Métodos de la Minería de Datos

Dr. Oldemar Rodríguez Rojas

1 de noviembre de 2005

Contents

Clasificación jerárquica ascedente

1. Introducción

Para utilizar plenamente los métodos de análisis descriptivo de datos multivariados se requiere la elaboración de histogramas, mapas, árboles, índices,...
Con vistas a la producción de esos materiales es necesario formular los algoritmos correspondientes de una forma precisa, de manera que sea posible su implementación computacional. Es claro esntonces que la herramienta por excelencia del análisis de datos es el instrumento informático, y la posibilidad de que sea utilizado por los profesionales no matemáticos depende de que dispongan de un software *amigable.*
En este artículo presentamos una descripción de la teoría, los métodos, algoritmos, y de la herramienta informática, atinente a la clasificación automática ascendente. Los principales elementos téoricos, metodológicos y algorítmicos se presentan en la segunda y tercera secciones. En la cuarta sección se describen los aspectos informáticos principales que intervienen en el diseño y creación del sistema *Pimad-Clasifica* en ambiente Windows. Finalmente en la quinta sección se incluye un ejemplo de ilustración.

2. Definiciones básicas

Sea X la matriz de datos cuyas n filas o p columnas, forman el conjunto del cual se busca una buena partición. Supondremos que X es una matriz de n individuos por p variables continuas, una tabla de contingencia, o alguna otra forma de datos asimilables a los anteriores.

2.1. Disimilitudes y agregaciones

Con el propósito de encontrar una clasificación de las filas o de las columnas de X , el primer problema a resolver es cómo cuantificar la similitud entre esos objetos o entre grupos de objetos.

2.2. Indices de disimilitud

Un índice de disimilitud entre un conjunto de objetos I (filas de la tabla de datos) es una función d tal que

$$d : I \times I \longrightarrow [0, +\infty[$$

y

$$d(x, y) = d(y, x) \text{ para todo } x, y \in I.$$

También llamamos a esta función, por abuso de lenguaje, una distancia. La distancia seleccionada depende, en general, de la naturaleza de los datos. Así, en nuestro caso tendremos tres clases de distancias:

- *Distancia euclídea clásica*: supongamos que $x_i = (x_{i1}, \ldots, x_{ip})$ y $x_s = (x_{s1}, \ldots, x_{sp})$ son dos filas cualesquiera de la matriz X, entonces la distancia euclídea diagonal es

$$d(x_i, x_s) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{sj})^2}$$

  Este tipo de disimilitud se usa comunmente cuando las variables observadas son continuas.

- *Distancia euclídea de las varianzas*: cuando las variables tienen varianzas muy desiguales, la magnitud del término $(x_{ij} - x_{sj})^2$ puede depender de la varianza σ_j^2 de la variable x^j, haciendo depender la distancia entre filas, de la estructura de varianzas más que de la estructura de correlaciones. Para corregir este efecto se usa la fórmula

$$d(x_i, x_s) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sigma_j^2} (x_{ij} - x_{sj})^2}$$

  Obsérvese que lo anterior equivale a dividir cada columna x^j por su desviación estándar σ_j y usar la distancia euclídea clásica sobre los datos así transformados.

- *Distancia Chi-dos* (χ^2): si X es una tabla de contingencia o una tabla en la cual tiene sentido la suma por filas y por columnas, se acostumbra usar el índice Chi-dos en virtud de las propiedades que posee. El cuadrado de esta distancia se define por la fórmula

$$\chi^2(p_i, p_s) = \sum_{j=1}^{p} \frac{1}{c_j} \left(\frac{x_{ij}}{f_i} - \frac{x_{sj}}{f_s} \right)^2$$

donde $p_i = (\frac{x_{i1}}{f_i}, \ldots, \frac{x_{ip}}{f_i})$ es el i-ésimo perfil de fila, $c_j = \sum_{i=1}^{n} x_{ij}$ el total de la columna x^j, y $f_i = \sum_{j=1}^{p} x_{ij}$ es el total de la fila x_i.

A cada individuo se le asigna el peso $p_i = \frac{f_i}{T}$. Si el interés fuese hacer un análisis de las columnas entonces se trabaja con la tabla transpuesta y todas las definiciones anteriores valen. El índice Chi-dos goza de la propiedad de reducibilidad, la cual, grosso modo, expresa: s*i dos o más columnas de X son muy correlacionadas, entonces se pueden sustituir por su suma, y la distancia Chi-dos entre las nuevas filas es aproximadamente igual a las distancias entre las filas originales.* Cuando la correlación es perfecta, las distancias se mantienen invariables. Lo anterior significa que el índice χ^2 elimina la contribución repetitiva de las variables correlacionadas, sobre la magnitud de las distancias.

2.3. Índices de agregación

Para cuantificar la similitud entre grupos de objetos del conjunto a clasificar, se usan unas funciones llamadas índices de agregación o, simplemente, agregaciones.
Una agregación es una función δ tal que

$$\delta : P(I) \times P(I) \longrightarrow [0, +\infty[$$

$$\delta(x, x) = 0 \ \ \forall\, x \in P(I)$$

$$\delta(x, y) = \delta(y, x),$$

donde $P(I)$ es el conjunto de partes de I, no vacías y disjuntas dos a dos. Se ofrece a continuación una lista de agregaciones:

1. *Agregación de Ward*:

$$\delta_w\,(x, y) = \frac{|x| \cdot |y|}{|x| + |y|}\, \|g_x - \ g_y\|^2$$

donde g_x y g_y son el baricentro de $x \in P(I)$ y $y \in P(I)$ respectivamente.

2. *Agregación del salto mínimo*:

$$\delta_{\min}\,(x,\, y) = \min\{d\,(h,\, k) \mid h \in x \text{ y } k \in y\}\,.$$

3. *Agregación del salto máximo*:

$$\delta_{\max}\,(x, y) = \max\{d\,(h,\, k) \mid h \in x \text{ y } k \in y\}$$

4. *Agregación del promedio de las disimilitudes*:

$$\delta_{\text{prom}}(x, y) = \frac{1}{|x| \ |y|} \sum \{d(h, k) \mid h \in x \text{ y } k \in y\}.$$

Se encuentran en la literatura especializada una gran cantidad de fórmulas de agregación. Gracias al Francés *Michel Jambu* disponemos de una fórmula general unificadora de esta diversidad, la cual transcribimos:

$$\begin{aligned} \delta(x \cup y, z) \quad = \quad & a_1\delta(x, z) + a_2\delta(y, z) + a_3\delta(x, y) + a_4 f(x) \\ & + a_5 f(y) + a_6 f(z) + a_7 \left|\delta(x, z) - \delta(y, z)\right| \end{aligned}$$

donde $a_1, \ldots, a_7$ son unas constantes que dependen de la agregación δ.

1. *Agregación de Ward*: Los a_i son: $a_1 = \frac{|x|+|z|}{|x|+|y|+|z|}$, $a_2 = \frac{|y|+|z|}{|x|+|y|+|z|}$, $a_3 = -\frac{|z|}{|x|+|y|+|z|}$, $a_i = 0$; $i = 4, 5, 6, 7$.

2. *Agregación del salto mínimo*: Los a_i son: $a_1 = a_2 = \frac{1}{2}$, $a_i = 0$ para $i = 3, 4, 5, 6$ y $a_7 = -\frac{1}{2}$ y $\delta_{\min}(\{h\}, \{k\}) = d(h, k)$.

3. *Agregación del salto máximo*: Los a_i son: $a_1 = a_2 = \frac{1}{2}$, $a_7 = \frac{1}{2}$ y $a_i = 0$ si $i = 3, 4, 5, 6$.

4. *Agregación del promedio de las disimilitudes*: Los a_i son: $a_1 = \frac{1}{|x| + |z|}$, $a_2 = \frac{1}{|y| + |z|}$ y $a_i = 0$ para $i = 3, 4, 5, 6, 7$.

3. Jerarquías binarias

Una jerarquía binaria sobre un conjunto de objetos denotado por I es una colección H de partes no vacías de I, llamadas nodos o clases que poseen las siguientes propiedades:

- $\{x\} \in H$ para todo $x \in I$.
- $I \in H$.
- Para todo $x \in H$ tal que $\text{card}(x) > 1$, existen y, $z \in H$ tales que $x = y \cup z$ y $y \cap z = \Phi$. Esto significa que toda parte de la jerarquía H, con más de un elemento, es la unión disjunta de dos partes pertenecientes también a H.

Para construir una jerarquía binaria se utiliza el *algoritmo general* que resumimos así: Sea $I = \{1, \ldots, n\}$ el conjunto del cual se busca una buena partición.

1. *Inicialización*: el procedimiento empieza con las clases que se reducen a un solo elemento, es decir con la partición $P_h = \{\{1\}, \ldots, \{n\}\}$, con $h = 0$.

2. *Formación de nuevos nodos*: se fusionan los dos nodos de P_h más cercanos en el sentido de la agregación δ. Es decir, si x y y son estos dos nodos entonces $\delta(x, y) = \min\{\delta(l, k) \mid l, k \in P_h\}$.

3. *Actualización de* P_h: sea $h \leftarrow h + 1$ y $P_h \leftarrow [P_h \cup \{x \cup y\}] - \{x,\ y\}$.

4. *Test*: Si $h < n - 2$ regresar a 2. En otro caso, hacer la última fusión y terminar.

Una jerarquía binaria se llama débilmente indexada si existe una función $f : H \longrightarrow [0, +\infty[$ con las siguientes propiedades:

1. $f(x) = 0 \ \forall\ x \in H$ tal que card $(x) = 1$.

2. $f(x) \leq f(y) \ \forall\ x,\ y \in H$ tales que $x \subseteq y$.

- Si para una función f asociada a una jerarquía falla la segunda propiedad, decimos que la jerarquía tiene inversiones.

- La forma práctica de construir jerarquías débilmente indexadas por medio del algoritmo general es definiendo

$$f(x \cup y) = \max\{f(x), f(y), \delta(x, y)\}.$$

- Para esta definición, E. Diday obtuvo varios resultados relativos a la existencia e inexistencia de inversiones. Los resultados más elaborados se refieren a dar condiciones sobre los valores de los a_i de la fórmula de Jambu.

Se ofrece a continuación una lista de agregaciones con información sobre los correspondientes a_i y la existencia de inversiones.

1. *Agregación de Ward*: $\delta_w(x,\ y) = \frac{q_x\ q_y}{q_x\ +\ q_y} \|g_x -\ g_y\|^2$ donde q_x y g_x son el peso y el baricentro de x respectivamente. Los a_i son: $a_1 = \frac{q_x\ +\ q_z}{q_x\ +\ q_y\ +\ q_z}$, $a_2 = \frac{q_y\ +\ q_z}{q_x\ +\ q_y\ +\ q_z}$, $a_3 = -\frac{q_z}{q_x\ +\ q_y\ +\ q_z}$, $a_i = 0$; $i = 4, 5, 6, 7$.

2. *Agregación de la inercia*: $\delta_I(x,\ y) = \sum_{k \in x \cup y} p_k \|k - g_{x \cup y}\|^2$, los a_i son: $a_1 = \frac{q_x + q_z}{q_x\ +\ q_y\ +\ q_z}$, $a_2 = \frac{q_y + q_z}{q_x\ +\ q_y\ +\ q_z}$, $a_3 = \frac{q_x + q_y}{q_x\ +\ q_y\ +\ q_z}$, $a_4 = -\frac{q_x}{q_x\ +\ q_y\ +\ q_z}$, $a_5 = -\frac{q_y}{q_x\ +\ q_y\ +\ q_z}$, $a_6 = -\frac{q_z}{q_x\ +\ q_y\ +\ q_z}$ y $a_7 = 0$.

3. *Agregación de la varianza*: $\delta_{\text{var}}(x,y) = \frac{1}{q_x + q_y}\delta_I(x,\, y)$ que es una variante de la agregación anterior. Lo que se quiere es tomar en cuenta el tamaño de los nodos, puesto que entre clases con inercia del mismo orden de magnitud, son mejor agrupadas las de mayor cardinalidad. Los a_i se deducen inmediatamente de los de δ_I, multiplicando cada uno por el factor $\frac{1}{q_x + q_y + q_z}$.

4. *Agregación del aumento ponderado de la varianza*: $\delta_{aum}(x,\, y) = \text{var}(x \cup y) - \frac{q_x}{q_x+q_y}\text{var}(x) - \frac{q_y}{q_x+q_y}\text{var}(y)$ lo que naturalmente es igual a $\frac{1}{q_x+q_y}\delta_w(x,\, y)$. Los a_i se deducen de los a_i de δ_w, multiplicando cada uno por $\frac{1}{q_x+q_y+q_z}$.

5. *Agregación de la diferencia de los centros de gravedad*: $\delta_{cg}(x,\, y) = \|g_x - \ g_y\|^2$, los a_i son: $a_1 = \frac{q_x}{q_x + q_y}$, $a_1 = \frac{q_y}{q_x + q_y}$, $a_3 = -\frac{q_x\ q_y}{(q_x + q_y)^2}$.

6. *Agregación del salto mínimo*: $\delta_{\min}(x,\, y) = \min\{d(h,\, k) \mid h \in x \text{ y } k \in y\}$, los a_i son: $a_1 = a_2 = \frac{1}{2}$, $a_i = 0$ para $i = 3, 4, 5, 6$ y $a_7 = -\frac{1}{2}$ y $\delta_{\min}(\{h\},\ \{k\}) = d(h, k)$.

7. *Agregación del salto máximo*: $\delta_{\max}(x, y) = \max\{d(h,\, k) \mid h \in x \text{ y } k \in y\}$, los a_i son: $a_1 = a_2 = \frac{1}{2}$, $a_7 = \frac{1}{2}$ y $a_i = 0$ si $i = 3, 4, 5, 6$.

8. *Agregación del promedio de las disimilitudes*: $\delta_{\text{prom}}(x, y) = \frac{1}{|x|\ |y|}\sum\{d(h,\, k) \mid h \in x \text{ y } k \in y\}$, los a_i son: $a_1 = \frac{1}{|x| + |z|}$, $a_2 = \frac{1}{|y| + |z|}$ y $a_i = 0$ para $i = 3, 4, 5, 6, 7$.

Para las agregaciones $\delta_w, \delta_I, \delta_{\text{var}}, \delta_{\text{aum}}$ y δ_{cg} se usa una métrica $\|\cdot\|$ euclídea. En las otras d es cualquier índice de disimilitud. En cuanto a la existencia de inversiones en la jerarquía binaria edificada con el algoritmo general, tenemos:

- $\delta_w, \delta_I, \delta_{\text{aum}}, \delta_{\min}, \delta_{\max}$ y δ_{prom} producen jerarquías sin inversiones.
- δ_{var} y δ_{cg} pueden producir jerarquías con inversiones.