DEL ANALISIS ESTADISTICO A LA MINERIA DE DATOS

1 Introducción

En esta tema se presentan dos tipos de problemas –clasificación no supervisada y clasificación supervisada– que son de propósito general para buen número de paradigmas tanto provenientes de la Estadística como de una rama de la Inteligencia Artificial conocida como Aprendizaje Automático.

A lo largo del tema se mostrarán varios problemas provenientes de distintos dominios, en los que estos paradigmas pueden ser aplicados, tratando de recalcar las diferencias y similitudes entre los paradigmas provenientes de la Estadística y aquellos cuyo origen está en el Aprendizaje Automático.

2 Clasificación no supervisada frente a clasificación supervisada

En este apartado vamos a mostrar las características básicas de dos técnicas −*clasificación no supervisada* y *clasificación supervisada*– que se engloban dentro de la denominación de *reconocimiento de patrones*.

Ambas técnicas son de una amplia aplicación en muy diversos campos, tal y como se desprende de los ejemplos que se van a citar más adelante.

2.1 Clasificación no supervisada

En la clasificación no supervisada –véase tabla 9.1– cada uno de los N objetos que constituyen el fichero de casos, viene caracterizado por n variables, que denotamos por $X_1, \ldots, X_n$. Se denota por x_i^j el valor que el j–ésimo objeto toma en la i–ésima variable, con $i = 1, \ldots, n$ y $j = 1, \ldots, N$.

Tabla 9.1: Fichero de casos de partida para la clasificación no supervisada.

	X_1	...	X_i	...	X_n
1	x_1^1	...	x_i^1	...	x_n^1
...	...	...	...	...	...
j	x_1^j	...	x_i^j	...	x_n^j
...	...	...	...	...	...
N	x_1^N	...	x_i^N	...	x_n^N

El objetivo que se persigue con la clasificación no supervisada (*clustering*) de dicho fichero es el de agrupar objetos que presenten características similares, de tal forma que en cada grupo (*cluster*) se tenga una alta similaridad entre los elementos que forman parte del mismo y al mismo tiempo exista una alta disimilaridad entre los distintos grupos.

Los distintos métodos desarrollados en la clasificación no supervisada se pueden organizar según diferentes criterios. Si nos fijamos en el tipo de variables podemos hablar de *taxonomía numérica* (cuando todas las variables son cuantitativas) versus *clustering conceptual* (variables cualitativas). Fijándonos en el resultado de la clasificación no supervisada, distinguiremos entre *clustering particional* (cada objeto pertenece a un sólo grupo) y el *clustering jerárquico* (se obtiene una jerarquía de particiones). Podemos también considerar una *clasificación particional probabilística* (en la cual se establece la probabilidad de pertenencia de cada objeto a cada uno de los grupos). En esta última, la aplicación del algoritmo EM a los modelos mixtos resulta de especial interés.

En estos apuntes se presentarán en el tema 12 algunos algoritmos de clustering particional y clustering jerárquico para el caso en que los objetos están caracterizados por variables numéricas.

Típicos ejemplos de aplicación de la clasificación no supervisada los tenemos en los siguientes dominios:

- Biología (clasificación de animales, árbol filogenético)
- Nutrición ('rueda' de alimentos)
- Imágenes (segmentación)
- Marketing (tipología de clientes)
- Sociología (tipología de encuestados)

2.2 Clasificación supervisada

En la clasificación no supervisada –véase tabla 9.2– cada uno de los N objetos que constituyen el fichero de casos, viene caracterizado por $n+1$ variables, de las cuales las n primeras, $X_1, \ldots, X_n$, son *variables predictoras*, y la $n+1$–ésima, C, es la *clase*. Dicha variable C es la variable a predecir. Se denota por x_i^j el valor que el j–ésimo objeto toma en la i–ésima variable, con $i = 1, \ldots, n$ y $j = 1, \ldots, N$. Asimismo, c^j denota la clase a la que pertenece el j–ésimo objeto.

Tabla 9.2: Fichero de casos de partida para la clasificación supervisada.

	X_1	...	X_i	...	X_n	C
1	x_1^1	...	x_i^1	...	x_n^1	c^1
...	...	...	...	...	...	...
j	x_1^j	...	x_i^j	...	x_n^j	c^j
...	...	...	...	...	...	...
N	x_1^N	...	x_i^N	...	x_n^N	c^N

El objetivo que persigue la clasificación supervisada es el inducir automáticamente un *modelo clasificatorio*. Dicho modelo clasificatorio se utilizará con posterioridad para predecir la clase de pertenencia de objetos que tan están caracterizados por sus variables predictoras.

Aspectos importantes en dicha modelización son, entre otros, los siguientes:

1. La medición honesta de la bondad del modelo clasificatorio (tema 10).
2. La selección de un subconjunto de las n variables con el que construir modelos parsimoniosos. Visto desde otro punto de vista, se trata de encontrar aquellas variables irrelevantes o redundantes con objeto de llevar a cabo la inducción del modelo con el resto de las variables (tema 11).
3. Determinación del *paradigma clasificatorio* más adecuado para unos datos concretos (temas 13 al 21).

Algunos ejemplos de aplicación de los métodos de clasificación supervisada ordenados según dominios son los siguientes:

- Medicina (diagnóstico de mamografías a partir de imágenes digilatilizadas; supervivencia a distintas enfermedades; predicción de la concentración de glucosa en pacientes diabáticos, ...)
- Finanzas (fidelización de clientes; predicción de bancarrota en empresas; concesión de créditos, ...)
- Computación (predicción de fallos en discos duros; búsqueda de información en la World Wibe Web; *collaborative filtering*, ...)

3 Métodos Estadísticos versus Aprendizaje Automático

El término *minería de datos* ha surgido recientemente dentro de la Inteligencia Artificial referido a un proceso que partiendo de un fichero de casos –recogiendo cantidades más o menos voluminosas de datos– trata de extraer conocimiento de dicho fichero de casos.

Dicho proceso de minería de datos consta de 5 pasos básicos:

1. Recopilización y almacenamiento de la información.
2. Filtrado de la información.
3. Extracción de datos para llevar a cabo el estudio.
4. Modelado.
5. Toma de decisiones.

En este apartado nos vamos a centrar en el cuarto paso y más concretamente en modelos cuya finalidad es la de clasificación supervisada. Comentaremos las diferencias entre las aproximaciones a la misma provenientes de la Estadística en relación con las que tienen su origen en el Aprendizaje Automático.

Las características generales en problemas de clasificación supervisada de los Métodos Estadísticos en contraposición con los métodos provenientes del Aprendizaje Automático se han reflejado en la tabla 9.3. A continuación se desarrolla cada uno de los items recogidos en dicha tabla.

Tabla 9.3: Características generales en problemas de clasificación supervisada de los Métodos Estadísticos y de los provenientes del Aprendizaje Automático.

	Métodos Estadísticos	Aprendizaje Automático
Asunciones	si	no
Score	verosimilitud	porcentaje bien clasificados
Búsqueda	heuri. determinista + test hipót.	heuri. no deterministas
Transparencia	poca	mucha
Validación	no	si
Selección variables	filter	wrapper

- Asunciones
  Los Métodos Estadísticos presuponen que los datos constituyen una muestra aleatoria simple de tamaño N extraida de una determinada población, la cual sigue una distribución probabilística previamente especificada. En los paradigmas de Aprendizaje Automático no es necesraio el efectuar este tipo de suposiciones de partida.

- Score
  En la modelización estadística la bondad de cada posible modelo se mide por medio de la verosimilitud que el mismo asigna al fichero de N casos. Por el contrario suele ser habitual que en los paradigmas de clasificación supervisada englobados dentro del denominado Aprendizaje Automático, se guie la búsqueda por medio del porcentaje de bien clasificados que proporciona el modelo evaluado. Es decir, en los paradigmas de clasificación supervisada provenientes del Aprendizaje Automático, la evaluación de cada uno de los modelos candidatos se efectúa con un criterio que coincide con la finalidad del clasificador.

- Búsqueda
  Esta es una característica que diferencia claramente a los modelos clasificatorios según su procedencia. En los desarrollados dentro de la Estadística suele ser habitual que el modelado se efectúe por medio de un algoritmo *hill–climbing* en combinación con un test de hipótesis basado en la razón de verosimilitud.

Una solución típica, denominada modelización hacia adelante (*forward*), consiste en contemplar la modelización como un proceso iterativo, que partiendo del modelo nulo, va en cada paso construyendo el modelo correspondiente con las variables ya seleccionadas en el paso anterior, a las cuales se les une la variable –seleccionada de entre las no escogidas hasta ese momento– cuya inclusión hace que el nuevo modelo tenga una verosimilitud que sobrepasa de manera significativa a la verosimilitud del modelo en el paso anterior. Esta significatividad se chequea por medio del test de la razón de verosimilitud. El proceso iterativo para cuando la inclusión de ninguna de las variables no seleccionadas hasta ese momento, hace que el nuevo modelo mejore significativamente en verosimilitud.

La modelización hacia atrás (*backward*) actúa de forma análoga a la anterior. En este caso se parte de un modelo construido con todas las variables, y en cada paso nos preguntamos cual de las variables –a partir de las que se construye el modelo actual– puede ser quitada sin que la verosimilitud del mismo se reduzca significativamente. El proceso para cuando la exclusión de cualquiera de las variables con las que se ha construido el modelo nos lleva a que la verosimilitud del nuevo modelo decrezca significativamente.

En la modelización hacia adelante (atrás) una vez que la variable ha sido incluida en (excluida del) modelo, esta no puede volver a ser excluida del (incluida en el) mismo. Una modelización mas flexible, y que tambien se utiliza en Estadística, es la denominada paso a paso (*stepwise*). En la misma en cada paso una variable incluida en el modelo en un paso previo, puede ser excluida del mismo, análogamente una variable excluida del modelo en un paso previo puede llegar a formar parte del mismo en un paso posterior.

Nótese que en cualquiera de las tres modelizaciones anteriores (hacia adelante, hacia atrás, paso a paso) se lleva a cabo una selección de las variables conlas que luego se construye el modelo. En cualquier caso, las tres modelizaciones anteriores constituyen desde un punto de vista de optimización, métodos locales, basados en heurísticos deterministas.

Tal y como se refleja en la tabla 9.3, en los paradigmas provenientes del Aprendizaje Automático, la búsqueda sobre el espacio de posibles modelos puede llevarse a cabo por medio de heurísticos no deterministas, aunque obviamente si este es el caso, el costo computacional asociado aumenta considerablemente.

- Transparencia
  En términos generales podemos decir que los paradigmas clasificatorios provenientes del Aprendizaje Automático son más transparentes y fácilmente interpretables que aquellos cuyo origen se sitúa en la Estadística.

- Validación
  Si bien algunos procedimientos estadísticos utilizan métodos de validación para testar los modelos inducidos, lo habitual es que la validación se utilice tan sólo en los paradigmas provenientes del Aprendizaje Automático. Nótese que la modelización estadística, tal y como se ha explicado anteriormente, evita modelos que se sobreajusten a los datos, cuestión que puede peligrar en el caso de la modelización por Aprendizaje Automático.

- Selección de variables
  Tal y como se ha explicado en el punto relacionado con la búsqueda de modelos, en la modelización habitual en Estadística se lleva a cabo, al unísono, una selección de variables. Conviene tener presente que para seleccionar dichas variables no se tiene en cuenta el porcentaje de bien clasificados obtenido con el modelo, sino la verosimilitud del mismo. La selección de variables que está guiada por un criterio que no coincide con la finalidad del modelo, se denominada aproximación *filter*, y es una característica de la selección de variables en Estadística.

Esta aproximación contrasta con la denominada *wrapper*, mas propia de paradigmas de Aprendizaje Automático, en la cual la selección de los subconjuntos de variables se lleva a cabo teniendo en cuenta la capacidad clasificatoria del modelo que se construye con dicho subconjunto de variables.

4 Notas bibliográficas

Una referencia general y de fácil lectura es Weiss y Kulikowski (1991). Hand (1999) y Glymour et al. (1996) tratan las similitudes y diferencias entre métodos con raices estadísticas y métodos provenientes del Aprendizaje Automático.

5 Recursos en internet

- http://gubbio.cs.berkeley.edu/mlpapers
  Página web del *Computer Science Department* de la Universidad de Berkeley, donde se puede acceder a un gran número de trabajos relacionados con el Aprendizaje Automático (*Machine Learning*)

Lista de trabajos optativos

Escribir un informe que recoja los trabajos encontrados en la página web anterior sobre la aplicación de métodos de Aprendizaje Automático a un problema de tu interés.

Referencias

1. D. J. Hand (1999). Statistics and data mining: intersecting disciplines. *ACM SIGKDD*, Vol. 1, Issue 1, 16–19.

2. C. Glymour, D. Madigan, D. Pregibon, P. Smyth (1996). Statistical inference and data mining. *Communications of the ACM*, **39**, 35–41.

3. S. M. Weiss, C. A. Kulikovski (1991). *Computer Systems that Learn.* Morgan Kaufman.