Tema1. Introducción a la Minería de Datos

Abdelmalik Moujahid

Grupo de Inteligencia Computacional
Universidad del País Vasco UPV/EHU
Curso 2014-2015

Índice

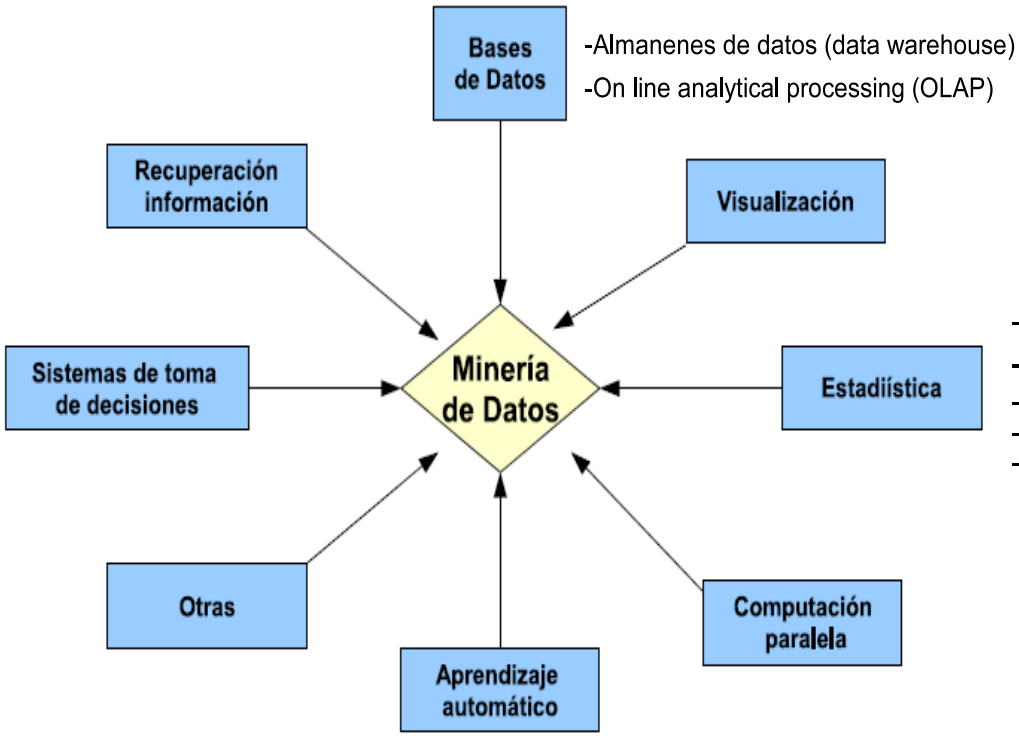
Índice

Minería de datos

Definición

- Data mining. Mineria de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten y Frank, 2000)
- Knowledge discovery in databases. Descubrimiento de conocimiento en bases como proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles, y en última instancia, comprensibles a partir de los datos (Fayyad y col. 1996)

Relación con otras disciplinas



Índice

Aplicaciones

Segmentación de clientes

Considera la información sobre el consumo de productos y servicios de los clientes para detectar grupos con necesidades y valor distintos, y analiza las dinámicas de vinculación y desvinculación de cada grupo para ayudar a definir:

- Objetivos comerciales por segmento.
- Estrategias comerciales por defecto para cada segmento (pautas de actuación tanto para venta como para retención)
- Creación de nuevos productos y ofertas

Aplicaciones

Sistemas de recomendación

Hace años que Amazon popularizó las recomendaciones en Internet, basandose en técnicas de data mining que haciendo uso de la información del consumidor, sus preferencias y comportamientos, son capaces de modificar contenidos y ofertas en tiempo real.
Actualmente, existen soluciones concretas basadas en sistemas de recomendación:

Aplicaciones

Sistemas de recomendación

- Next best activity (NBA):es un sistema de recomendación personalizada que permite asignar a cada cliente, de forma dinámica, la mejor oferta posible, basándose en las sendas de vinculación y desvinculación observadas en clientes similares en el pasado.
- AQUA SOCIAL NETWORKS: permite conocer y explotar comercialmente cómo se propagan la información, las opiniones y los comportamientos entre los individuos que forman parte de una comunidad familiar, de amistad o profesional.

Para más información:
`http://www.neo-metrics.com/webnm/`

Aplicaciones

Instituciones financieras

El **Falcon Fraud Manager** es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para intentar detectar y paliar el número de fraudes.

Para más información:
`http://www.fico.com/en/Pages/default.aspx`

Aplicaciones

Medicina

- Diagnóstico de enfermedades
- Detección de pacientes con riesgo de sufrir una patología concreta
- Tratamiento de imágenes médicas

Aplicaciones

Bioinformática

Análisis de datos de expresión genética para fines diversos como:

- Celección de genes relevantes,
- Caracterización de enfermedades,
- Predicción, localización de patrones de comportamiento genéticos.

Para más información:
`http://www.sc.ehu.es/ccwbayes/members/inaki/DM-applications.htm`

Índice

Minería de datos

Tipos de datos

- Datos estructurados (bases de datos)
- Datos en forma de grafos
  - World Wide Web
  - Estructuras moleculares
  - Redes sociales
  - etc.
- Otros tipos de bases de datos
  - Espaciales
  - Temporales
  - Secuencias genéticas

Minería de datos

Classification problem data matrix

	a_1	...	a_j	...	a_m	*Clase*
X_1	x_{11}	...	x_{1j}	...	x_{1m}	c_1
...	...	...	...	...	...	...
X_i	x_{i1}	...	x_{ij}	...	x_{im}	c_i
...	...	...	...	...	...	...
X_n	x_{n1}	...	x_{nj}	...	x_{nm}	c_n

Problema de clasificación supervisada.

Machine learning repository: `http://archive.ics.uci.edu/ml/`

Minería de datos

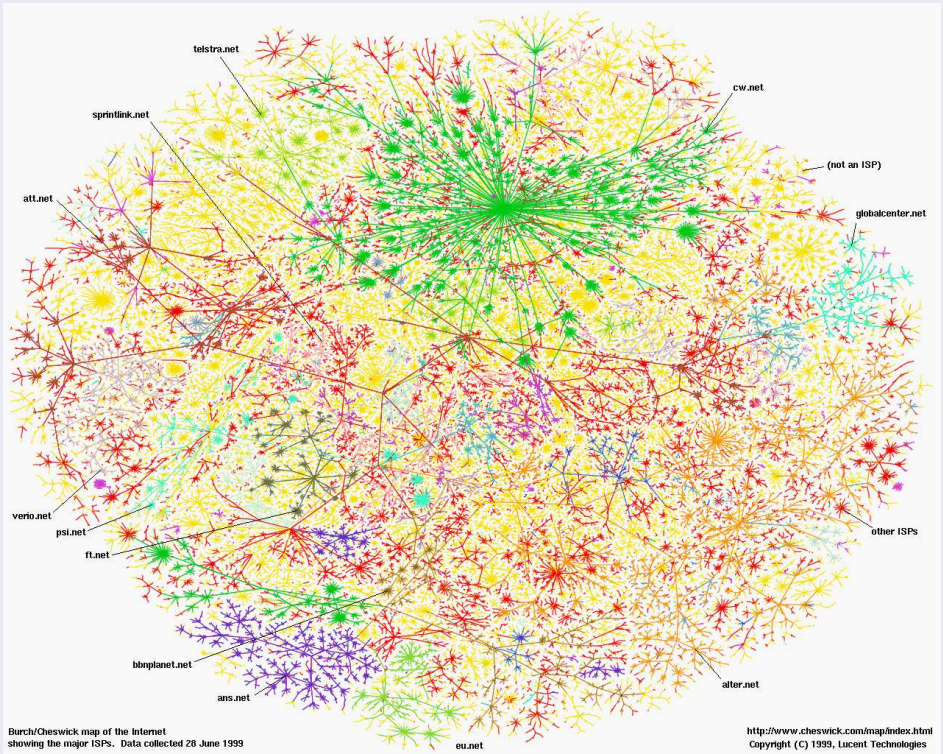
Gene expression data matrix

	condition 1	...	*condition* j	...	*condition* m
Gene 1	x_{11}	...	x_{1j}	...	x_{1m}
...	...	...	...	...	
Gene i	x_{i1}	...	x_{ij}	...	x_{im}
...	...	...	...	...	...
Gene n	x_{n1}	...	x_{nj}	...	x_{nm}

European Bioinformatics Institute:
`http://www.ebi.ac.uk/2can/databases/microarray2.html`
Gene Expression Atlas: `http://www.ebi.ac.uk/gxa/`

Minería de datos

Datos en forma de grafos



Network data sets: `http://www-personal.umich.edu/~mejn/netdata/`

Minería de datos

Datos en forma de secuencias

El ADN, es un ácido nucleico que contiene instrucciones genéticas necesarias para construir otros componentes de las células, como las proteínas y las moléculas de ARN responsables del desarrollo y funcionamiento de todos los organismos vivos (Wikipedia).
Por ejepmlo, *AAAGTCTGAC* es una secuencia de ADN compuesta por una sucesión de los 4 posibles nucleótidos representados por las letras A (Adenina), C (Citocina), G (Guanina), T (Timina).

Índice

Knowledge Discovery from Databases (KDD)

Fases del proceso iterativo e interactivo

1. Integración y recopilación de datos
2. Selección, limpieza y transformación
3. Minería de datos
4. Evaluación e interpretación
5. Difusión y uso

Knowledge Discovery from Databases (KDD)

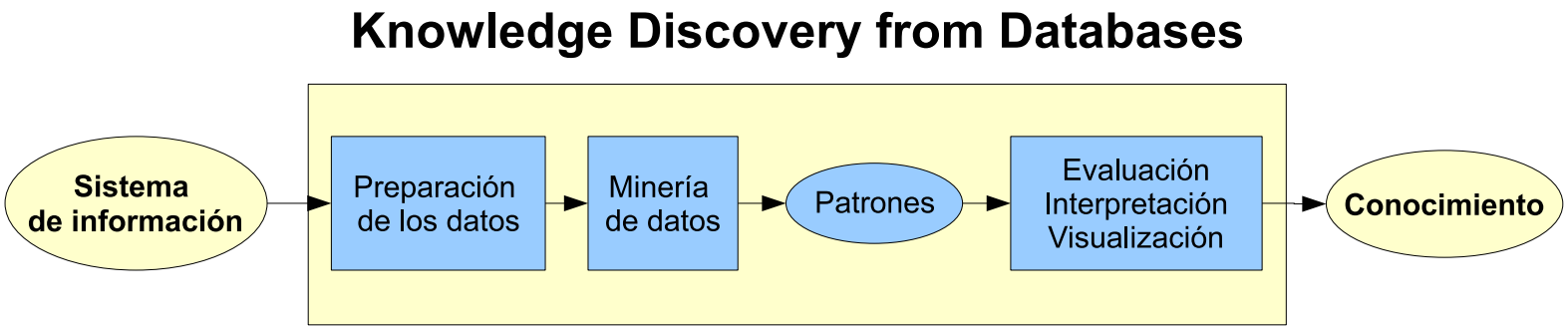


Figura: Proceso de extracción de conocimiento

Knowledge Discovery from Databases (KDD)

1. Integración y recopilación de datos

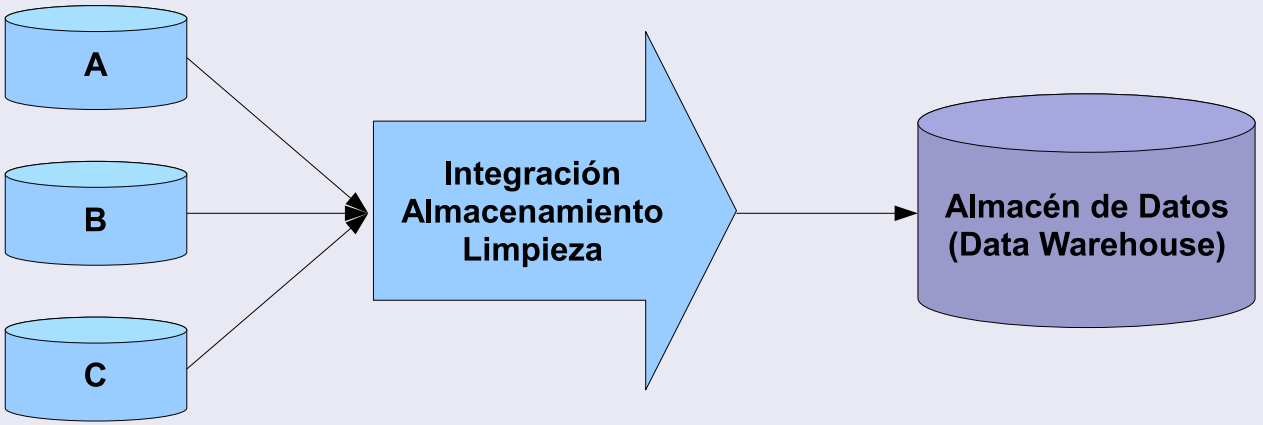


Figura: Integración en un almacén de datos

Knowledge Discovery from Databases (KDD)

2. Selección, limpieza y transformación

- Calidad del conocimiento descubierto depende (además del algoritmo de minería) de la calidad de los datos analizados
- Presencia de datos que no se ajustan al comportamiento general de los datos (outliers)
- Presencia de datos perdidos (missing values)
- Selección de variables relevantes (feature subset selection)
- Construcción automática de nuevas variables que faciliten el proceso de minería de datos
- Discretización de variables continuas

Knowledge Discovery from Databases (KDD)

3. Minería de datos: Modelos descriptivos

- Reglas de asociación
- Clustering: particional, probabilístico, jerárquico, conceptual
- Biclustering: Biclusters con valores constantes, biclusters con valores constantes en las filas o columnas, biclusters con valores coherentes.

Knowledge Discovery from Databases (KDD)

3. Minería de datos: Modelos predictivos

- Regresión: regresión lineal, regression tree, model tree, additive regression
- Clasificación supervisada: clasificadores Bayesianos, regresión logística, redes neuronales, árboles de clasificación, inducción de reglas, K-NN, combinación de clasificadores

Knowledge Discovery from Databases (KDD)

4. Evaluación e interpretación

- Técnicas de evaluación: validación simple (training + test), validación cruzada con k-rodajas, *bootstrapping*
- Clustering: variabilidad intra y entre
- Regresión: error cuadrático medio
- Clasificación supervisada: porcentaje de bien clasificados, matriz de confusión, análisis ROC
- Modelos precisos, comprensibles (inteligibles) e interesantes (útiles y novedosos)

Knowledge Discovery from Databases (KDD)

5. Difusión y uso

- Difusión: necesario distribuir, comunicar a los posibles usuarios, integrarlo en el *know-how* de la organización
- Medir la evolución del modelo a lo largo del tiempo (patrones tipo pueden cambiar)
- Modelo debe cada cierto tiempo de ser:
  - Reevaluado
  - Reentrenado
  - Reconstruido

Tema1. Introducción a la Minería de Datos

Abdelmalik Moujahid

Grupo de Inteligencia Computacional
Universidad del País Vasco UPV/EHU
Curso 2014-2015