Minería de datos[1]

por

José A. Lozano, Universidad del País Vasco–Euskal Herriko Unibertsitatea

1. Introducción

La minería de datos (ver [1] y [2]) es un área que se encuentra en la intersección de diversas disciplinas, la inteligencia artificial, la estadística, las ciencias de la computación y las matemáticas. El objetivo de la minería de datos es obtener conocimiento mediante la construcción de modelos computacionales a partir de datos. Este objetivo involucra diferentes tareas: desde el descubrimiento de patrones en conjuntos de datos, pasando por la estructuración de un dominio hasta la creación de modelos predictivo-descriptivos. La minería de datos ha sido objeto de gran investigación en las dos últimas décadas, este interés por la disciplina tiene dos bases, por un lado el incremento en la capacidad de proceso y almacenamiento de los computadores personales actuales y por otro la gran cantidad de aplicaciones prácticas exitosas de las técnicas que se manejan dentro de la disciplina.

A continuación, y con el objetivo de dar una idea general del tipo de problemas y el tipo de situaciones en las que se puede utilizar la minería de datos, se verán algunos ejemplos de aplicaciones reales de técnicas de minería de datos:

Reconocimiento automático de dígitos. Una de las aplicaciones clásicas de la minería de datos es el reconocimiento automático de dígitos: es decir, a partir de una imagen de un dígito escrito a mano, conocer cuál es el dígito que representa. Esta aplicación es particularmente

[1]Trabajo realizado en colaboración con Ekhine Irurozki, Departamento de Ciencias de la Computación e Inteligencia Artificial, UPV/EHU, ekhine.irurozki@ehu.es

de interés para los servicios de correos, ya que el reconocimiento automático de dígitos escritos a mano, automatiza la tarea de clasificar los envíos postales. En relación con esta aplicación es importante reseñar que el objetivo es *clasificar* cada dígito dentro de una las siguientes 10 categorías $\{0, 1, \ldots, 9\}$.

Segmentación de clientes. En la actualidad es muy común el uso de tarjetas de fidelización de clientes, probablemente la más famosa de ellas sea la travel club. Estas tarjetas tienen por objetivo recoger la información de las compras que realizan los poseedores de la tarjeta. Una vez recogida la información de las transacciones comerciales se puede llevar a cabo un trabajo de *agrupar* clientes con patrones de compra similares. Una vez agrupados los clientes es posible estudiar cada grupo para, observando sus hábitos de consumo, realizar publicidad específica para dicho grupo, ofrecer productos que con alta probabilidad van a ser comprados, etc. En este caso la tarea es agrupar datos similares.

Sistemas de recomendación. Dentro de Internet es posible encontrar muchos servidores que se dedican a vender libros como Amazon u otros para escuchar música como LastFM. En ambos casos los servidores utilizan "sistemas de recomendación" para ofrecer al usuario ciertos productos. En el primer caso, una vez que se realiza una compra, el sistema nos ofrece otros productos, que *curiosamente* nos interesan. En el segundo caso, el sistema propone unas canciones al usuario y el usuario las evalúa (mala, buena, muy buena, etc.). Pasado un tiempo el servidor únicamente propone canciones que son del agrado del usuario. El tipo de técnicas que se utilizan en ambos casos son similares: se buscan sujetos que tengan comportamientos similares y a partir de ellos se realizan propuestas.

Biología. La biología ha sufrido un gran cambio en los últimos años. Ha pasado de ser una disciplina con escasez de datos a un campo donde existe una gran abundancia de ellos. Este cambio ha sido el producto de la aparición de nuevas tecnologías que permiten recoger gran cantidad de información. Una de estas tecnologías permite medir la expresión de miles de genes al mismo tiempo. A partir de estos datos es posible tratar de buscar que genes están asociados con una enfermedad concreta. Para ello se parte de un conjunto de datos donde existen personas afectadas por la enfermedad y personas que no lo están. Utilizando dicho conjunto de datos y un conjunto de técnicas matemático-computacionales es posible establecer qué genes están relacionados con la enfermedad.

Medicina. En el campo de la medicina existe gran cantidad de problemas donde establecer un diagnóstico es muy complicado. Uno de estos problemas, por ejemplo, es saber cuando una persona padece apendicitis o no. Una de las aplicaciones de la minería de datos es la construcción de un sistema donde, a partir de datos del paciente (temperatura corporal, presión sanguínea, etc...) se pueda predecir con gran probabilidad si el paciente padece apendicitis o no.

Virus informáticos. Existen aplicaciones de la minería de datos que consisten en la localización de virus informáticos. El objetivo no solo es detectar los virus que se conocen, sino sobre todo detectar virus que aún no se conocen.

A la vista de los ejemplos anteriores es posible establecer ciertas características y tareas de las aplicaciones de la minería de datos. Común a todas las aplicaciones es la existencia de un conjunto de datos del que se quiere extraer información y la existencia de incertidumbre inherente al dominio de aplicación (las cosas nos son ni blancas ni negras, sino que existen muchos niveles de grises). Las tareas son principalmente de tres tipos, construir modelos predictivos, crear grupos dentro de datos (*clustering*) y seleccionar variables o características que describan los datos.

2. Problemas de la minería de datos

Tal y como se ha comentado en la sección anterior existen *básicamente* tres escenarios dentro de la minería de datos. Un primer escenario donde el objetivo es crear un modelo predictivo a partir de un conjunto de datos. Es decir, se trata de crear un modelo cuyo objetivo es clasificar un nuevo dato en una clase dentro de un conjunto de clases predeterminadas. En este caso el aspecto del conjunto de datos a partir del que se construye el modelo viene dado en la Figura 1.

	X_1	...	X_n	C
Caso 1	$x_1^{(1)}$	...	$x_n^{(1)}$	$c^{(1)}$
Caso 2	$x_1^{(2)}$	...	$x_n^{(2)}$	$c^{(2)}$
...		...		...
Caso N	$x_1^{(N)}$	...	$x_n^{(N)}$	$c^{(N)}$

Figura 1: Ejemplo de fichero de datos en problemas de clasificación.

En un segundo escenario el objetivo es descubrir grupos dentro de un conjunto de datos. El producto final en este caso debe ser el número de grupos

existentes en el conjunto de datos así como el grupo al que pertenece cada dato. El fichero de datos del que se dispone en un problema de agrupamiento debe ser parecido al que se muestra en la Figura 2.

	X_1	...	X_n	C
Caso 1	$x_1^{(1)}$	...	$x_n^{(1)}$	?
Caso 2	$x_1^{(2)}$	...	$x_n^{(2)}$	?
...		...		...
Caso N	$x_1^{(N)}$	...	$x_n^{(N)}$	?

Figura 2: Ejemplo de fichero de datos en problemas de agrupamiento o clustering.

En medio de estas dos situaciones está lo que se denomina la clasificación semisupervisada, donde se dispone de un fichero de casos, algunos de los cuales tienen asociados una etiqueta y otros no (véase la Figura 3).

	X_1	...	X_n	C
Caso 1	$x_1^{(1)}$	...	$x_n^{(1)}$	c_1
Caso 2	$x_1^{(2)}$	...	$x_n^{(2)}$	c_2
Caso 3	$x_1^{(3)}$	...	$x_n^{(3)}$	?
Caso 4	$x_1^{(4)}$	...	$x_n^{(4)}$	?
...		...		...
Caso N	$x_1^{(N)}$	...	$x_n^{(N)}$	?

Figura 3: Ejemplo de fichero de datos en problemas de clasificación semisupervisada.

Finalmente, el último problema consiste en la selección de aquellas variables o características que son relevantes para un problema concreto. En los ficheros que hemos visto en las figuras anteriores el objetivo sería seleccionar algunas X_i como las relevantes. El ejemplo de la selección de genes asociados a enfermedades se encuentra dentro de este marco.

3. El proceso de la minería de datos

Aunque el objetivo de la minería de datos es la construcción de los modelos descriptivo-predictivos que hemos comentado anteriormente existen muchas tareas que hay que realizar antes de la construcción de dichos modelos y también posteriormente a la construcción del modelo. El conjunto de actividades que hay que realizar un proyecto de minería de datos se agrupan

principalmente en tres familias de actividades: el preproceso, la construcción del modelo y la evaluación del mismo.

El preproceso trata de preparar los datos para la posterior construcción del modelo. Esta limpieza implica ciertas actividades como pueden ser:

- **Limpieza de datos.** Existen muchas bases de datos donde los registros están repetidos. En estos casos es necesario chequear la existencia de los mismos y borrar los registros repetidos. También pueden darse situaciones donde hay variables que tienen asignados valores que no tienen sentido (por ejemplo una temperatura corporal igual a 0).

- **Tratamiento de datos perdidos.** Existen gran cantidad de bases de datos que no están completas sino que poseen una cantidad importantes de datos perdidos. Esto puede deberse a diversas razones, desde errores en el proceso de grabado de los datos hasta la no existencia de valores para algunas medidas. En todo caso, la mayoría de los algoritmos de construcción de modelos no pueden tratar con datos perdidos por lo que es necesario realizar algún tratamiento con dichos datos. Existen diferentes alternativas, desde el borrado de aquellos registros con datos perdidos hasta la imputación, es decir, asociar un valor basándose en ciertos criterios estadísticos.

- **Selección de variables.** A pesar de que intuitivamente se puede pensar que cuantas más variables describan a cada caso mejor será el modelo que se puede construir, esto no es del todo cierto. De hecho, muchos estudios han demostrados que existen modelos donde la inclusión de variables irrelevantes o redundantes pueden disminuir la capacidad predictora del modelo. Por lo tanto, un paso importante antes de construir el modelo es decidir cuáles son las variables necesarias y de interés.

- **Otras tareas.** Existen situaciones donde el modelo que se quiere construir sólo admite variables discretas (toman una cantidad finita de valores), sin embargo las variables de las que se dispone son continuas (pueden tomar cualquier valor en un intervalo). En estos casos es necesario aplicar técnicas que discretizan los datos, es decir, convierten variables continuas en variables discretas. Otra actividad que es necesaria en algunas ocasiones es llevar a cabo una selección de casos. Existen bases de datos con una cantidad inmensa de los mismos. La utilización de todos estos datos en la construcción del modelo conllevaría en muchas ocasiones tiempos computacionales demasiado grandes. Por lo tanto, es necesario aplicar un conjunto de técnicas que puedan seleccionar el conjunto de datos más adecuado para construir un modelo.

Una vez realizado el preproceso el siguiente paso consiste en la construcción del modelo. Existen diferentes tipos de modelos, algunos de ellos se mostrarán en la siguiente sección. Habitualmente la construcción del modelo se lleva a cabo de forma automática mediante la utilización de un algoritmo que, partiendo de los datos, obtiene el modelo descriptivo-predictivo.

Un último paso del proceso de la minería de datos es el de la evaluación del modelo. Supongamos que hemos creado un modelo predictivo a partir de un conjunto de datos, la pregunta ahora es, ¿cuál es la probabilidad de que el modelo acierte el valor de la clase de un nuevo caso que nunca ha visto? Dicha probabilidad es lo que se conoce como la precisión del modelo. El cálculo de la precisión del modelo es un problema complicado, ya que queremos calcular cómo se va a comportar el modelo en situaciones que no se han visto hasta el momento. Desde un punto de vista matemático es un problema de estimación. Existen diferentes técnicas para realizar dicha estimación. La más común es la denominada *hold-out* que consiste en tomar el conjunto de datos y dividirlo en dos partes. Con la primera parte se aprende el modelo y luego se calcula la precisión de dicho modelo con los datos de la segunda parte. Una técnica más sofisticada es el *k-fold crossvalidation*. En este caso el conjunto de datos es dividido en k partes cada una de ellas con el mismo número de casos. A continuación se aprenden k modelos cada uno de ellos con $k-1$ partes y cada modelo se testea (calculándose un valor de precisión) en la parte con la que no se ha construido el modelo, obteniéndose al final k valores de precisión. La precisión final asociada al modelo con todos los datos es el valor medio de las k precisiones.

4. Modelos de minería de datos

Existe una gran cantidad de modelos de minería de datos, desde modelos basados en probabilidad, pasando por modelos basados en lógica fuzzy, hasta modelos basados en reglas. A continuación, y centrándonos en modelos predictivos, veremos con cierto detalle algunos de estos modelos. Los tres modelos elegidos difieren en su filosofía y también en el resultado final que obtiene el modelo.

4.1. Los k vecinos más cercanos

Uno de los modelos de clasificación más comunes es el de los k vecinos más cercanos. De hecho, no se llega a construir un modelo sino que existe una regla de clasificación: dado un nuevo caso se busca, dentro del conjunto de entrenamiento, los k casos más cercanos al actual (la medida para evaluar la cercanía entre dos casos depende del tipo de datos que se esté utilizando).

Se comprueba cuáles son las clases asignadas a dichos k casos y aquella clase que aparece más veces es con la que se clasifica al nuevo caso. La Figura 4 muestra un ejemplo de aplicación de la técnica de los k vecinos más cercanos.

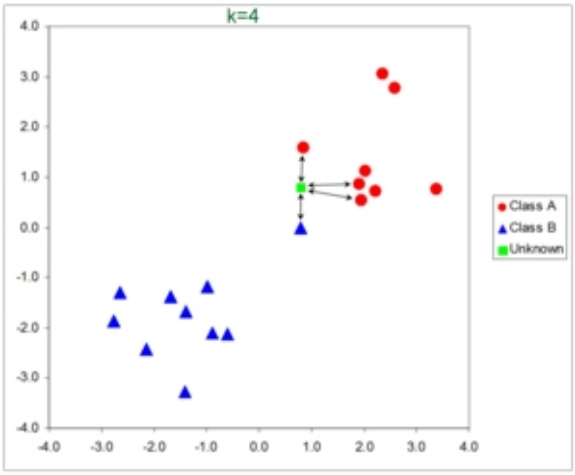

Figura 4: Ejemplo de clasificación de mediante la técnica de los k vecinos más cercanos.

El algoritmo de los k vecinos más cercanos posee ciertas características de gran interés: por una lado es una técnica muy sencilla que puede ser fácilmente comprendida y también implementada en un ordenador. Por otro es posible demostrar que la regla de clasificación es óptima cuando el tamaño del conjunto de entrenamiento tiende a infinito. Los puntos más débiles de los k vecinos más cercanos son el coste computacional. Cuando el número de casos en el conjunto de entrenamiento es demasiado alto, la búsqueda de los k vecinos más cercanos se convierte en una tarea que consume tal cantidad de recursos computacionales que se hace inviable su uso.

4.2. Árboles de clasificación

Los árboles de clasificación construyen un modelo de árbol a partir de los datos. En cada nodo del árbol se toma una decisión en relación a una de las variables involucradas en los datos: por ejemplo, a partir de un nodo es posible crear dos ramas: una se corresponde con $temperatura \geq 38$ y la otra con $temperatura < 38$. Cada hoja del árbol se corresponde con una de las posibles clases en las que cada dato puede ser clasificado. Puede existir más de una hoja que se corresponde con la misma clase. La Figura 5 representa un árbol de clasificación.

Dado un nuevo caso, éste recorre el árbol comenzando por la raíz hasta que llega hasta una de las hojas. La clase asociada a la hoja es la que le corresponde al nuevo caso.

Este modelo es muy diferente al anterior ya que en este caso se tiene un modelo explícito que puede proporcionar información acerca de cómo está estructurado el dominio. Es decir, en este caso, no sólo se consigue un modelo predictivo sino que el modelo describe y proporciona información acerca del dominio en el que se está trabajando.

La construcción del modelo se realiza mediante un algoritmo de aprendizaje. Dicho algoritmo, a partir del conjunto de datos, va construyendo el árbol determinando cuál es el atributo a seleccionar en cada nodo y qué ramas crear en relación a los valores del atributo.

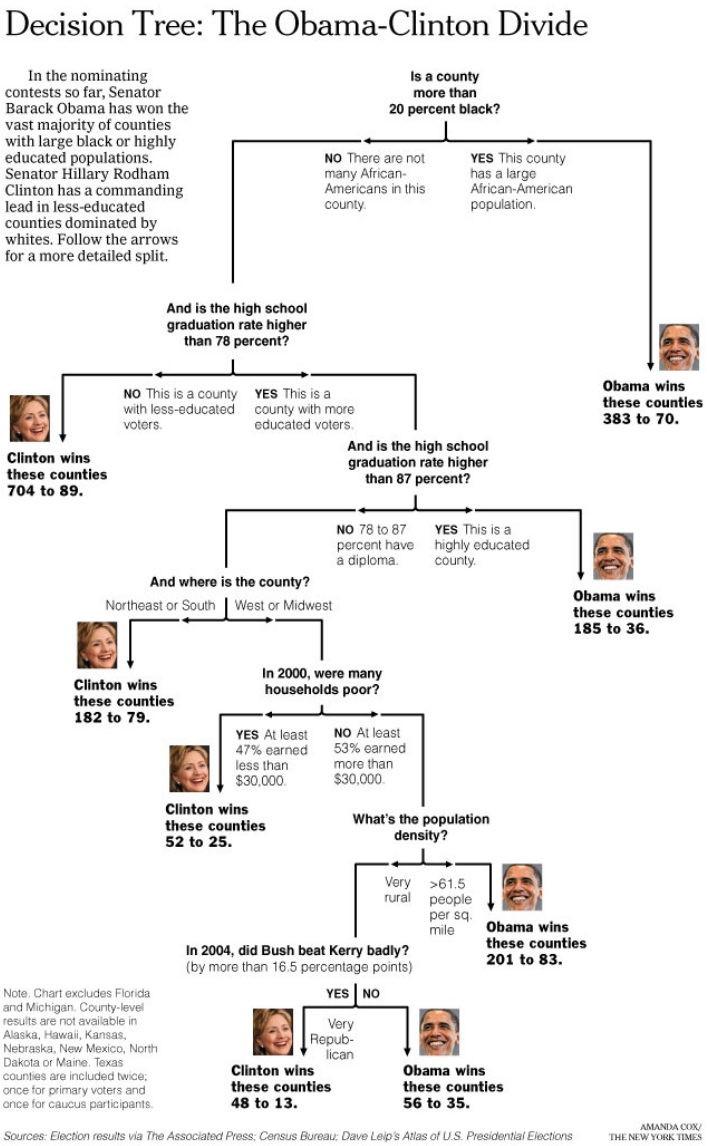

Figura 5: Ejemplo de clasificación de mediante un árbol de clasificación.

4.3. El modelo naive-Bayes

El modelo naive-Bayes es un modelo probabilístico. En este caso se supone que todos los casos de los que se dispone (y los que puedan venir) provienen de una distribución de probabilidad que involucra a todas las variables del modelo, incluida la variable a clasificar. El modelo naive-Bayes construye una distribución de probabilidad a partir del conjunto de datos y posteriormente utiliza dicha distribución de probabilidad para clasificar nuevos casos. Suponiendo que el modelo ha aprendido la distribución $p(x_1, \ldots, x_n, c)$ y que únicamente podemos clasificar un caso como 0 o 1, entonces dado un nuevo caso $(x_1, \ldots, x_n)$ se calculan las probabilidades siguientes:

$$p(C = 0|x_1, \ldots, x_n) \quad \text{y} \quad p(C = 1|x_1, \ldots, x_n)$$

si la probabilidad asociada a $C = 0$ es mayor que la asociada a $C = 1$ entonces la clase que se asocia al nuevo caso es 0, en caso contrario es 1[2].

El cálculo de una distribución de probabilidad conjunta completa es inviable, tanto desde el punto de vista computacional (se necesita una cantidad exponencial de parámetros que calcular y almacenar) como estadística (para realizar una estimación razonable de todos los parámetros sería necesaria una cantidad exponencial de datos). Por lo tanto, el modelo naive-Bayes realiza suposiciones sobre la dependencia probabilística de los datos de cara a simplificar la distribución de probabilidad de los mismos. Particularmente, este modelo asume que cualquier par de variables predictoras son condicionalmente independientes dada la variable clase. De esta forma, el número de parámetros a estimar es proporcional al número de variables. Sin embargo, estas suposiciones son demasiado severas en algunos casos y los resultados que se obtienen con este modelo pueden no ser muy buenos. En la Figura 6 es posible ver un ejemplo de un clasificador naive-Bayes junto con la factorización de la distribución de probabilidad asociada.

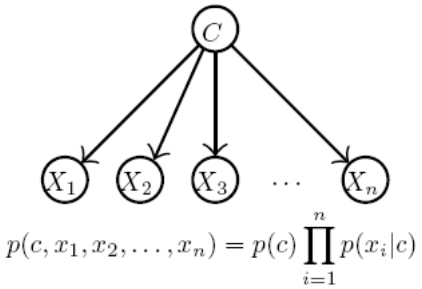

Figura 6: Ejemplo de un clasificador naive-Bayes.

[2]Obviamente no es necesario calcular ambas probabilidades, se ha realizado de esta manera por motivos pedagógicos

En este caso el aprendizaje del modelo consiste en el aprendizaje de los parámetros de la distribución de probabilidad. La forma más común de aprender dichos parámetros es utilizando el método de máxima verosimilitud.

4.4. Redes neuronales

La redes neuronales son otro modelo de clasificación ampliamente utilizado en aplicaciones prácticas de minería de datos. En este caso el modelo trata de imitar el comportamiento de las neuronas humanas. El modelo está compuesto de un conjunto de unidades de cómputo muy sencillas, las neuronas, que están unidas entre sí formando una estructura de red. Habitualmente las neuronas se dividen en tres conjuntos, neuronas de entrada (existe una por cada variable predictora en el conjunto), variables ocultas (no es posible acceder a ellas de forma directa) y variables de salida (donde se obtiene la clasificación). Dado un nuevo caso, la neuronas de entrada toman el valor del caso y estos valores se propagan a través de la red hasta que las variables de salida toman un valor. Dicho valor es el valor de clasificación del caso. En la Figura 7 se puede ver un ejemplo de una red neuronal y en la Figura 8 se muestra una neurona.

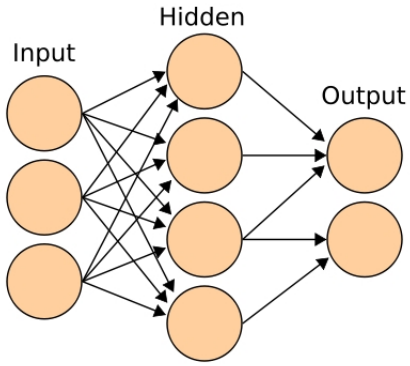

Figura 7: Ejemplo de una red neuronal.

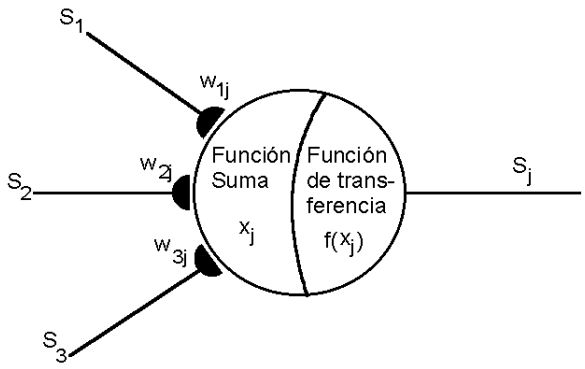

Figura 8: Ejemplo de una neurona.

5. Conclusiones

El conjunto de técnicas que constituye la minería de datos se aplica a un número tan diverso de ámbitos que actualmente se encuentra casi a diario en nuestra vidas, desde que abrimos el ordenador hasta que compramos en un supermercado. Este conjunto de técnicas se está imponiendo en aplicaciones prácticas dada su flexibilidad y su amplio abanico de posibilidades en el tratamiento de datos. Aunque es una disciplina relativamente nueva, sus posibilidades de desarrollo son grandísimas y van estrechamente unidas al desarrollo de las computadoras y del modelado matemática de datos. En los próximos años veremos grandes avances que nos sorprenderán y modificarán nuestros hábitos de vida.

Bibliografía

[1] J. Hernandez, M.J. Ramírez y C. Ferri: *Introducción a la minería de datos*, Pearson Educación, 2008.
[2] B. Sierra: *Aprendizaje Automático: Conceptos Básicos y Avanzados*, Pearson Educación, 2006.

José A. Lozano
Universidad del País Vasco
Euskal Herriko Unibertsitatea
Facultad de Informática
Departamento de Ciencias de la Computación
e Inteligencia Artificial
Manuel de Lardizabal 1, 20018 San Sebastián
e-mail: ja.lozano@ehu.es
http://www.sc.ehu/ccwbayes/members/jalozano