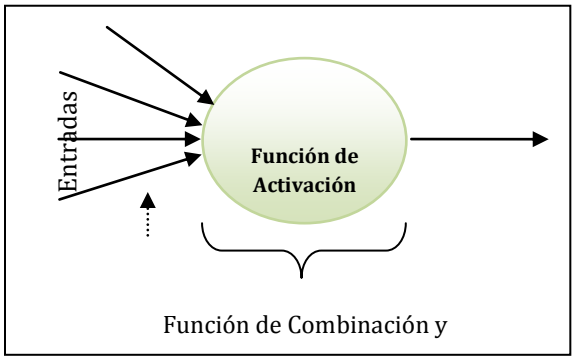
- Crecimiento exponencial (de cómputo y de datos)
- Soporte limitado para atributos de datos
- Los ítems correctos son difíciles de determinar
- Los ítems poco frecuentes son problemáticos

REDES NEURONALES

Las redes neuronales son una serie de neuronas artificiales conectadas entre sí que permiten realizar predicciones, clasificaciones y *clustering*. Cada neurona artificial tiene entradas que produce un único valor de salida. Las unidades están conectadas entre sí en forma similar a las neuronas naturales, sirviendo las salidas de algunas como entrada de otras. Las redes neuronales aprenden a través de ejemplos.

La neurona artificial, cuya estructura se puede observar en la figura 7, cuenta con una función de activación. La misma está compuesta por dos partes: la función de combinación y la función de transferencia. La primera, como su nombre indica, toma los valores de entrada y los combina en un único valor. Cada entrada tiene su propio peso, por lo que comúnmente se multiplica el valor de entrada por tal peso propio y luego se realiza la sumatoria de todos los productos. Ésta constituye la función de combinación más común. La segunda calcula el valor de salida a partir del resultado de la función de combinación.

La función de transferencia más común es la Función Sigmoidea. Dentro de este tipo, las más comunes para redes neuronales son la Función Logística y la Tangente Hiperbólica. La principal diferencia entre ambas es el rango de sus salidas, siendo entre cero y uno para la Logística y entre -1 y 1 para la Tangente Hiperbólica.



Representación de una Neurona Artificial

Existen distintos tipos de redes neuronales. Las redes unidireccionales son las más comunes y se utilizan para solucionar problemas de estimaciones, predicciones y clasificaciones. Otros tipos de redes son los mapas autoorganizados (utilizados para encontrar *clusters*) y las utilizadas para series de tiempo.

Las redes unidireccionales se estructuran en tres capas. La primera se denomina *input layer* o capa de entrada. Cada unidad en la capa está conectada a un campo de origen y copia el valor de entrada directamente a su salida. A nivel práctico representa el mapeo de los valores a un rango adecuado.

El valor *bias* es una constante utilizada en la entrada de las neuronas cuyo valor es 1 y posee su propio peso. El mismo permite un mejor desempeño de la red.

Se conoce como *hidden layer* o capa oculta a aquella capa de unidades neuronales que se encuentra entre otras dos capas de unidades y no conectada a las entradas ni salidas de la red. Si bien el incremento de las capas ocultas puede devenir en errores, permite aumentar el poder de la misma. Usualmente solo una capa oculta es necesaria.

Se recomienda empezar con un nivel en tal capa y luego comenzar a incrementar el número si la red no es muy precisa. En caso de sobre entrenamiento, reducir el tamaño de la capa.

La última capa se la denomina *output layer* o capa de salida, la cual, como su nombre indica, está conectada a la salida de la red neuronal. Se conecta a su vez con todas las neuronas de la capa oculta. La capa de salida puede tener una o varias neuronas artificiales, dependiendo de la cantidad de valores a obtener de la red. Usualmente se utiliza solo una neurona artificial en tal capa.

Existen variaciones de la topología, pudiendo tener solo dos capas, conectándose la capa de entrada directamente con la de salida.

No se trata de una técnica explicativa debido a que las neuronas artificiales se comportan como cajas negras[1]. De todas formas es posible averiguar la importancia relativa de las entradas a través de una técnica denominada análisis sensitivo.

El entrenamiento de redes neuronales es el proceso de establecer los mejores pesos en las entradas de las neuronas artificiales incluido el valor *bias*. Se recurre para ello a un set de entrenamiento.

Hay algunos aspectos en los que se debe hacer especial hincapié cuando se trabaja con redes neuronales. Los sets de entrenamiento deben contener gran cantidad de registros. A su vez, la preparación de éste debe ser realizada cuidadosamente, ya que los datos de los que se valen las redes deben pertenecer a un rango definido, y la información por ello debe ser mapeada. Cuando se trabaja con valores continuos la labor es relativamente sencilla, pero en casos como valores ordenados y discretos, categóricos u otros como fechas la labor a realizar para preparar el set de entrenamiento requiere un detallado análisis y trabajo.

Se recomienda limitar los valores de las entradas a un rango entre -1 y 1, para lo cual se debe recurrir al proceso denominado mapeo, que modifica los valores que servirían de entrada para adecuarlos al rango indicado, evitando así que valores muy altos dominen a otros muy bajos.

Así como los distintos valores de entrada deben ser mapeados para que la red funcione correctamente, los valores de salida también deben ser interpretados, realizando el mapeo de las salidas. En estos casos se lleva a cabo el proceso inverso al mapeo de los valores de entrada, siendo necesario en ocasiones recurrir a un set de validación distinto del set de entrenamiento para calibrar las salidas.

MODALIDAD: directa/indirecta

RESULTADO: Clasifica, predice o estima / construye grupos homogéneos (Clusters)

EXPRESIVIDAD: baja

FASES: dos fases de aplicación: entrenamiento y prueba. En la primera se usa un conjunto de datos para determinar los pesos (parámetros de diseño de la red). Una vez hecho esto, se pasa a la fase de prueba donde los patrones de prueba se procesan siendo la entrada habitual de al red y se analizan las prestaciones definitivas.

APLICACIONES: generalización: su objetivo es dar una respuesta correcta a la salida para un estímulo de entrada que no ha sido entrenado con anterioridad.

[1]El análisis de cajas negras es aquel que tiene en cuenta las salidas de un sistema pero no sus procesos internos.

FUNCIONES:

MODIFICACION DE VALORES CONTINUOS DEL INPUT: (dato –Mínimo) / (Máximo – Mínimo + 1)

MODIFICACION DE VALORES CATEGORICOS DEL INPUT: Se asigna una fracción del intervalo (0,1) a cada distinto valor

MODIFICACION DEL OUTPUT SI SE ESPERA CATEGORICO PESO DE CADA DATO DEL INPUT: Lo descubre la herramienta, con aprendizaje: al inicio son random.

BACKPROPAGATION: recalcula los pesos para ajustar el resultado; tiende a reaccionar lento a los cambios en dirección opuesta a cómo vienen dándose; controla que los cambios sean decrecientes

ACTIVACION: COMBINACION: suma ponderada del input

ACTIVACION: TRANSFERENCIA: función lineal, sigmoidea (f(x)=1/(1+e-x)) o Exponencial.

BUSQUEDA DE EXPRESIVIDAD: ANALISIS SENSITIVO: tomar el promedio de cada input y generar el output, luego ir modificando de a uno cada input y ver su impacto en el output.
MIEMBRO PROMEDIO (en clusters): para entender el cluster

UNIDADES DEL INPUT: cada una se conecta a un campo (una source); el output puede ser el input idéntico o modificado.

UNIDADES DE NIVEL OCULTO: cada unidad se conecta a todas las del input, las recibe con un cierto peso, transfiere la sumatoria a un output.

UNIDADES DEL OUTPUT: genera(n) la salida

FORTALEZAS:

- Son versátiles (se puede agregar cualquier tipo de datos)
- Pueden producir buenos resultados en dominios complicados
- Manejan tipos de datos continuos y categóricos
- Disponibles en varios paquetes de soft

DEBILIDADES:

- Todas las salidas o entradas deben estar convertidas a 0 y 1
- Los resultados son poco expresivos
- Probablemente convergen en una solución inferior

RAZONAMIENTO BASADO EN MEMORIA

El razonamiento basado en memoria se centra en la idea de similitud, es decir, compara los valores de los parámetros de un objeto actual con objetos con características y contexto similar del pasado y predice las consecuencias futuras del mismo.

Una particularidad de esta herramienta, a diferencia de otras del *data mining,* es que no se necesita dar formato a los registros para ser utilizados. Esto es particularmente útil en conjuntos de datos geográficos y solo requiere la existencia básica de dos operaciones: La función de distancia, capaz de calcular la distancia entre dos registros cualquiera y la función de combinación, capaz de combinar resultados de varios vecinos para arribar a una respuesta.

Una desventaja de la herramienta es el alto costo que implica obtener un resultado, ya que deberán ser leídos muchos registros históricos para identificar vecinos candidatos a ser clasificados.

La efectividad de la herramienta estará dada por la aplicación de algunos factores determinantes:

1) Correcta selección del conjunto de datos de entrenamiento: El mismo deberá estar balanceado e incluirá una cantidad adecuada de registros que representen las potenciales categorías que pudieran encontrarse en la población, para encontrar la cantidad suficiente de vecinos de nuevos registros no clasificados y así proveer una respuesta adecuada.

Es importante remarcar nuevamente que el conjunto de datos debe estar correctamente balanceado. Es decir, si el mismo posee muchos registros, probablemente brinde un nivel de clasificación óptimo, ya que encontrará muchos vecinos para inferir la respuesta adecuada. El gran problema sería la performance general del sistema (cada registro será comparado con el nuevo registro a clasificar). Por otro lado, un conjunto de datos de entrenamiento insuficiente proveerá una performance muy buena, pero los resultados serán demasiado pobres por falta de detalle.

Una potencial solución a este problema, en caso que las categorías estén lo suficientemente separadas, sería identificar los *clusters* que representen a cada una de ellas y utilizar los centroides de las mismas para conformar un conjunto de datos de entrenamiento reducido. Existen otros métodos para reducir conjuntos de datos de entrenamiento con categorías superpuestas y han sido motivo de recientes investigaciones. Una buena práctica es definir que un conjunto de datos debería caber dentro de una hoja de cálculo, de manera que cualquier equipo con características estándares pueda procesar los registros de manera aceptable.

2) Determinación de la función de distancia: La función de distancia es la encargada de medir la similitud entre objetos y determina cuán cercano o “similar” es un caso del set de entrenamiento respecto al caso a predecir.

Esta función puede basarse en valores únicos o matrices, las cuales podrán ser de distintos tipos de datos.

Cualquier función de distancia cuenta con cuatro propiedades claves a tener en cuenta:

Bien definida: La distancia entre cada punto será un número real positivo.

Identidad: La distancia entre un punto a si mismo siempre será cero.

Conmutatividad: La distancia entre dos puntos siempre será la misma, es decir que la distancia de A a B será la misma que de B a A.

Inequidad triangular: Las suma de las distancias de un punto intermedio C entre A y B no será nunca menor a la distancia entre los puntos A y B.

Cabe recordar que cada punto es en realidad un registro de una base de datos y en ocasiones se pueden obtener buenos resultados a pesar de no cumplir alguna de las propiedades antes descriptas.

El típico caso y más simple de demostrar es aquel que mide distancias métricas, pero una función de distancia podría medir muchos tipos de datos además de numéricos.

3) Selección de la cantidad de vecinos a considerar: La cantidad de vecinos son aquellos casos de similares características que dan soporte a la predicción, es decir, cuanto mayor sea la cantidad de vecinos mayor será la certeza y precisión de la predicción.

4) Determinación de la función de combinación: La función de combinación determina el valor a predecir en base a los votos de los vecinos preseleccionados, pudiendo ser:

Democrática: Todos los vecinos elegidos lo hacen con igual peso y el resultado tiene un % de certeza (100% si hay coincidencia).

Ponderada: La incidencia del voto es inversamente proporcional a la distancia al nuevo vecino (cuanto menor la distancia, mayor incidencia en el voto).

Un ejemplo simple de una función de combinación es una función de promedio, la cual predice una variable en base al promedio de los valores conocidos de esa variable en los vecinos preseleccionados.

MODALIDAD: directa

RESULTADO: Clasifica según alguna categoría, elige vecino más cercano

EXPRESIVIDAD: No presenta

FUNCIONES:

DISTANCIA: entre registros de la base; se calcula para cada campo y luego se suma para el registro (O se suma dividiendo por la máxima suma o...)

Ejemplos:

- Para valores numéricos |a-b| /máxima distancia

- Para sexo 0 ó 1

ELECCION DE VECINOS: cuántos?

COMBINACION: de los votos de los vecinos sobre la categoría a determinar

Democrática: todos los elegidos lo hacen con igual peso y el Resultado tiene un % de certeza (100% si hay coincidencia,....)

Ponderada: la incidencia del voto es inversamente proporcional a la distancia al nuevo vecino.

MEDIDAS: (Para el caso de asignaciones múltiples de categorías)

PRECISION: de los códigos que asignó la herramienta, que % corresponde a códigos correctos?
RECALL: del total de correctos a asignar, que % asignó? (pueden faltar)

FORTALEZAS:

- Resultados entendibles y claramente justificables
- Independiente de la representación de los datos
- Preformase independiente del training set (cualquier número de campos)
- Requiere mínimo esfuerzo de mantenimiento

DEBILIDADES:

- Requiere alto poder de cómputo en las etapas de predicción y clasificación.
- Requiere gran capacidad de almacenamiento para el training set (mientras más grande mejor el resultado)
- Los resultados dependen de la función distancia. Combinación y número de vecinos.

ALGORITMOS GENETICOS

Los algoritmos genéticos son herramientas de búsqueda y optimización basada en la teoría de la selección natural de Darwin, por la cual los individuos más aptos de una población son los que sobreviven, dado que se adaptan con mayor facilidad a los cambios en su entorno. Las diferencias entre cada individuo están en sus genes, los cuales son transmitidos a sus descendientes al reproducirse.

Consisten en funciones matemáticas o rutinas de software que toman los ejemplares de entrada y devuelven aquellos que deberían generar nueva descendencia para la próxima generación. Este proceso puede ser iterativo, realimentándose tantas veces como fue contemplado en su diseño.

Los algoritmos genéticos son principalmente utilizados en optimizaciones, dado que analizan las posibles soluciones en base a una función de evaluación y deciden cuales serán las seleccionadas para la próxima generación.

Se supone que los individuos (posibles soluciones del problema) pueden representarse como un conjunto de parámetros (que se denominan genes), los cuales agrupados forman un conjunto de valores (a menudo referido como cromosoma).

Tomando como referencia el algoritmo genético simple, deberíamos considerar que la función de evaluación cumple un papel fundamental, ya que es la que evalúa si un individuo será apto o no para reproducirse, por esto debe ser diseñada para cada problema de manera específica. Dado un cromosoma particular, la función de evaluación le asigna un número real, que se supone refleja el nivel de adaptación al problema del individuo representado por el cromosoma.

Durante la fase reproductiva se seleccionan los individuos de la población para cruzarse y producir descendientes, que constituirán, una vez mutados, la siguiente generación de individuos. La selección de padres se efectúa al azar usando un procedimiento que favorezca a los individuos mejor adaptados, ya que a cada individuo se le asigna una probabilidad de ser seleccionado que es proporcional a su función de evaluación.

La población inicial, es decir la cantidad de individuos que alimentaran inicialmente el algoritmo, podrá ser seleccionada al azar y en general deberá existir un balance para que pueda hallarse la cantidad suficiente de soluciones, sin afectar la performance por el volumen de información a manejar.

El cruce de los individuos seleccionados en base a la probabilidad de reproducción devuelta por la función de evaluación se realizará por medio del cruce basado en un punto, donde se cortaran los “padres” en un punto y se unirán ambas partes de cada padre para formar nuevos individuos, derivando así en nuevas generaciones. Existen otros métodos de cruce que contemplan cruces de más de un punto que no serán considerados en este análisis.

Tal como sucede en la naturaleza, existen mutaciones, que en el caso de los algoritmos genéticos podrían traducirse en la introducción de mínimos cambios aleatorios en los genes de los individuos, lo cual podría o no producir mejorías al momento de ser evaluadas por la función de evaluación.

MODALIDAD: directa

RESULTADO: función optimizada (individuo mejor adaptado)

EXPRESIVIDAD: baja

FUNCIONES:

SELECCIÓN Y EVALUACION (supervivencia del más apto): Toma los más aptos para una primera selección y no utiliza el resto (fitness).

CRUZA: se genera una población nueva que intercambia material cromosómico (más aptos de la selección) y sus descendientes forman la siguiente generación.

MUTACION: cuando un sistema se ha detenido en una estructura genética no optima, o cuando el sistema se ha viciado de cadenas muy parecidas es necesario infiltrar mutaciones que reanimen el sistema. En la mutación no permite la estabilización de poblaciones en soluciones locales.

CLONACION: La clonación consiste en la duplicación de la estructura genética de un cromosoma para la generación siguiente.

INVERSION: La inversión consiste en la operación contraria a la clonación.

REVERSION: La reversión consiste en la operación que cambia el cromosoma por sí mismo ordenado de atrás para adelante.

FORTALEZAS:

- Poseen buena capacidad de trabajo con cajas negras.
- Operan simultáneamente con varias soluciones.
- Utilizan operadores probabilísticas.
- Resuelve el problema de la estabilización por máximos locales.

DEBILIDADES:

- Es difícil seleccionar una función de adaptación adecuada (se utilizas redes neuronales con los pesos para solucionarlo).
- Pueden converger perpetuamente.
- Pueden tardar demasiado tiempo en converger.

CRISP

La metodología **CRISP** (**CRoss-Industry Standard Process for Data Mining**) es un proceso jerárquico formado por varias tareas que ofrece a las organizaciones la estructura necesaria para obtener mejores y más rápidos resultados en la minería de datos.

La organización está dada de la siguiente manera:

Fases:

- **Comprensión de negocio**
  - Determinar objetivos (de negocio y factores de éxito)
  - Entorno (restricciones, recursos, riesgos, costo-beneficio)
  - Determinar objetivos de data mining
  - Producir el plan de proyecto
- **Comprensión de datos**
  - Recolección inicial
  - Descripción
  - Exploración
  - Verificación de calidad
- **Preparación de datos**
  - Selección
  - Limpieza
  - Construcción de datos
  - Integración de datos
  - Formateo de datos

- **Modelado**
    - Elección de técnica (según: problema, requerimientos, restricciones)
    - Testeo del modelo
    - Construcción de modelos
    - Selección de modelo definitivo
- **Evaluación**
    - Evaluar resultado
    - Revisar procesos
    - Plantear nuevos pasos
- **Implementación**
    - Plan de implementación
    - Plan de monitoreo y mantenimiento
    - Reporte final
    - Documentación

TEXT MINING

La minería de texto (text mining), también conocida como minería de datos textuales o descubrimiento de conocimiento desde bases textuales no estructurales, pretende algo similar a la minería de datos: identificar relaciones y modelos en la información, pero a diferencia de la minería de datos, lo hace a partir de información no cuantitativa. Es decir, apunta a proveer una visión selectiva y perfeccionada de la información contenida en documentos, sacar consecuencias para la acción y detectar patrones no triviales e información sobre el conocimiento almacenado en las mismas.

Parte de grandes volúmenes de información en formato textual, datos que carecen de una estructura intrínseca, donde establecer relaciones es difícil de sistematizar.

Con text mining una organización, ante una situación determinada, puede saber no solo qué ha ocurrido; sino que también dirá por qué, permitiendo así tomar mejores decisiones de cara al futuro.

Las organizaciones que implementen esta tecnología obtendrán los siguientes beneficios:

- Conocer el significado de un texto sin la necesidad de leer el texto por completo.
- Poder leer resúmenes exactos de los textos antes de hundirse en documentos enteros.
- Navegar eficientemente en grandes bases con textos no estructuradas.
- Realizar una recuperación efectiva de datos en el idioma original.

POBLEMATICA: el lenguaje natural es ambiguo. Posee conocimiento sintáctico, semántico y retórico que esta implícito y es difícil de capturar. Además, de errores de errores de tipeo, sinónimos y múltiples acepciones nos encontramos con grandes masas de información en distinto idiomas.

METODOLOGIA: Dos etapas: un **pre-procesamiento**, donde los textos se transforma a un tipo de representación estructurada más sencilla de utilizar y luego la etapa de **descubrimiento**, donde se analizan las representación para descubrir patrones en ellos.

MODALIDAD: indirecto.

WEB MINING

Web mining se refiere al proceso global de descubrir conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web (Etzioni 1996). Es un campo multidisciplinar donde convergen áreas como la recuperación de información, el data mining, la estadística, la visualización de datos, lenguajes de etiquetas, tecnología web, etc. con el objetivo de descubrir redes de relaciones existentes en la WWW, utilizando su información desestructurada o semi-estructurada.

En otras palabras, el web mining es simplemente aprovechar las técnicas de Data Mining para obtener conocimiento de la información disponible en Internet.

Existen 3 clases de web mining:

1) **Web content:** se puede clasificar en:
   - Text Mining Si los documentos son textuales (planos)
   - Hipertext Mining: si los documentos contienen enlaces a otros documentos
   - Markup Mining: si los documentos son semi-estructurados (con marcas)
   - Multimedia Mining: para imágenes, audio y video
2) **Web Structure**: se intenta descubrir un modelo a partir de la topología de enlaces de red.
3) **Web Usages**: Se intenta extraer (hábitos y preferencias de los usuarios. O contenidos y relevancias de documentos) a partir de las sesiones y comportamientos de los usuarios y navegantes.

FORTALEZAS:

- Mejoran la performance del server.
- Permiten una reestructuración del sitio para optimizar su navegabilidad.
- Descubrir potenciales clientes de comercio electrónico.
- Identifica las horas pico de acceso y los lugares preferidos por los usuarios para permitir colocar publicidad estratégica.

DEBILIDADES:

- No hay forma exacta de determinar el inicio y el fin de sesión del usuario.
- No se tiene información sobre el acceso a páginas almacenadas en la caché local de los clientes de navegación.
- La información registrada puede ser ambigua si hay cambio de nombre de servidores o reubicación de páginas.