QÜESTIIÓ, vol. 25, 3, p. 479-498, 2001

LA MINERÍA DE DATOS, ENTRE LA ESTADÍSTICA Y LA INTELIGENCIA ARTIFICIAL

TOMÀS ALUJA
Universitat Politècnica de Catalunya*

En la pasada década hemos asistido a la irrupción de un nuevo concepto en el mundo empresarial: el «data mining» (minería de datos). Algunas empresas han implementado unidades de minería de datos estrechamente vinculados a la dirección de la empresa y en los foros empresariales las sesiones dedicadas a la minería de datos han sido las protagonistas. La minería de datos se presenta como una disciplina nueva, ligada a la Inteligencia Artificial y diferenciada de la Estadística. Por otro lado, en el mundo estadístico más académico, la minería de datos ha sido considerada en su inicio como una moda más, aparecida después de los sistemas expertos, conocida desde hacía tiempo bajo el nombre de «data fishing».

¿Es esto realmente así? En este artículo abordaremos las raíces estadísticas de la minería de datos, los problemas que trata, haremos una panorámica sobre el alcance actual de la minería de datos, presentaremos un ejemplo de su aplicación en el mundo de la audiencia de televisión y, por último, daremos una visión de futuro.

Data mining, between statistics and artificial intelligence

Palabras clave: Data mining, análisis de datos, modelización, inteligencia artificial, KDD, redes neuronales, árboles de decisión

Clasificación AMS (MSC 2000): 62-07, 68T10, 62P30

* Departamento de Estadística e Investigación Operativa. Universitat Politècnica de Catalunya (UPC). E-mail: tomas.aluja@upc.es

– Recibido en abril de 2001.
– Aceptado en noviembre de 2001.

1. INTRODUCCIÓN

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información de las organizaciones. Esto es aún más evidente en las empresas de la nueva economía, el e-comercio, la telefonía, el marketing directo, etc. Los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan. Este es el objetivo de la Minería de Datos.

En primer lugar situemos la minería de datos a partir de algunas definiciones que se ha dado sobre la misma:

> Data Mining: «the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners» (Hand, 1998)
>
> «Iterative process of extracting hidden predictive patterns from large databases, using AI technologies as well as statistics techniques» (Mena, 1999)

La última, sin ser la definición más popular, enfatiza, sin embargo, cuáles son las raíces de la minería de datos: la Inteligencia Artificial (en particular *Machine learning)* y la Estadística.

Si buscamos a su vez definiciones de estas dos disciplinas:

> *Machine learning*: «a branch of AI that deals with the design and application of learning algorithms» (Mena, 1999)
>
> Estadística: «a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data . . . Statistics may be regarded as
>
> 1. the study of populations
> 2. the study of variation
> 3. the study of methods of the reduction of data»
>
> (Fisher, 1925)
>
> «methodology for extracting information from data and expressing the amount of uncertainity in decisions we make» (C. R. Rao, 1989).

Observemos que ya en 1925, Sir R. Fisher consideró la estadística bajo tres ópticas diferentes, como el estudio de poblaciones, lo cual está en el propio origen de la disciplina, como el estudio de la variabilidad que permite la modelización de los fenómenos

teniendo en cuenta la aleatoriedad presente en la naturaleza y como métodos de síntesis de la información contenida en los datos. La definición más moderna de C. R. Rao permite resaltar las coincidencias con la definición de minería de datos presentada más arriba.

Es claro que para cualquier persona vinculada con la estadística puede hablarse de dos tipos de estadística, una que podemos denominar Estadística Exploratoria ($\cong$ Data Analysis) y otra que podemos denominar Estadística Inferencial ($\cong$ modelling).

Si bien las fronteras entre ambos tipos de estadística no siempre es fácil de establecer, y a menudo la primera se presenta como la fase previa de la segunda (Cox & Snell, 1982) (Rao, 1989), existe una diferencia conceptual importante entre ambos tipos de Estadística. La Estadística Inferencial se refiere al paradigma central del quehacer estadístico, esto es, decidir entre varias hipótesis a partir de las consecuencias observadas. Consiste en incorporar la aleatoriedad dentro de la decisión (ya sea en su forma paramétrica mediante el razonamiento deductivo para delucidar la verosimilitud de cada hipótesis o por métodos computacionales).

«Data analysis» tiene un sentido muy general de estadística aplicada (con una connotación de aproximación pragmática e informatizada). Jean Paul Benzecri expresaba perfectamente el espíritu de este enfoque cuando afirmaba en sus cursos de 1965 que «le modèle doit suivre les données et non l'inverse» (si bien «Data Analysis» es más amplio que el equivalente francés de «analyse des données», el cual queda circunscrito a Análisis Multivariante Exploratorio). En este enfoque, no se trata de no tener en cuenta la naturaleza aleatoria de los datos (es obvio, por ejemplo, que cuando se seleccionan las componentes principales «significativas» para realizar una clasificación se está tratando de eliminar la parte aleatoria de los datos), sino que primero son los datos y es a partir de estos que se busca manifestar la información relevante para los problemas planteados.

Se puede constatar, sin embargo, que muchos de los problemas abordados en Análisis de Datos son comunes con la Inteligencia Artificial. Estas dos disciplinas, como a menudo sucede en el entorno académico, se han desarrollado la una a espaldas de la otra, dando lugar a nomenclaturas totalmente diferentes para problemas iguales. La Tabla 1, elaborada por el profesor L. Lebart, muestra las equivalencias para el problema de la predicción con redes neuronales.

Resumiendo mucho, podemos decir que la Inteligencia Artificial ha estado más preocupada en ofrecer soluciones algorítmicas con un coste computacional aceptable, mientras que la Estadística se ha preocupado más del poder de generalización de los resultados obtenidos, esto es, poder inferir los resultados a situaciones más generales que la estudiada.

Tabla 1. Equivalencias de nomenclatura entre la Estadística y la Inteligencia Artificial para el problema de predicción por redes neuronales.

Inteligencia Artificial	*Estadística*
red (network)	modelo
ejemplos (patterns)	observaciones, individuos
features, inputs, outputs	variables
inputs	variables explicativas
outputs, targets	variables de respuesta
errores	residuos
training, learning	estimación
función de error, coste	criterio de ajuste
pesos, coef. sinápticos	parámetros
aprendizaje supervisado	regresión, discriminación
aprenendizaje no supervisado	clasificación

Como mera ilustración de las aportaciones de ambas disciplinas al problema de la predicción, señalemos los hitos históricos de la regresión para la predicción de una variable continua (Galton, 1890), el análisis discriminante para la predicción de una variable nominal (Fisher, 1937), el AID para la construcción de árboles de decisión (Sonquist y Morgan, 1964), MARS (Friedman, 1991)... en Estadística, mientras que en el campo de la Inteligencia Artificial podemos citar el perceptrón, antecedente de las modernas redes neuronales (Rosemblat, 1958), los sistemas expertos, secuencias de reglas «if - then – else», para la toma de decisiones en los años setenta, los algoritmos genéticos (Holland, 1970), también los árboles de decisión (Quinlan, 1986)...

1.1. Nuevos problemas

La progresiva utilización de los avances tecnológicos por las empresas e instituciones hace aparecer nuevas colectas de datos y nuevos problemas. «Development in hardware have contributed to statistics by giving us many new and interestings sorts of data to analyse. Data have been able to be captured and stored quickly and cheaply by spectrometers, telescopes, process measuring devices, ... From these instruments have come new research problems. New applications have not arisen in science alone. Hardware changes have led to sophisticated point-of-sales terminals, bar-code readers and the ability to store and recall the huge volumes of data that are constantly being collected in warehouses, retail stores, government departments and financial institutions. Attempts to use such data to improve business performance have led to the field of data mining» (Cameron, 1997).

Un campo privilegiado de aplicación de las técnicas de minería de datos es el marketing, concretamente todo aquello que se agrupa bajo el nombre de CRM (*costumer relationship management*), donde el objetivo es conocer lo mejor posible los clientes para poder satisfacerlos mejor y asegurar así la rentabilidad de las empresas. Problemas tales como estimar el potencial económico de los clientes, modelizar la probabilidad de baja, medir la satisfacción por el servicio, descubrir nuevos segmentos de clientes potenciales, etc., son problemas que los responsables de la acción comercial de las empresas deben afrontar.

Pero no sólo las empresas o las instituciones son generadoras de nuevos problemas que afrontar, otros campos científicos también generan nuevos problemas donde la minería de datos se convierte en imprescindible, tales como las investigaciones originadas a raíz del proyecto Genoma, ¿qué secuencias de genes motivan la aparición de enfermedades?, ¿lo hacen de forma determinista o en probabilidad? También la información transmitida por satélite puede proporcionar avances a fenómenos hasta hoy difíciles de explicar, tales como la vulcanología, los terremotos o el clima, etc. La Tierra está dejando de ser el marco de referencia único para serlo cada vez más el sistema solar, como lo prueba la influencia que tienen las erupciones solares en las telecomunicaciones vía satélite. Otros campos de gran actualidad son encontrar métodos de predicción fiables, rápidos y baratos sobre la composición de los alimentos a partir del análisis del espectro infrarrojo de estos alimentos u otros análisis químicos.

Todo esto comporta la necesidad de tratar tablas de datos complejos y de tamaño inimaginable hasta ahora. Esta situación es nueva para el estadístico y bastante alejada de la clásica muestra aleatoria de observaciones independientes formada por algunas decenas de variables y unos cuantos millares de individuos. Tal como señala D. Hand (1998), ahora los datos son «secondary, messy, with many missings, noisy and not representative». Esto supone un reto para la estadística que obligará a repensar los esquemas clásicos de la inferencia estadística y de significación de los resultados observados. Si bien el nivel de significación fisheriano continua siendo válido para detectar la discrepancia entre los datos observados y la hipótesis formulada, es obvio que el orden de magnitud de los «*p*-value» es ahora muy inferior al acostumbrado. También es claro que en este contexto cobran renovada importancia los métodos de inducción computacionales, de simulación por Monte Carlo, bootstrap, etc.

2. ¿QUÉ PROBLEMAS ABORDA LA MINERÍA DE DATOS?

Cualquier problema para el que existan datos históricos almacenados es un problema susceptible de ser tratado mediante técnicas de Minería de Datos. Sin pretender ser exhaustivos la siguiente es una lista ilustrativa:

Búsqueda de lo inesperado por descripción de la realidad multivariante. Un principio clásico de la Estadística, el principio de la parsimonia, ya no es ahora válido (si bien siempre serán preferibles los modelos simples). Para describir un fenómeno cuantas más variables tengamos mejor, más ricas, más globales y más coherentes serán las descripciones y más fácil será detectar lo inesperado, esto es, aquello que no habíamos previsto y que resulta valioso para entender mejor el comportamiento de algún grupo de individuos, lo cual se ve favorecido por el hecho de trabajar con muestras grandes. Las muestras aleatorias son suficientes para describir la regularidad estadística global, pero no para detectar comportamientos particulares de subgrupos.

Búsqueda de asociaciones. Un cierto suceso, ¿está asociado a otro suceso?, ¿podemos inferir que determinados sucesos ocurren simultaneamente más de lo que seria esperable si fuesen independientes?, ¿es posible sugerir un producto, sabiendo que otro ha sido adquirido?

Definición de tipologías. Los consumidores son, a efectos prácticos, infinitos, pero los tipos de consumidores distintos son un número mucho más pequeño. Detectar estos tipos distintos, su perfil de compra y proyectarlos sobre toda la población, es una operación imprescindible a la hora de programar una política de marketing. Por otro lado, las tipologías no tienen que ser necesariamente de consumo, pueden ser de opiniones, valores, condiciones de vida, etc.

Detección de ciclos temporales. Todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se sitúa cada consumidor ayudará a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

Predicción. A menudo deberemos efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿lloverá mañana? Estas y muchas más son preguntas que deberemos responder, para ello construiremos un modelo a partir de los datos históricos. Si la variable de respuesta es continua (p. e. la rentabilidad de un cliente) diremos que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

3. LAS TÉCNICAS

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y sólo puede crecer en el futuro. También aquí, sin pretender ser exhaustivos, la siguiente es una lista de técnicas con una breve reseña.

Análisis Factoriales Descriptivos. Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

«*Market Basket Analysis*» o análisis de la cesta de la compra. Permite detectar qué productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes del de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

Técnicas de «clustering». Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

Series Temporales. A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

Redes bayesianas. Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

Modelos Lineales Generalizados. Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo la preferencia entre productos concurrentes en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos: (GAM, Projection Pursuit, PLS, MARS, . . .).

Previsión local. La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

Redes neuronales. Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.

Árboles de decisión. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

Algoritmos genéticos. También aquí se simula el modelo biológico de la evolución de las especies, sólo que a una velocidad infinitamente mayor. Es una técnica muy prometedora. En principio cualquier problema que se plantee, como la optimización de una combinación entre distintas componentes, estando estas componentes sujetas a restricciones, puede resolverse mediante algoritmos genéticos.

Un enriquecimiento de las posibilidades de análisis son los sistemas híbridos, esto es, la combinación de dos o más técnicas para mejorar la eficiencia en la resolución de un problema, como por ejemplo, utilizar un algoritmo genético para inicializar una red neuronal, o bien utilizar un árbol decisión como variable de entrada en una regresión logística.

En el futuro, el campo de actuación de la minería de datos no puede sino crecer. En particular debemos mencionar en estos momentos el análisis de datos recibidos por internet y «on line», dando lugar al *web mining*, donde las técnicas de *data mining* se utilizan para optimizar las interacciones a través de la *web*. ¿Cuáles son las secuencias de páginas más visitadas?, ¿qué páginas visitan los que compran?, ¿los que compran, vuelven a conectarse?, ¿cuales son las «killer pages»?, ¿una vez efectuada una adquisición, qué productos puedo sugerir?, son algunas de las preguntas que los responsables de comercio electrónico de las empresas se están formulando en estos momentos.

También los datos objeto de análisis pueden ser textos, dando lugar al *text mining.* Esto es particularmente útil en el análisis de las encuestas de satisfacción percibida por los usuarios. La utilización de las frases realmente escritas supone un enriquecimiento de los análisis realizados sólo con información numérica. También la utilización del *text mining* para la síntesis y la presentación de la información encontrada en la web es un campo actual de investigación. Más a largo plazo podrán utilizarse la voz o las imágenes.

Otra de las nuevas vías de investigación es el *fuzzy mining*, esto es, la utilización de las técnicas de minería de datos con objetos simbólicos, que representen más fidedignamente la incertidumbre que se tiene de los objetos que se estudian.

La tendencia actual más prometedora sería la de integrar los dos puntos de vista, provinientes de la estadística y de la Inteligencia Artificial, en las soluciones algorítmicas propuestas, de forma de aprovechar los puntos fuertes de ambas disciplinas. En consecuencia los algoritmos deberían contemplar las dos siguientes propiedades básicas:

Poder de generalización a poblaciones diferentes de la observada. Lo cual implica implementar técnicas eficientes de validación de resultados, ya sea a partir del conocimiento de la distribución muestral de los estadísticos del modelo o por métodos computacionales como la validación cruzada, etc.

Escalabilidad. Dado el volumen de datos a tratar, el coste de los algoritmos ha de ser todo lo lineal que sea posible respecto de los parámetros que definen el coste, en particular respecto del número de individuos.

4. COMPARACIÓN DE TÉCNICAS

Una pregunta que nos podemos formular es cuál es el mejor método para resolver un problema. La experiencia nos muestra que excepto ciertos problemas específicos y difíciles, la mayoría de problemas abordados en minería de datos dan resultados comparables cualquiera que sea la técnica utilizada. Hemos realizado una prueba con un fichero de 4000 individuos y 15 variables para explicar dos variables de respuesta sobre la adquisición de un cierto producto, el primero es un producto que podríamos calificar de relativamente «fácil» de predecir, mientras que el segundo es claramente más difícil. Hemos efectuado la predicción de ambas variables mediante 4 técnicas alternativas:

- Análisis Discriminante
- Redes neuronales
- Árboles de decisión
- Regresión Logística

Para medir la calidad de la predicción por cada método, hemos seleccionado al azar 4 muestras de 1000 individuos cada una como muestras de aprendizaje, utilizando los 3000 restantes como muestra de validación. Para cada método hemos realizado 3 ejecuciones cambiando ligeramente los parámetros del modelo. Por tanto, en total disponemos de 12 ejecuciones por método. Tomando el promedio de la probabilidad de acierto en las muestras de aprendizaje y en las de validación obtenemos los resultados que se muestran en la Tabla 2:

Tabla 2. Comparación de la probabilidad de acierto según 4 métodos de predicción.

Problema 1	*Apren.*	*Test*
Análisis Discriminante	71.13%	**69.71%**
Redes Neuronales	71.63%	**69.12%**
Árboles de Clasificación	72.94%	**70.31%**
Regresión logística	74.18%	**71.33%**
Problema 2	*Apren.*	*Test*
Análisis Discriminante	62.18%	**61.39%**
Redes Neuronales	62.29%	**60.19%**
Árboles de Clasificación	62.70%	**61.03%**
Regresión logística	65.28%	**59.36%**

Observando los resultados vemos que las probabilidades de acierto en la muestra de validación son bastante parecidas para los cuatro tipos de modelos utilizados.

5. EJEMPLO DE APLICACIÓN. DEFINICIÓN DE TARGETS COMPORTAMENTALES DE CONSUMO TELEVISIVO

Las innovaciones tecnológicas en el mundo audiovisual, producen el almacenamiento de una cantidad ingente de datos. El análisis de estos datos permite una mejora en la toma de decisiones por parte de las organizaciones implicadas.

En audiometría se dispone de información minuto a minuto de la audiencia realizada por un panel de familias. Estas observaciones pasan por un proceso de validación y enriquecimiento a partir de los datos sociodemográficos disponibles sobre los panelistas y por el minutado de programas y spots. Posteriormente, la muestra obtenida se afecta con un factor de elevación para obtener datos a nivel poblacional.

El problema planteado es el de definir targets compuestos explicativos del consumo de programas del género «Revistas del corazón», el cual ha experimentado un notable aumento en los últimos años en la programación televisiva.

Los datos analizados han sido todos los programas de este género emitidos durante el año 1997. La variable de respuesta ha sido los minutos semanales de visión de los programas de tipo *rosa*, en las cadenas estatales y para todos los individuos mayores de 3 años. Las variables explicativas son todas las sociodemográficas.

En el año 1997 se realizaron un total de 707 emisiones para este tipo de programas con una audiencia promedio de 32 minutos semanales por individuo.

El interés del problema planteado es claro, tanto para las propias Televisiones y Productoras, como técnica alternativa para definir targets afines a cualquier programa, como para las empresas de publicidad, anunciantes, agencias o centrales, al poder efectuar un «matching» entre los targets de consumo televisivo con el target consumidor del producto anunciado y así poder ser introducido en un sistema de «media-planing» para la compra de publicidad.

Clásicamente, este problema se soluciona mediante técnicas estadísticas simples, como es la distribución por variable o análisis de perfiles simples. La simplicidad de esta técnica es su principal ventaja. Así, por ejemplo, en el histograma de la Figura 1, mostrando el perfil de la audiencia de dibujos animados respecto de la edad, es clara la preferencia del segmento de niños (4 a 15 años) de este tipo de programas, pero también se observa una cierta afinidad con el segmento de personas jubiladas (más de 65 años). El histograma no revela si esta afinidad es propia del segmento o es debida a la presencia de niños en el hogar.

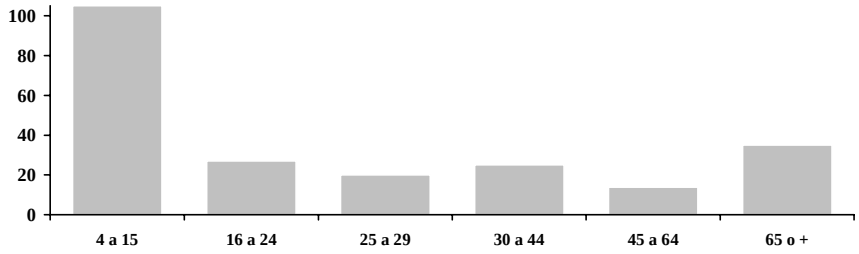


Figura 1. Perfil de audiencia de dibujos animados según la edad.

Una forma de obtención directa de targets compuestos del consumo televisivo de los programas *rosa*, es la utilización de árboles de decisión para explicar la audiencia de este tipo de programas. Escogemos esta metodología por su aplicabilidad inmediata de los resultados obtenidos. Estos resultados se obtienen de forma visual. La Figura 2 esquematiza el proceso de Minería de Datos en audiometría.

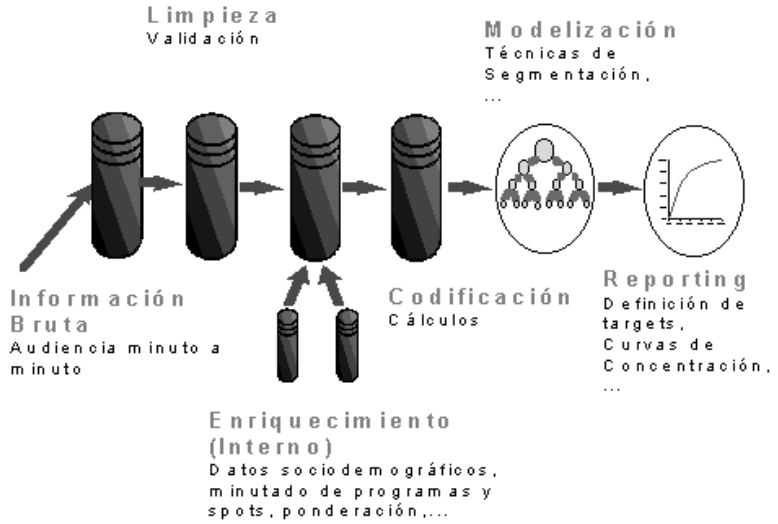


Figura 2. Sistema de KDD en audiometría.

El proceso de generación de un árbol es un proceso iterativo. Empieza situando toda la muestra disponible en el nodo raíz, a partir del cual, por sucesivas particiones, se obtienen las ramas del árbol hasta los nodos terminales u hojas, formadas por conjuntos de individuos que han visto un número similar de minutos los programas *rosa*.

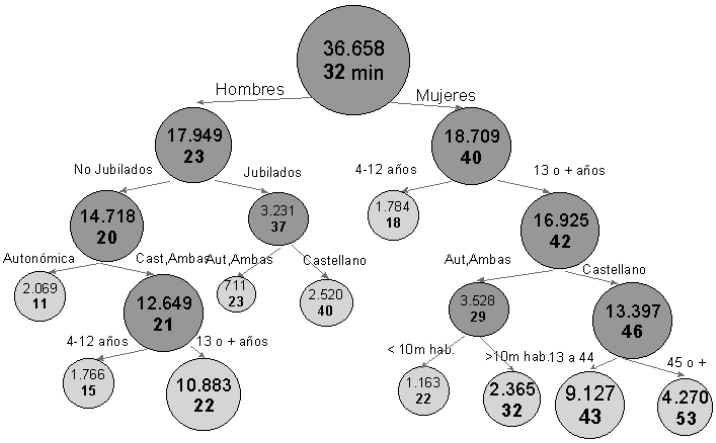


Figura 3. Ejemplo de árbol de decisión.

El algoritmo de construcción de un árbol de decisión implementa el siguiente bucle:

Hacer para cada nodo:

1. Verificar el criterio de parada del proceso en el nodo.
2. Definir la lista de todas las particiones posibles del nodo.
3. Seleccionar la partición óptima.
4. Generar la partición seleccionada.

La Figura 3 ilustra las primeras particiones del árbol generado.

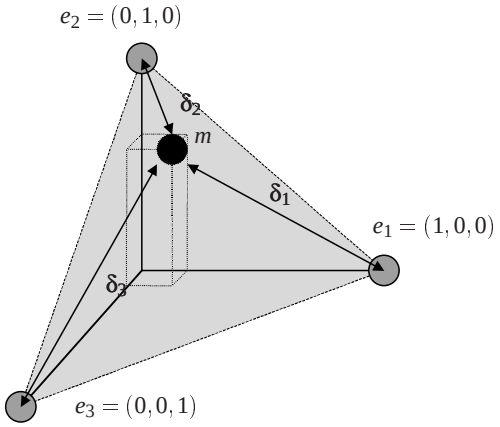
El árbol obtenido ilustra bien el proceso seguido, cada nodo da el número de individuos que contiene y el promedio de audiencia de estos individuos en los programas *rosa*. La obtención visual de los resultados permite a su vez su crítica, en efecto, la programación ofrecida por cada cadena condiciona la visión que puede hacerse de sus programas.

Existen varios algoritmos para la construcción de árboles de decisión: AID, CHAID, CART, C4.5. La diferencia básica es, aparte del hecho de que los árboles generados sean binarios o n-arios, la definición de la partición óptima de un nodo. Ciertamente seleccionar la partición óptima implica previamente definir un criterio de optimalidad.

Nosotros expresamos el criterio a optimizar en función de la pureza del nodo $i(t)$, definida por la siguiente fórmula:

$$i(t) = \frac{\sum_{i \in t} w_{it}\, \delta(i, m_t)}{\sum_{i \in t} w_{it}}$$

Función de los pesos de los individuos w_{it} del nodo y de las distancias de estos individuos al representante del nodo m_t.



Para el caso de variables de respuesta continuas y utilizando la métrica euclídea (norma L_2), la fórmula anterior se reduce a la conocida fórmula de la variancia de la variable de respuesta (y_i) en el nodo:

$$i(t) = \frac{\sum_{i \in t} (y_i - \bar{y}_t)^2}{n_t}$$

En este caso es obvio que efectuar una partición de un nodo implica descomponer la variancia total (V_T) del nodo original en dos componentes, una es la variancia *intra* (V_w) y la otra es la variancia *inter* (V_b):

$$V_T = V_b + V_w$$

Por tanto, maximizar la pureza de los nodos hijos implica minimizar V_w y por consiguiente maximizar V_b, esto es, encontrar dos nodos con la diferencia de medias lo más significativa posible (teorema de Huyghens).

El otro criterio a verificar en cada nodo es el criterio de parada, ya que en caso contrario podemos hacer crecer un árbol hasta que todos los nodos sean puros o contengan un solo individuo. Es evidente que entonces habríamos sobreparametrizado el árbol. Cuanto más avanzamos en la construcción del árbol, menos fiables son las particiones que se obtienen. Una manera de evitar esto es utilizar una técnica de validación como criterio de parada. Cuando se produzcan diferencias significativas entre la muestra de aprendizaje y la de validación, significa que las particiones no son estables.

La figura 5 muestra la calidad del árbol en función de su tamaño. En la muestra de aprendizaje esta medida es siempre monótona creciente, mientras que en la muestra de

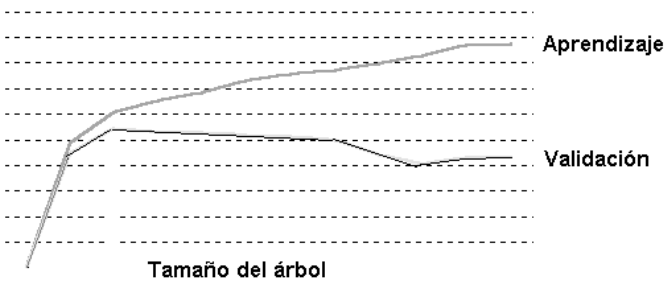


Figura 5. Calidad de un árbol en función de su tamaño.

validación, a partir de un cierto tamaño se estabiliza y puede llegar a decrecer, indicando que las particiones efectuadas más allá de este nivel son producto del azar.

6. CONCLUSIONES

La experiencia práctica muestra claramente la aptitud de las técnicas de minería de datos para resolver problemas empresariales. También es clara su aportación para resolver problemas científicos que impliquen el tratamiento de grandes cantidades de datos.

La minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de Análisis de Datos. Existe, además, una aportación propia de técnicas específicas de Inteligencia Artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste.

A diferencia de la IA, que es una ciencia joven, en Estadística se viene aprendiendo de los datos desde hace más de un siglo, la diferencia consiste que ahora existe la potencia de cálculo suficiente para tratar ficheros de datos de forma masiva y automática. Esta es una realidad que cada vez será más habitual. Sin abandonar ninguno de los campos previamente abordados, la Estadística ha evolucionado de ocuparse de la contabilidad de los estados a ser la metodología científica de las ciencias experimentales, hasta ser un «problem solver» para las organizaciones modernas. Es por esta razón el énfasis dado a que los resultados sean accionables.

Por otro lado y en relación a la amplia panoplia de técnicas disponibles, conviene tener claro de que no existe la técnica más inteligente, sino formas inteligentes de utilizar una técnica y que cada uno utiliza de forma inteligente aquello que conoce. También que para la mayoría de problemas no existen diferencias significativas en los resultados obtenidos.

Por todo lo dicho, es nuestra opinión de que la minería de datos no es una moda pasajera, sino que se entronca en una vieja tradición estadística y que cada vez más debe servir para hacer más eficiente el funcionamiento de las organizaciones modernas, ayudar a resolver problemas científicos y ampliar los horizontes de la Estadística.

7. BIBLIOGRAFÍA

Adriaans, P. & Zantige, D. (1996). *Data mining*. Addison-Wesley.

Aluja, T. & Morineau, A. (1999). *Aprender de los datos: el análisis de componentes principales, una aproximación desde el data mining*. EUB. Barcelona.

Aluja, T. & Nafria, E. (1996). «Automatic segmentation by decision trees». *Proceedings on Computational Statistics COMPSTAT 1996,* ed. A. Prat. Physica Verlag.

Aluja, T. & Nafria, E. (1998*a*). «Robust impurity measures in Decision Trees». *Data Science, Classification and related methods*, ed. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H-H. Bock and Y. Baba. Springer.

Aluja, T. & Nafria, E. (1998*b*). «Generalised impurity measures and data diagnostics in decision trees». *Visualising Categorical Data*, ed. Jörg Blasius and M. Greenacre. Academic Press.

Aluja, T. (2000). «Los nuevos retos de la estadística, el Data Mining». *Investigación y Marketing*, 68, 3, 34-38. AEDEMO.

Benzécri, J.-P. & *coll.* (1973). *La Taxinomie, Vol. I, L'Analyse des Correspondances, Vol. II*, Dunod, Paris.

Berry, M. J. A. & Linoff, G. (1997). *Data mining techniques for marketing, sales and customer support*. J. Wiley.

Beveridge, W. H. (1944). *Full employed in a free society*. George Allen and Unwin.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Booker, L. B., Goldberg, D. E. & Holland, J. H. (1989). *Classifier systems and genetic algorithms*. Springer-Verlag.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Waldsworth International Group, Belmont, California.

Cameron, M. (1997). «Current influences of Computing on Statistics». *International Statistical Review*, 65, 3, 277-280.

Celeux, G. (Ed.) (1990). *Analyse discriminante sur variables continues*, coll. didactique, INRIA.

Celeux, G. & Lechevallier, Y. (1982). «Méthodes de Segmentation non Paramétriques». *Revue de Statistique Appliquée*, XXX(4), 39-53.

Celeux, G. & Nakache, J. P. (1994). *Analyse discriminante sur variables qualitatives*. Polytechnica.

Ciampi, A. (1991). «Generalized Regression Trees». *Computational Statistics and Data Analysis*, 12, 57-78. North Holland.

Cox, D. R. & Snell, E. J. (1982). *Applied Statistics. Principles and Examples*. Chapman and Hall.

Elder, J. F. & Pregibon, D. (1996). «A statistical perspective on Knowledge Discovery in Databases». *Advances in Knowledge Discovery and Data Mining*, 83-116. AAAI Press.

Fayad, U. M., Piatetsky-Shapiro, G. & Smuth, P. (1996). «From Data Mining to Knowledge Discovery: an overview». *Advances in Knowledge Discovery and Data Mining*, 1-36. AAAI Press.

Fisher, R. A. (1925). *Statistical Methods, Experimental Design and Scientific Inference*. Oxford Science Publications.

Friedman, J. H. (1991). «Multiple Adaptative Regression Splines». *Annals of Statistics* 19, 1-141.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*. Academic Press.

Gueguen, A. & Nakache, J. P. (1988). «Méthode de discrimination basée sur la construction d'un arbre de décision binaire». *Revue de Statistique Appliqu ée*, XXXVI (1), 19-38.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. J. Wiley.

Hand, D. J. (1998). «Data Mining: Statistics and more?». *The American Statistician*, 52, 2, 112-118.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Kass, G. V. (1980). «An Exploratory Technique for Investigating Large Quantities of Categorical Data». *Applied Statistics*, 29, 2, 119-127.

Lebart, L., Morineau, A. & Piron, M. (1995). *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart, L., Salem, A. & Berry, E. (1998). *Exploring Textual Data*, Kluwer, Boston.

Lebart, L. (1998). «Correspondence Analysis, Discrimination and Neural Networks». *Data Science, Classification and related methods*, ed. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H-H. Bock and Y. Baba. Springer.

Lefébure, R. & Venturi, G. (1998). *Le data mining*. Eyrolles.

McCullagh, P. & Nelder, J. A. (1986). *Generalized Linear Models*. Chapman and Hall.

Mena, J. (1999). *Data Mining your website*. Digital Press.

Mola, F. & Siciliano, R. (1992). «A two-stage predictive splitting algorithm in binary segmentation». *Computational Statistics*, 1. Y. Dodge and J. Whittaker ed. Physica Verlag.

Murthy, S. K. (1998). «Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey». *Data Mining and Knowledge Discovery*, 2, 345-389.

Quinlan, J. (1988). *C4.5: Programs for machine learning*. Morgan Kaufman.

Rao, C. R. (1989). *Statistics and Truth*. CSIR, New Delhi.

Ripley, B. D. (1996). *Neural Networks and pattern recognition*. Wiley, New York.

Sarle, W. S. (1994). «Neural Networks and Statistical Models». *Proc. 9th. Annual SAS Users Group International Conference*. SAS Institute.

Sonquist, J. A. & Morgan, J. N. (1964). *The Detection of Interaction Effects*. Ann Arbor: Institute for Social Research. University of Michigan.

ENGLISH SUMMARY

DATA MINING, BETWEEN STATISTICS AND ARTIFICIAL INTELLIGENCE

TOMÀS ALUJA
Universitat Politècnica de Catalunya*

In the last decade a new concept had raised in the entrepreneurial side: data mining. Some companies have created data mining units directly linked to the CRM direction and in the professional forums data mining sessions have gained appeal. Data mining has appeared as a new discipline linked to Machine Learning, Artificial Intelligence and Data Bases, clearly differentiated from Statistics. On the other side, on the well-established statistics academia, data mining has been seen as the last fashion of a bad-known trend of data fishing and data dredging.

Is it really so? In this paper we will focus on the statistical roots of data mining, we will try to make and overview of the actual scope of data mining, we will present an application in the TV audience measure and we give some insights for the next future.

Keywords: Data mining, data analysis, modelling, artificial intelligence, KDD, neural networks, decision trees

AMS Classification (MSC 2000): 62-07, 68T10, 62P30

* Department of Statistics and Operations Research. Technological University of Catalonia (UPC). E-mail: tomas.aluja@upc.es

– Received April 2001.
– Accepted November 2001.

Statistics had risen in the beginning of XX century as a response to problems of society. Problems like defining a optimum fertiliser, the optimal conditions of production in an industry or assessing the effect of a drug, etc. Innovation always occurred due to stated problems. Anyway, we shall agree that statistics has been manipulating data for a most part of XX century without having a real computer device. Also a certain style of statistics installed in academia favoured a theoretical concept of the discipline. Nowadays, development in hardware have contributed to new and interesting sorts of data to analyse, which statistics should face. One of this problem is to come into knowledge the information hidden in the stored data by the information systems put in work for companies in the last two decades, coming up what is called the field of «data mining». It is no question to analyse small data files, but gigas or terabytes of data, with a precise goal, to take a managerial decision. This caused the appearance of data mining units in firms and its increasing interest in scientific meetings which devoted sessions to data mining. Anyway, data mining appeared linked to Artificial Intelligence, mainly machine learning, disciplines. Whereas it was considered by statisticians as a new version of the bad-known «data fishing» or «data dredging». It is really so?. I will establish that data mining stems from an old statistics tradition. In fact statistics lump together what can be called «data analysis» and «inferential statistics», the first being the first phase for the second (Cox, 1982, Rao, 1989). The difference between both approaches was wisely stated by Benzecry in his courses of 1964, «data is first, then follows the model», whereas for the «inferential» approach it is just the opposite. This is a sound difference, but it is clear that new problems arise, from the retail bar-code readers, transactions with a banking card, calls from a mobile telephone, or from the genome project, or satellite data, etc. This data very often is, as D. Hand (1998) pointed out, very large (huge), secondary, messy, with many missings, noisy and not representative. But it is absolute clear that statistics could play a central role for handling their associated uncertainty. These problems constitute a challenge for statisticians, pushing them to think again statistics, in particular the central issue of the tests of significance, and also to establish bridges of co-operation with our competitors of Artificial Intelligence, taking advantage of the strongness of both disciplines. Scientific disciplines, splitted in locked knowledge areas, have been developing isolated ones from others, leading to apparent different disciplines for the same problems. Here we show this applied to the case of Statistics and Artificial Intelligence, following the L. Lebart theory of two languages.
Finally we present a data mining application to the problem of finding good targets for TV program audience using decision trees. Decision trees, although they are not within the best classifier performers, like neural networks, generalised linear models, support vector machines, etc. have the appeal of being directly actionable, which in practice can overcome its shortcoming. Also for not very complex problems there is very little difference among different methods, making trees very useful for managerial applications. We follow the previously presented methodology (Aluja *et al*, 1998*b*) of building stable trees taking into account the individual contribution to impurity whithin the CART framework of tree building (Breiman *et al*, 1984).

Finally a point of humility from lord William Beveridge (1940) in this starving for knowledge:

> «Nobody believes a theory, except the one that has formulated it.
> Everybody believes a figure, except the one who has calculated it».