Inteligencia Artificial

Programa de Ingeniería de Sistemas

Tema: Análisis de Conglomerados

LuzStellaRoblesPedrozo

Análisis de Conglomerados

Definición

Es un conjunto de objetos que poseen características similares.

Características

- La palabra conglomerado es la traducción más cercana al término 'Cluster'
- El análisis de conglomerados busca particionar un conjunto de objetos en grupos, de tal forma que los objetos de un mismo grupo sean similares y los objetos de grupos diferentes sean disímiles

Análisis de Conglomerados

Los propósitos más frecuentes para la construcción y análisis de conglomerados son los siguientes:

- La identificación de una estructura natural en los objetos
- La búsqueda de esquemas conceptuales útiles que expliquen el agrupamiento de algunos objetos
- La verificación de hipótesis, o la confirmación de si estructuras definidas mediante otros procedimientos están realmente en los datos

Análisis de Conglomerados

Un psicólogo clínico emplea una muestra de un determinado número de pacientes alcohólicos admitidos a un programa de rehabilitación, con el fin de construir una clasificación. Los datos generados sobre estos pacientes se obtienen a través de una prueba. La prueba contiene 566 preguntas de respuestas dicotómicas, las cuales se estandarizan y resumen en 13 escalas que dan un diagnóstico. Mediante una medida de similitud y la consideración de homogeneidad dentro y entre grupos, se conformaron cuatro grupos de alcohólicos: (1) emocionalmente inestables de personalidad, (2) psiconeuróticos con ansiedad-depresión,(3) de personalidad psicópata (4) alcohólico con abuso de drogas y características paranoicas.

Ejemplo 1

Análisis de Conglomerados

En taxonomía vegetal, el análisis de conglomerados se usa para identificar especies con base en algunas características morfológicas, fisiológicas, químicas, etológicas, ecológicas, geográficas y genéticas. Con esta información se encuentran algunos conglomerados de plantas, dentro de los cuales se comparten las características ya indicadas.

Ejemplo 2

Análisis de Conglomerados

El análisis de conglomerados puede emplearse con propósitos de muestreo. Así por ejemplo, un analista de mercados está interesado en probar las ventas de un producto nuevo en un alto número de ciudades, pero no dispone de los recursos ni del tiempo suficientes para observarlos todos. Si las ciudades pueden agruparse en conglomerados, un miembro de cada grupo podria usarse para la prueba de ventas; de otra parte, si se generan grupos no esperados esto puede sugerir alguna relación que deba investigarse.

Ejemplo 3

Análisis de Conglomerados

Para alcanzar los propósitos ilustrados anteriormente, se deben considerar los siguientes aspectos:

- Cómo se mide la similitud?

  Comparación y registro de la proximidad entre pares de objetos de tal forma que la distancia entre las observaciones indique la similitud.

Análisis de Conglomerados

Para alcanzar los propósitos ilustrados anteriormente, se deben considerar los siguientes aspectos:

- Cómo se forman los conglomerados?

  Método o procedimiento mediante el cual se agrupan las observaciones que son más similares dentro de un determinado conglomerado.

Análisis de Conglomerados

Para alcanzar los propósitos ilustrados anteriormente, se deben considerar los siguientes aspectos:

- Cuántos grupos se deben formar?
  - El criterio decisivo es la homogeneidad 'media', alcanzada dentro de los conglomerados.
  - A medida que el número de conglomerados disminuye, la homogeneidad dentro de los conglomerados necesariamente disminuye.

Análisis de Conglomerados

Son dos los elementos requeridos en el análisis de conglomerados:

- La medida que señale el grado de similitud entre los objetos
- El procedimiento para la formación de grupos o conglomerados

Análisis de Conglomerados

Medidas de Similitud

- Reconocer objetos como similares o disímiles es fundamental para el proceso de clasificación.
- Las medidas de similitud se pueden clasificar en dos tipos:
  - Las que reúnen las propiedades de métrica, como la distancia
  - Los coeficientes de asociación

Análisis de Conglomerados

Medidas de Similitud

Una métrica d es una función (o regla) que asigna un número a cada par de objetos de un conjunto Ω (omega):

$$\Omega \text{ X } \Omega : \xrightarrow{d} \text{R}$$

$$(x, y) \longrightarrow d(x, y)$$

Análisis de Conglomerados

Medidas de Similitud

Lo anterior satisface las siguientes condiciones sobre los objetos x, y, z del conjunto Ω:

1. No Negatividad. d(x, y) = 0, si y sólo si, x = y
2. Simetría. Dados dos objetos x, y, la distancia d, entre ellos satisface:

$$d(x, y) = d(y, x)$$

3. Desigualdad Triangular.

Análisis de Conglomerados

Medidas de Similitud

Lo anterior satisface las siguientes condiciones sobre los objetos x, y, z del conjunto Ω:

3. Desigualdad Triangular. Para tres objetos x, y, z las distancias entre ellos satisface la expresión:

$$d(x, y) \leq d(x, z) + d(z, y)$$

La longitud de uno de los lados de un triángulo es menor o igual que la suma de las longitudes de los otros dos lados.

Análisis de Conglomerados

Medidas de Similitud

Lo anterior satisface las siguientes condiciones sobre los objetos x, y, z del conjunto Ω:

4. Identificación de no Identidad. Dados los objetos x, y:

$$\text{si } d(x, y) \neq 0, \text{ entonces } x \neq y$$

5. Identidad. Para dos elementos idénticos, x, x', se tiene que:

$$d(x, x') = 0$$

Si los objetos son idénticos, la distancia entre ellos es cero.

Análisis de Conglomerados

Medidas de Similitud

Las medidas de similitud, de aplicación más frecuente, son las siguientes:

- Medidas de Distancia
- Coeficientes de Correlación
- Coeficientes de Asociación
- Medidas Probabilísticas de Similitud

Análisis de Conglomerados

Medidas de Similitud

- Antes de utilizar alguna de las medidas anteriores, se debe encontrar el conjunto de variables que mejor represente el conjunto de similitud, bajo el estudio a desarrollar.
- Idealmente, las variables deben escogerse dentro del marco conceptual que explícitamente se usa para la clasificación.
- La teoría en cada campo, es la base racional para la selección de las variables a usar en el estudio.

Análisis de Conglomerados

Medidas de Similitud

Las medidas de similitud, de aplicación más frecuente, son las siguientes:

- Medidas de Distancia
- Coeficientes de Correlación
- Coeficientes de Asociación
- Medidas Probabilísticas de Similitud

Análisis de Conglomerados

Medidas de Distancia

- Distancia Euclidiana, definida por:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}.$$

Análisis de Conglomerados

Medidas de Distancia

- Distancia Euclidiana
- Distancia D^2 de Mahalanobis, definida por:

$$D^2 = d_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$$

Análisis de Conglomerados

Medidas de Distancia

- Distancia Euclidiana
- Distancia D^2 de Mahalanobis
- Distancia de Manhattan, definida por:

$$d_{ij} = \sum_{k=1}^{p} |X_{ik} - X_{jk}|.$$

Análisis de Conglomerados

Medidas de Distancia

- Distancia Euclidiana
- Distancia D^2 de Mahalanobis
- Distancia de Manhattan
- Distancia de Minkowsky, definida por:

$$d_{ij} = \left(\sum_{k=1}^{p} |X_{ik} - X_{jk}|^r \right)^{1/r} \quad \text{con } r = 1, 2, \ldots$$

Análisis de Conglomerados

Medidas de Distancia

Ejemplo

Supóngase que se tienen cuatro personas cuya edad X_1 (en años), estatura X_2 (en metros), peso X_3 (en kilogramos), son los siguientes:

Persona	Edad	Estatura	Peso
A	23	1.69	61
B	40	1.70	72
C	26	1.65	68
D	38	1.68	70

Análisis de Conglomerados

Medidas de Distancia

Ejemplo

La matriz de distancias euclidianas es:

	A	B	C	D
A	0	20.25	7.62	17.49
B	20.25	0	14.56	2.83
C	7.62	14.56	0	12.17
D	17.49	2.83	12.17	0

Donde la distancia entre A y B, por ejemplo, resulta del siguiente cálculo:

$$d_{AB} = \sqrt{(23-40)^2 + (1.69-1.70)^2 + (61-72)^2} = 20.25$$

Análisis de Conglomerados

Medidas de Distancia

Ejemplo

La matriz de distancias euclidianas es:

	A	B	C	D
A	0	20.25	7.62	17.49
B	20.25	0	14.56	2.83
C	7.62	14.56	0	12.17
D	17.49	2.83	12.17	0

Se puede notar que los individuos más similares o cercanos son B y D. Resalta fácilmente de los datos.

Análisis de Conglomerados

Medidas de Distancia

Ejemplo

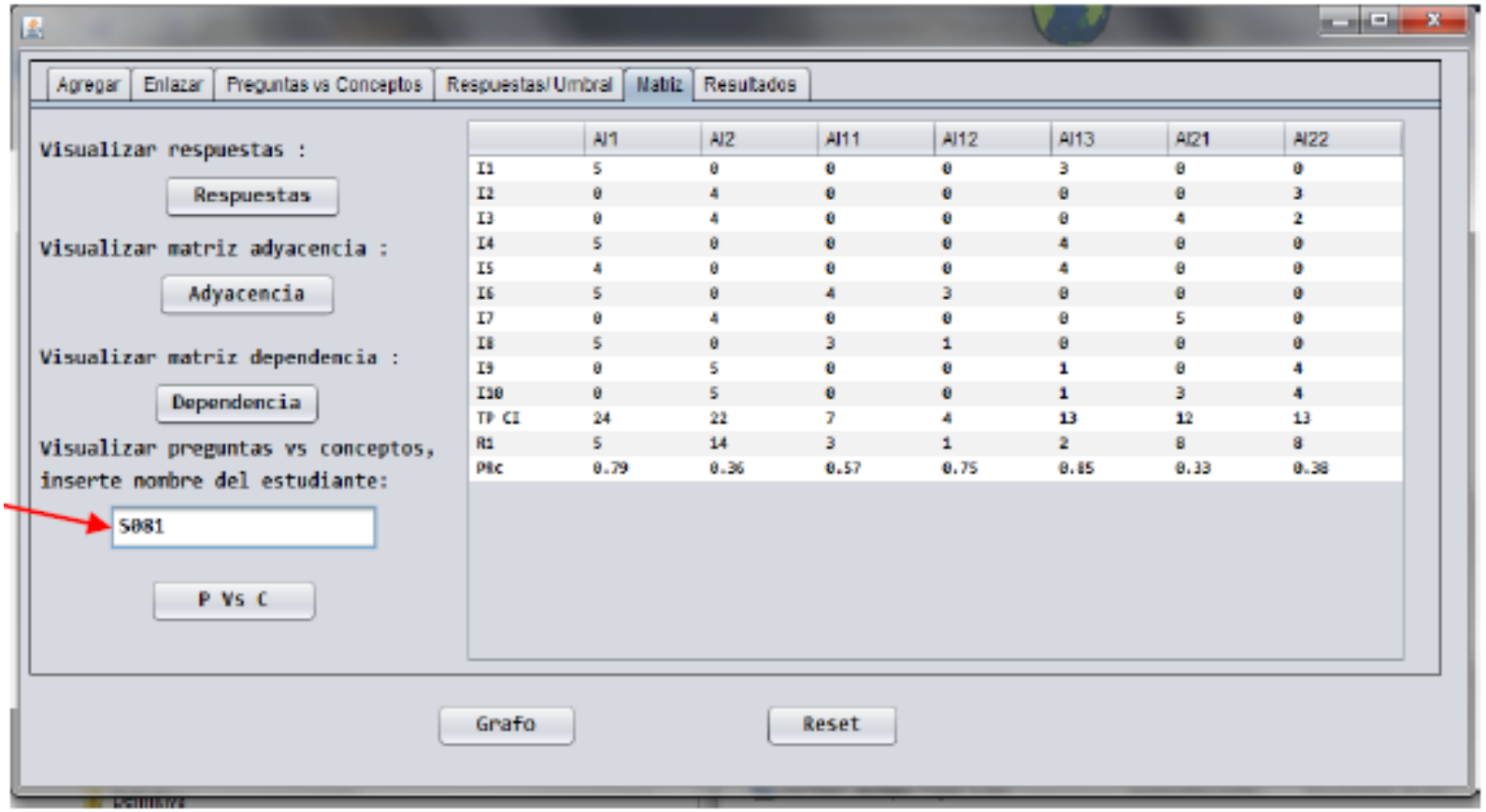
La matriz de distancias de Mahalanobis es:

	A	B	C	D
A	0	7.21	6.36	10.01
B	7.21	0	8.89	15.62
C	6.36	8.89	0	7.96
D	10.01	15.62	7.96	0

La distancia entre B y D, mientras que con la distancia euclidiana B y D son los más cercanos, con la distancia de Mahalanobis resultan con los valores más lejanos.

Análisis de Conglomerados

Cluster sobre resultados de un test



	AI1	AI2	AI11	AI12	AI13	AI21	AI22
I1	5	0	0	0	3	0	0
I2	0	4	0	0	0	0	3
I3	0	4	0	0	0	4	2
I4	5	0	0	0	4	0	0
I5	4	0	0	0	4	0	0
I6	5	0	4	3	0	0	0
I7	0	4	0	0	0	5	0
I8	5	0	3	1	0	0	0
I9	0	5	0	0	1	0	4
I10	0	5	0	0	1	3	4
TP CI	24	22	7	4	13	12	13
R1	5	14	3	1	2	8	8
PBc	0.79	0.36	0.57	0.75	0.85	0.33	0.38

Análisis de Conglomerados

Cluster sobre resultados de un test

Conceptos debiles por sujeto evaluado

S012 A11 A12 A111 A112 A113 A121 A122
S021 A11 A12 A111 A112 A113 A121 A122
S024 A11 A12 A111 A112 A113 A121 A122
S025 A11 A12 A111 A112 A113 A121 A122
S027 A11 A12 A111 A112 A113 A121 A122
S033 A11 A12 A111 A112 A113 A121 A122
S034 A11 A12 A111 A112 A113 A121 A122
S039 A11 A12 A111 A112 A113 A121 A122
S042 A11 A12 A111 A112 A113 A121 A122
S044 A11 A12 A111 A112 A113 A121 A122
S048 A11 A12 A111 A113 A121 A122
S051 A11 A12 A111 A112 A113 A121 A122
⋮
S312 A11 A12 A111 A112 A113 A121 A122
S318 A12 A111 A112 A113 A121 A122
S325 A12 A111 A121 A122
S327 A12 A111 A121 A122
S331 A11 A12 A111 A112 A121 A122
S336 A11 A12 A111 A112 A121 A122
S337 A11 A12 A111 A112 A113 A121 A122
S345 A11 A12 A111 A112 A113 A121 A122
S348 A11 A12 A111 A112 A113 A121 A122
S351 A12 A111 A121 A122
S354 A11 A12 A111 A113 A121 A122
S357 A12 A111 A113 A121 A122
S358 A11 A12 A111 A112 A113 A121 A122
S360 A11 A12 A111 A112 A113 A121 A122

Análisis de Conglomerados

Cluster sobre resultados de un test

Es fundamental:

- El archivo de conceptos débiles por cada evaluado
- El peso total de cada concepto débil dentro del test (TP CI_d)
- Una vez construida la relación ítem-conceptos, con los pesos de cada concepto dentro del ítem asignados, verificar que la suma total del peso de un concepto dentro del test, no sea igual a la suma total del peso de algún otro concepto dentro del test.
- El peso total de los conceptos con falencia en el test (PTcd), por cada sujeto. Calculado como:

$$PTcd = \sum_{x=1}^{cd} (TP\ CI_d)_x$$

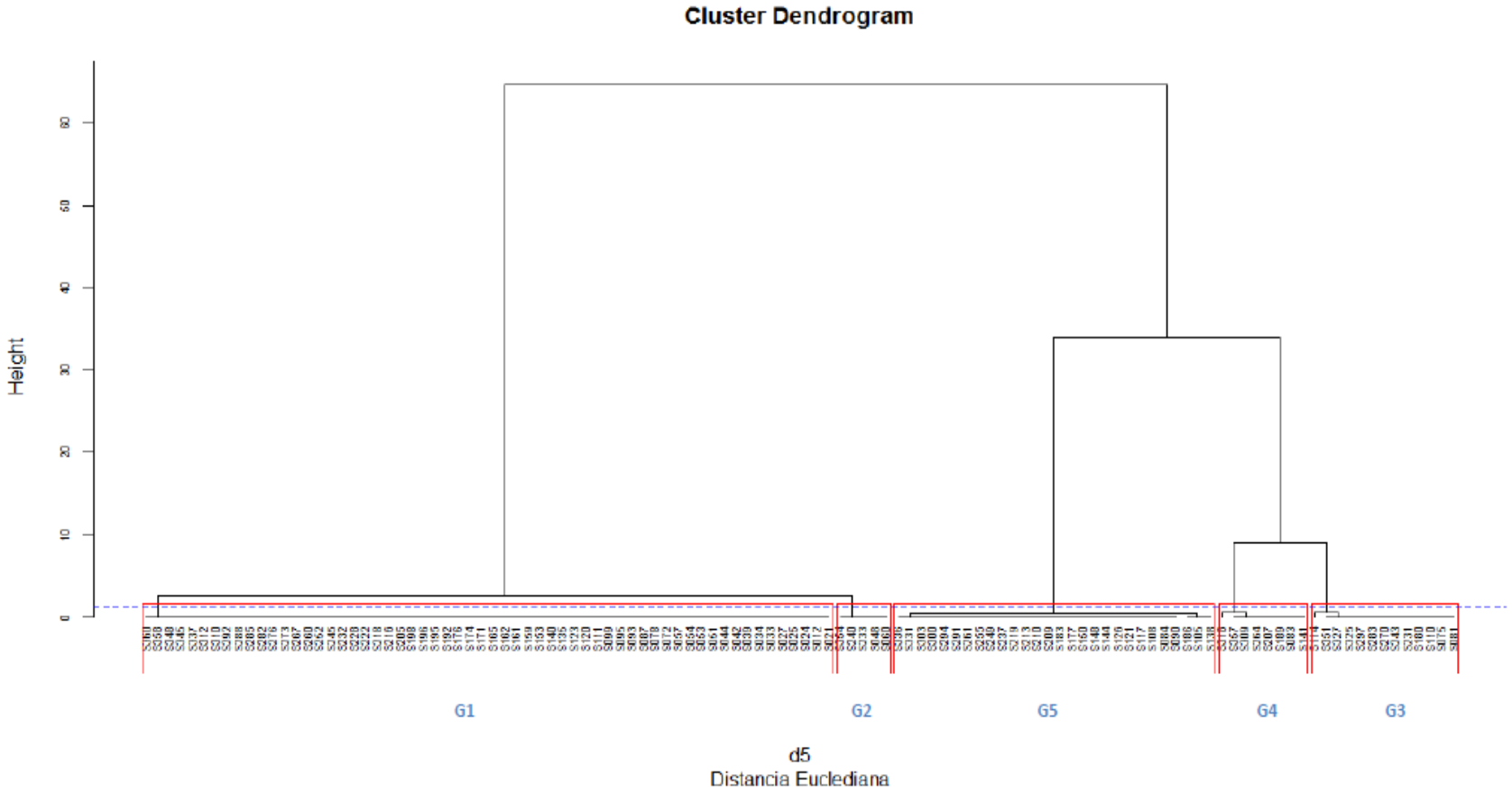
Análisis de Conglomerados

Cluster sobre resultados de un test

```
R> groups=cutree(cl,k=5)
R> groups
S012 S021 S024 S025 S027 S033 S034 S039 S042 S044 S048 S051 S053 S054
   1    1    1    1    1    1    1    1    1    1    2    1    1    1
S057 S060 S072 S075 S078 S081 S083 S084 S087 S090 S093 S095 S099 S105
   1    2    1    3    1    3    4    5    1    5    1    1    1    5
S108 S110 S111 S114 S117 S120 S121 S123 S126 S135 S138 S140 S141 S144
   5    3    1    3    5    1    5    1    5    1    5    1    4    5
S148 S150 S153 S159 S161 S162 S165 S171 S174 S176 S177 S180 S183 S186
   5    5    1    1    1    1    1    1    1    1    5    3    5    5
S189 S192 S195 S196 S198 S205 S207 S209 S210 S213 S216 S218 S219 S222
   4    1    1    1    1    1    4    5    5    5    1    1    5    1
S228 S231 S232 S233 S237 S240 S243 S245 S249 S252 S255 S260 S261 S264
   1    3    1    2    5    2    3    1    5    1    5    1    5    4
S267 S270 S273 S276 S282 S283 S285 S288 S291 S292 S294 S297 S300 S303
   1    3    1    1    1    3    1    1    5    1    5    3    5    5
S309 S310 S312 S318 S325 S327 S331 S336 S337 S345 S348 S351 S354 S357
   4    1    1    4    3    3    5    5    1    1    1    3    2    4
S358 S360
   1    1
```

Análisis de Conglomerados

Cluster sobre resultados de un test



Actividad Extra-clase

Práctica

Exposición!

Referencias

Huang, S. X. A content-balanced adaptive testing algorithm for computer-based training systems. En C. Frasson, G. Gauthier y A. Lesgold (Eds.), *Lecture notes in computer science 1086. Proceedings of the 3rd international conference on intelligent tutoring systems*. Its 1996 (pp. 306-314). New York: Springer Verlag. 1996.

Huapaya, C., Lizarralde, F., Vivas, J., Arona, G. Modelo de Evaluación del Conocimiento en un Sistema Tutorial Inteligente. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, No. 2. 2007.

Hwang, G. J. A test sheet generating algorithm for multiple assessment requirements, IEEE Trans. Educ., vol. 46, no. 3, pp. 329–337. Aug. 2003.

Hwang, G. J. A concept map model for developing intelligent tutoring systems. Computers & Education, 40(3), 217-235. 2003a.

Hwang, G. J., Hsiao, J. L., & Tseng, J. C. R. A computer-assisted approach for diagnosing student learning problems in engineering courses. Journal of Information Science and Engineering, 19(2), 229-248. 2003b.

Hwang, G. J. A Data Mining Algorithm for Diagnosing Student Learning Problems in Science Courses. International Journal of Distance Education Technology, 3(4), 35-50. 2005.

Gracias!