Nuevas perspectivas en el análisis de datos

1. INTRODUCCIÓN

Uno de los aspectos mas relevantes de los últimos diez años en el campo del análisis multivariante de datos ha sido la popularización del uso que estas técnicas han tenido en el mundo empresarial, sobre todo en las grandes corporaciones. La construcción de los llamados Data Warehouse o almacenes de datos con fines analíticos ha supuesto la necesidad de disponer de herramientas de análisis tanto para la explotación de los mismos como para su construcción.

Si la incorporación de la informática a la empresa supuso la mecanización de muchos procesos rutinarios y masivos, pensemos en las aplicaciones de nómina o de facturación, también supuso una inundación de datos que dio lugar a grandes bases de datos que almacenaban la historia de la empresa, clientes, empleados, costes de mercancías proveedores –dependiendo de los datos que manejara la distintas aplicaciones– . El potencial de conocimiento contenido en dichas bases de datos era enorme pero casi imposible de extraer debido a la forma en como se recogió y almacenó sin finalidad analítica y solo operativa. Superada la fase de mecanización en la empresa surge la necesidad de aprovechar el conocimiento escondido en esas masas de datos en gran parte debido al entorno cada vez más competitivo en el que se desenvuelve la actividad empresarial. Ya no era suficiente con facturar correctamente a un cliente se necesitaba saber lo que se conoce del mismo para evaluarlo y en su caso fidelizarlo. Tal y como se encontraban los datos en las distintas bases de datos de las diferentes aplicaciones hacía imposible responder a estas necesidades. Tuvieron que dar un giro en la forma de tratar y almacenar los datos teniendo en cuenta objetivos de análisis y se dio el primer paso en el que se contemplaba el dato organizado para su posterior análisis, en lo que se llamó paso del dato bruto al dato información a través de los repositorios de datos (Data Warehouse y Data Marts).

Desde el punto de vista del análisis multivariante de datos tanto el proceso de construcción de estos repositorios de datos como las necesidades de explotación de los mismos supusieron un nuevo tipo de problemas que dieron lugar al desarrollo de nuevas metodologías, mejora de algoritmos y de programas de análisis.

En este artículo nos vamos a centrar en algunos de estos problemas y en cómo se han tratado y a que resultados han dado lugar.

En lo que se refiere a los problemas que aparecen en la fase de construcción de los repositorios de información, los problemas surgen de la consolidación de los datos y en la limpieza de los mismos aparece estructuras de datos más complejas que dan lugar a una generalización del concepto de dato así como de su análisis. Una línea de investigación que lleva algunos años y ya ha dado frutos es el análisis de datos simbólicos al que dedicamos parte de este artículo.

Las necesidades del uso y explotación o análisis de la información de estos repositorios no es automática y el conocimiento que se esperaba encontrar escondido no es fácil de buscar; aparecen entonces conceptos nuevos como la minería de datos y el desarrollo de herramientas de navegación sobre bases de datos como medios para encontrar esos conocimientos.

2. HERRAMIENTAS DE NAVEGACIÓN SOBRE BASES DE DATOS

Este tipo de herramientas se conoce con el nombre de herramientas OLAP del inglés (On Line Analytical Processing). Se caracterizan por una exploración de los datos dirigida por el usuario. Las cuestiones típicas que el usuario se hace usando estas herramientas podrían ser del tipo: ¿Cuántas bicicletas vendimos el año pasado? ¿Cuál fue la facturación real en comparación a la planificada en los últimos 3 meses? ¿Cuáles son los 4 productos con mayores ventas clasificados por sus ingresos? El conocimiento se genera haciendo una serie de preguntas generalmente consecutivas a la base de datos. La diferencia de estas herramientas: con las habituales de consulta de una base de datos relacional es que están diseñadas para agilizar este tipo de consultas. Existen dos variantes principales de estas herramientas: las herramientas MOLAP, o multidimensionales, en las que se han predefinido una serie de dimensiones o variables sobre los que se ha construido una base de datos de datos preanalizados (base de datos multidimensional). Estas dimensiones suelen ser las que habitualmente consulta el usuario y al estar los datos preanalizados la consulta o navegación de los mismos es inmediata como contrapartida a la rapidez está la rigidez de las consultas pues resulta difícil si sale de las dimensiones predefinidas. En las herramientas ROLAP o relacionales la flexibilidad es mayor pero a costa de la rapidez pues durante la navegación se ha de acceder a todos los datos ya que no existen datos preelaborados. Existen además las HOLAP o híbridos que son una combinación de las anteriores.

Ejemplo

Como ejemplo de aplicación OLAP consideremos los datos resultantes de las ventas agregadas por región tipo de producto y canal de ventas. Una consulta típica a una aplicación OLAP pasaría por acceder a una base de datos enorme para obtener la tabla de todas las ventas realizadas por tipo de producto y región. Un analista después de revisar dicha tabla puede interesarle profundizar y querer saber las ventas realizadas por cada canal de venta dentro de cada tipo de producto y región. El siguiente paso podría ser la comparación de distintos años de cada canal de ventas para cada tipo de producto en cada región. Todo este proceso ha de poder ser

efectuado en tiempo real y rápidamente de forma que el proceso de análisis no se vea alterado. Es decir, podemos caracterizar a las consultas a una base de datos OLAP por:

- Acceso a grandes masas de datos; por ejemplo, varios años de datos sobre ventas.
- Análisis de relaciones entre distintas variables, ventas, productos canales de venta, regiones.
- Tratamiento de datos agregados como son unidades vendidas, presupuesto en euros, etc.
- Comparar datos agregados a lo largo de periodos de tiempo con estructura jerárquica años meses, quincenas, semanas.
- Presentación de la información en diferentes perspectivas, ventas por región, ventas por canales de distribución en cada región.
- Cálculos complejos como la ganancia esperada como función de los ingresos por cada canal de distribución para una región particular.
- Rapidez en la respuesta a las cuestiones planteadas por los usuarios de forma que éstos puedan proseguir su análisis sin que el sistema informático sea un obstáculo.

Las estructuras multidimensionales que usan las herramientas OLAP para representar la información pueden visualizarse como cubos de datos dentro de otros cubos. Cada lado de un cubo es una dimensión. Es por esto por lo que a este tipo de representación se la conoce como hipercubos. Cada dimensión representa una variable diferente como tipo de producto, región, canal de distribución. Cada celdilla dentro de la estructura multidimensional contiene el dato agregado en cada dimensión, así por ejemplo una única celdilla puede contener el total de ventas para un tipo de producto en una determinada región para un canal de distribución en un mes determinado.

Algunas de las funciones de análisis típicas de este tipo de herramientas son la consolidación que supone la posibilidad de agregación de datos, como, por ejemplo, las ventas por delegación pueden agregarse a ventas por distrito, y éstas a ventas por región. La operación inversa (Drill-down) o presentación de la información en detalle es también una funcionalidad habitual. También es propia la función de corte o la capacidad de observar la base de datos desde distintos puntos de vista. Un corte puede ser las ventas por producto en cada región. Otro corte las ventas por canal en cada tipo de producto; es frecuente que se hagan estos cortes a lo largo de un eje o dimensión temporal con el fin de identificar tendencias.

Las herramientas OLAP usan técnicas de reducción para maximizar la utilización del espacio de almacenamiento; así, los datos que ocupan un porcentaje alto de celdillas o datos densos se almacenan separadamente de los datos que aparecen en un porcentaje muy pequeño de celdillas. Esto supone minimizar el espacio físico de almacenamiento permitiendo analizar grandes cantidades de datos.

Podemos concluir que las bases de datos multidimensionales a diferencia de las bases de datos relacionales, que están diseñadas para operaciones transaccionales, están orientadas al análisis y la decisión.

3. LA MINERÍA DE DATOS

Comercialmente bajo este nombre la industria del software ofrece a los responsables de analizar la información de los data warehouse herramientas de análisis que les van a permitir encontrar patrones y regularidades en grandes conjuntos de datos. El ordenador es el responsable de encontrar estos patrones identificando las reglas y características escondidas en los datos. Con esta aproximación se pretendía evitar el cuello de botella que para las demandas de análisis de los nuevos repositorios de datos representaba el análisis de datos tradicional, caracterizado por un proceso lento y artesanal apoyado en paquetes de programas estadísticos que requerían un conocimiento experto tanto en estadística como en el uso de paquetes de programas. La minería de datos permite aunar el conocimiento experto sobre los datos a analizar con técnicas avanzadas de análisis por ordenador. Podría pretenderse eliminar el papel del estadístico aunque realmente no es así. La estadística sigue teniendo un papel que jugar pero ahora actúa sobre los resultados de la minería de datos. El uso de estas técnicas en los últimos años ha demostrado que son los expertos en análisis de datos los que más rendimiento sacan a una herramienta de minería de datos, debido a la dificultad inherente a la interpretación de algunas de las técnicas. Podríamos decir que si bien es verdad que las herramientas de minería de datos han permitido automatizar el proceso de un análisis de datos mediante nuevas aplicaciones mucho más amigables y con nuevos algoritmos de análisis que los paquetes de programas estadísticos tradicionales no han podido evitar el usuario experto.

Muchas de las técnicas que hay detrás de la minería de datos son técnicas estadísticas tradicionales, aunque las más características son las orientadas a tratar grandes volúmenes de datos, es típico de la minería de datos usar el máximo de datos posibles para llegar a conclusiones fiables. Otras disciplinas relacionadas con la minería de datos son el aprendizaje automático y la inteligencia artificial.

Funcionalidades más comunes en las herramientas de minería de datos

Clasificación

La clasificación, entendida como la obtención de un modelo a partir de la base de datos en el que en el caso de aprendizaje supervisado el usuario ha de definir dos o más clases, las variables o atributos de la base de datos se agrupan en aque-

llos que se han de predecir o dependientes y que una combinación de ellos identifican las clases, y los predictores. Las reglas de clasificación o clasificadores se generan a partir de los predictores y sobre un conjunto de registros de entrenamiento en los que se conoce el valor para todos los atributos. La bondad de los clasificadores se evalúa en un conjunto distinto de registros de la base de datos o conjunto test. El resultado de la evaluación dará reglas exactas o aquellas que clasifican correctamente todas las valores de la clase. O reglas aproximadas en función del porcentaje de malas clasificaciones.

Asociación

Dada una colección de items y un conjunto de registros que contienen uno o varios items, una función de asociación aplicada a ese conjunto de registros devuelve patrones o relaciones entre los items de la colección. Generalmente estos patrones se expresan mediante reglas del tipo: "el 87% de los registros que contienen los items A, B, C también contienen D y E". El porcentaje (87) se conoce como el factor de confianza de la regla.

Una aplicación muy común de esta funcionalidad es la llamada *Análisis de la cesta de la compra,* Una gran superficie de ventas aplica un asociador a los registros obtenidos a partir de transacciones que contienen identificadores de compradores y de productos adquiridos. La salida del asociador dará una serie de afinidades entre productos del tipo: "el 25% de las veces que se adquirió una bicicleta de montaña también se adquirió guantes y gafas deportivas".

Patrones secuenciales o temporales

Analizan un conjunto de registros a lo largo del tiempo para identificar tendencias. Si la identidad de un comprador se conoce se puede analizar el conjunto de registros relativo a las compras efectuadas en un periodo de tiempo. Es una situación típica de las ventas por catálogo o mediante tarjeta. Es útil para descubrir orden de compras: "¿Qué tipo de artículos suele preceder a la compra de una bicicleta de montaña?".

Son muy potentes para identificar conjuntos de clientes asociados a determinados patrones de compra. Por ejemplo, aplicado a un conjunto de registros de reclamaciones de una compañía de seguros puede servir para detectar fraude.

Segmentación

Entendemos por segmentación el proceso de particionar en clases un conjunto de forma que todos los miembros de una misma clase son semejantes de acuerdo a algún criterio. A las clases se les denomina clusters o conglomerados. Como en la clasificación, la segmentación sobre grandes bases de datos se suele realizar sobre una muestra de entrenamiento de la que se obtienen los clusters o clases que luego se proyectan a toda la base de datos mediante la generación de clasificadores; es así como se procede, por ejemplo, para la segmentación de los clientes de una compañía eléctrica que supone analizar bases de datos de millones de clientes.

Las técnicas mas usadas en minería de datos

Análisis de conglomerados

Es una técnica clásica del análisis multivariante usada para realizar las funciones de segmentación de bases de datos; las herramientas de minería de datos incorporan algoritmos muy eficientes para segmentar un gran numero de registros, aún así es necesario aplicar el procedimiento sobre muestras indicado más arriba cuando se tratan de millones de registros.

Inducción

Las técnicas que permiten inferir información que son generalizaciones de la información observada en los registros de una base de datos. Alguno de los métodos de inducción usados en minería de datos son:

– *Árboles de decisión:*
Los árboles de decisión representan clasificadores en forma de árbol; los nodos se etiquetan con el nombre de atributos o variables, las ramas que salen de un nodo con valores de dichas variables y las hojas se etiquetan con las clases. Los objetos se clasifican siguiendo un camino en el árbol a lo largo de las ramas que corresponden a los valores de los atributos del objeto.
El siguiente árbol clasifica los objetos, en este caso días del año, en las clases contenidas en el nodo inicial A1,A2,...,C3. Los nodos se etiquetan con las variables Tipo de Día, Promoción, Estación, etc. Así el día que sea laborable, que no haya tenido promoción y que sea de diciembre se etiqueta como A3.

– *Inducción de reglas:*
Generalmente provienen del aprendizaje automático y permiten generar reglas de clasificación del tipo: "si..., entonces..." Las reglas de producción han sido ampliamente usadas en representar conocimiento en los sistemas expertos y tienen la ventaja de ser fácilmente interpretables y pueden ser entendidas por los usuarios aisladamente sin necesidad de referirse a otras reglas.

– *Redes Neuronales:*
Las redes neuronales son una herramienta de cálculo que implica el desarrollo de estructuras matemáticas con capacidad de aprendizaje. Están basadas en aproximaciones del modelo de aprendizaje del cerebro.
Las redes neuronales utilizan un conjunto de procesadores o nodos análogos a las neuronas en el cerebro. Estos nodos están interconectados formando una red que puede identificar patrones, los datos una vez que actúa sobre éstos.
Las redes neuronales han sido aplicadas con éxito a muchos

problemas reales en diferentes campos. Por su especial habilidad para identificar patrones y tendencias en los datos se comporta bien en problemas de predicción como son la predicción, de ventas, gestión de riesgos, control de procesos industriales, validación de datos, etc.

El inconveniente que tienen es que actúan como una caja negra, en el sentido que no da explicaciones sobre los resultados a los que llega; esto hace que el usuario pueda desconfiar de los resultados y no usarlos. Otra de las desventajas que presentan es que los tiempos de aprendizaje son muy largos y empeoran cuando aumenta el volumen de datos.

4. ANÁLISIS DE DATOS SIMBÓLICOS

Uno de los campos de investigación, que se está desarrollando actualmente en el campo del análisis de datos, es el denominado análisis de datos simbólicos. En éste se generaliza la noción de dato clásico a estructuras más complejas, que se denominan datos simbólicos. En el análisis de datos clásicos los datos que se analizan vienen de observaciones únicas de variables sobre individuos únicos. Los datos simbólicos pueden encontrarse en muchas situaciones como las que resultan de tratar agrupaciones o agregados de individuos; la necesidad de describir las propiedades de estas clases necesitan de un nuevo tipo de variables distintas de las clásicas, y son las variables y datos simbólicos. Es frecuente encontrarse la necesidad de analizar datos simbólicos en grandes bases de datos como las de las Oficinas de Estadística o, como ya se ha mencionado, en los Data Warehouses de grandes compañías. No solamente los datos simbólicos permiten la descripción de clases de individuos sino que pueden provenir del conocimiento de un experto sin necesidad de datos individuales así como representar incertidumbre. También los datos simbólicos permiten representar metadatos como son las dependencias lógicas entre las variables y las dependencias jerárquicas entre las variables.

Diday ya en 1991 introdujo los objetos simbólicos como un medio para tratar conocimientos más ricos que los datos habituales, y estableció una relación con el modelo clásico de análisis de datos. Un objeto simbólico representa una intención o concepto y se define, en términos generales, como una conjunción de valores o conjunto de valores de variables que pueden ser ponderados. Constituye la descripción en intención de una clase de individuos, y éstos constituyen la extensión.

La siguiente tabla ilustra lo que podría ser una tabla de datos simbólicos.

En esta tabla encontramos distintos tipos de variables simbólicas:

- Variables Multievaluadas; por ejemplo, la variable "cliente" que puede tomar varios valores distintos en la misma celda.
- Variables Multievaluadas con pesos como, por ejemplo, "la ciudad". En el distribuidor P1 la variable "ciudad" toma valor "Madrid" con peso "1/2" y la variable "Burgos" con peso "1/2", que puede interpretarse como que el distribuidor P1 puede venir tanto de Madrid como de Burgos con la misma frecuencia.
- Reglas; además de la información contenida en la tabla podemos incorporar reglas a la entrada de datos como, por ejemplo, si la "ciudad" es Madrid y el "cliente" es A1, la "compañía" es C1.
- Taxonomías; por ejemplo, en la variable "ciudad" podríamos considerar "provincias" o "Comunidades Autónomas" que las sustituyan.

Distribuidor	*Compañía*	*Cliente*	*Horario semanal*	*Ciudad*	*Capacidad*
P1	C1	A1,A3	34	Madrid (1/2), Burgos (1/2)	(12,25)
P2	C2	A2	45	Sevilla	(20,20)
P3	C3	A4,A5	40	Madrid (1/3), Logroño (2/3)	(34,40)

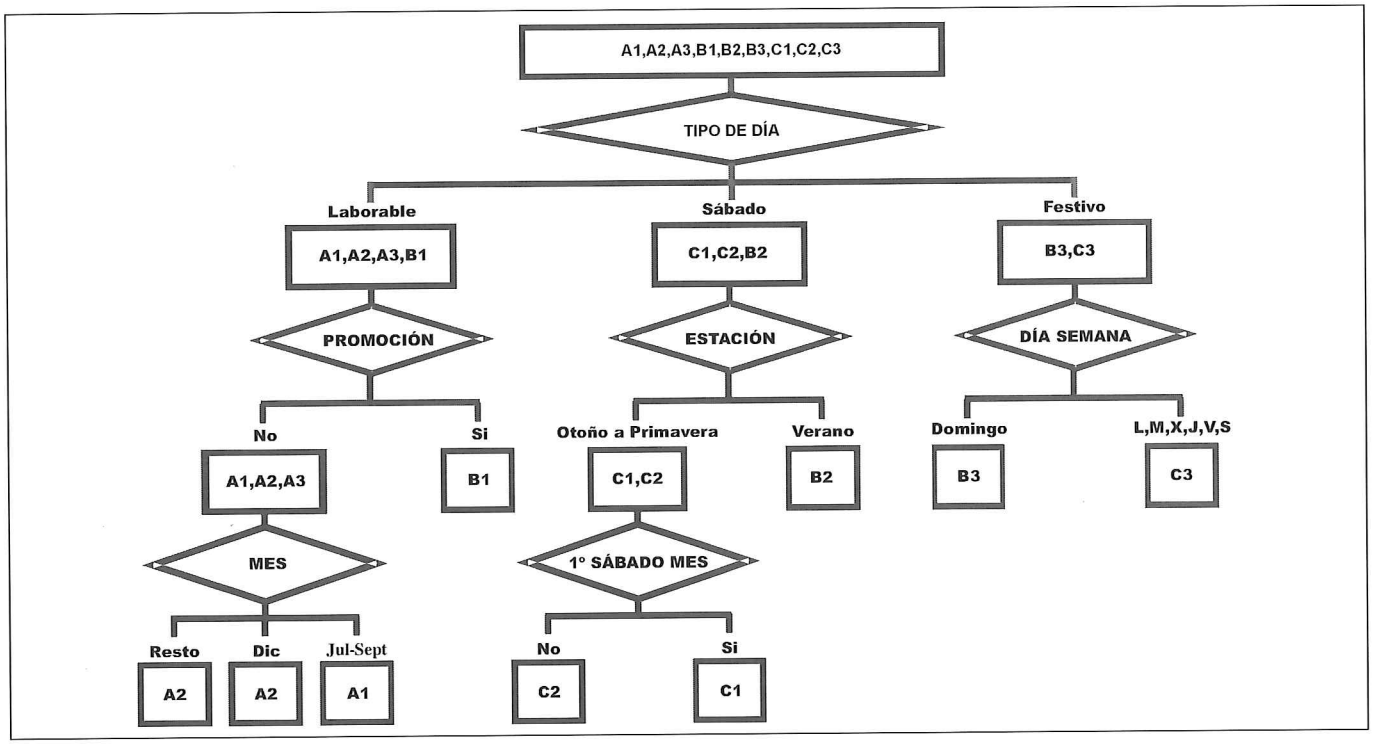
Se han desarrollado distintos métodos de análisis de datos simbólicos que generalizan los métodos sobre datos clásicos y que ya han dado lugar a un paquete de programas así como libros y artículos. En particular, y como consecuencia del proyecto SODAS (European Sprit Project 20821) en que han participado 17 grupos de investigación e institutos nacionales de estadística, se ha desarrollado el paquete SODAS.

El objetivo del mismo es, por un lado, generar bases de datos de objetos simbólicos a partir de bases de datos relacionales, además incorpora distintos métodos de análisis como son: Análisis de componentes principales y Análisis factorial discriminante de datos simbólicos. La salida de estos métodos conserva la variación interna de los datos de entrada, los individuos no se representan por un punto, como en el método clásico, sino por rectángulos que permiten la definición de objetos simbólicos con los ejes factoriales como variables explicativas.

Otros métodos de análisis son: Generalización de estadísticos descriptivos para datos simbólicos. Clustering mediante particiones jerarquía o pirámides de forma que cada clase se asocie a un objeto simbólico. Generalización de árboles de decisión, representaciones gráficas, etc.

5. OTROS RETOS PARA EL ANÁLISIS DE DATOS

Hay una serie de problemas en el mundo real que necesitan de solu-



ciones propias del análisis multivariante, pero que incorporan una característica que plantea una solución que en los tratamientos y técnicas clásicas no se consideran: la respuesta en tiempo real.

Cada vez se suelen condensar sistemas más complejos que, o bien se autocontrolan, o bien sirven recomendaciones a operadores externos a la hora de ejecutar una decisión. Esto se hace debido al avance tecnológico en la captación de medidas, analizando y procesando matrices de señales y parámetros, que cambian de manera continua y cuya propia complejidad es la del sistema integrado bajo consideración.

Todo ello conlleva la necesidad de una mejora en técnicas y algoritmos, tanto estadística (discriminación, clasificación, segmentación...), como propias de la I.O. y la I.A. (algoritmia, aprendizaje automático...).

Vamos a señalar los campos en que esta necesidad se plantea:

— *Data Fusión:*
De amplia utilización en sectores como el militar. Consiste en la recepción de múltiples señales de distintos sensores –por ejemplo, en un avión de combate sometido a distintas amenazas, las cuales son recogidas como señales–, que junto con técnicas de discriminación, para evaluar las verdaderamente influyentes en la decisión posterior, y con técnicas de clasificación, para ordenarlas en función de la importancia de la contingencia, poder dar una respuesta para que el sistema ejecute una acción –que el avión, por ejemplo, cambie de rumbo–.

– *Análisis de logs:*
Con el uso creciente de Internet y de la oferta de múltiples Portales Web Empresariales, etc., a que se puede acudir, los procesos de marketing de las webs de empresas que comienzan a ser visitados –y que se inicia con una señal respuesta al click del ratón, que nos introduce en la web– despierta la necesidad del reconocimiento del visitante para fidelizarle, ofreciéndole de manera fácil los siguientes espacios dentro del proceso de navegación en la web. Son necesarios técnicas de asignación de patrones, discriminación, clasificación...., con respuestas en tiempo real, son necesarios.

– *Mensajes en centro de mando y control:*
Sirva como referencia un dispatching de una operadora eléctrica cuya red está sometida a los avatares del consumo; continuamente se realizan dos tipos de operaciones en paralelo: una consistente en la generación de casos que recogiendo un instante y sucesivos de la red (topología, cargas...) nos indique cómo va a evolucionar, para en su caso "aprender" a tomar acciones preventivas o, si sucede una incidencia, acciones correctoras, y otra, al mismo tiempo, el conjunto de señales recogidas, nos debe llamar en tiempo real, –y de nuevo con algoritmos supereficientes–, a clasificar y averiguar alguno de los casos tratados, para operar de manera útil.
Esta sucesión de situaciones, nos indica que las técnicas estadísticas multivariantes deben añadir una preocupación más, y es mejorar la eficacia de los algoritmos y, también, que la búsqueda de soluciones, como en el caso de análisis de logs, se puede hacer con información parcial y correcciones sobre la misma, en tiempo real; por ejemplo, con cada señal del ratón en el caso citado.

Javier Martín Rodrigo
y José Miguel García-Santesmases
Universidad Complutense de Madrid