Reducción de Dimensiones en Bases de Datos usando el Enfoque Lógico Combinatorio

Jorge Ochoa Somuano, Gerardo Reyes Salgado

Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
{somuano, greyes}@cenidet.edu.mx

Abstract. El desarrollo de técnicas que permitan reducir el espacio de representación es un tema en el cual se ha puesto especial interés, ya que esto nos permite trabajar con menos variables y obtener los mismos resultados. Estas técnicas se utilizan principalmente para el tratamiento de grandes bases de datos en las cuales se maneja una inmensa cantidad de datos numéricos y cuyo objetivo final es la clasificación de dichos datos. En este artículo se presenta el resultado de la implementación de las técnicas "testores típicos", para realizar una reducción de dimensiones y "k-medias", para llevar a cabo la clasificación de objetos.

1. Introducción

El mayor problema que se presenta al utilizar datos n-dimensionales es la redundancia de la información, es decir, en muchos casos diferentes variables describen a un mismo objeto o grupo de objetos. Con la reducción de dimensiones (selección de variables) se trata de describir la información original de forma sintética o resumida. Se busca la simplicidad con base en conseguir una reducción de la complejidad del problema.

De manera general, los métodos para reducir el número de variables que se utilizan para la representación de un conjunto de datos consisten en condensar las variables originales en un número menor de nuevas variables creadas por el propio análisis, que contienen sin embargo gran parte de la información original. A este tipo de objetivo, y de método, se denomina *reducción de la dimensión* ya que los datos originales se expresan en un espacio de dimensión n mientras que como resultado del análisis podemos expresarlos en otro espacio de p, donde $p < n$ [2].

Uno de los objetivos de reducir dimensiones en un conjunto de objetos, es poder realizar la clasificación de dichos objetos de una manera más rápida y fácil. La clasificación es el primer paso para la comprensión de un fenómeno complejo. En algunos casos el interés está en determinar sobre el conjunto de objetos, un conjunto de clases tan diferenciadas como sea posible, este es el caso del análisis de agrupamiento. En otros casos, los grupos se diferencian de una forma natural e interesa determinar las

David Juárez Romero, J. Crispín Zavala-Díaz,(eds.). AGECOMP2004, Memorias del 3er Congreso de Cómputo AGECOMP, UAEM, México, ISBN(e) 968-878-205-X., pp. 35 - 45

características que provocan la separación existente entre los grupos, aquí se habla del análisis discriminante [8].

2. Análisis N-Dimensional

El análisis multivariante o n-dimensional es "la rama de la estadística que estudia las relaciones entre conjuntos de variables dependientes y los individuos para los cuales se han medido dichas variables" [5].

Un dato de tipo multidimensional es aquel que tiene n-dimensiones (atributos o características), que lo diferencian de otros datos o que lo asemejan a datos que comparten las mismas características. A partir de estos atributos surgen ciertas técnicas, las cuales se utilizan para el análisis de este tipo de datos.

Estas técnicas tratan de visualizar y simplificar la estructura de los datos. En unos casos, se pretende proyectar los puntos dados en un espacio de dimensión inferior, conservando ciertas propiedades del conjunto inicial de datos, especialmente las referidas a las distancias entre individuos.

Entre las técnicas estadísticas que pueden emplearse con un enfoque exploratorio, sin duda las de mayor utilidad son las que pueden englobarse dentro del análisis multidimensional, caracterizado por el estudio conjunto de un grupo de *n* variables, medida cada una de ellas sobre *m* objetos [8].

En otros casos, se trata de hacer un cambio de base. Las observaciones originales de los objetos están expresadas en la base formada por las variables, que pueden o no estar correlacionadas, es decir, en general no es una base ortogonal. Al transformar el conjunto inicial de variables en otro incorrelado, se hace un cambio de base a una nueva ortogonal.

Entre los métodos que se pueden incluir en este tipo se hallan *el análisis de componentes principales, análisis de componentes independientes, análisis de componentes curvilíneos análisis factorial y testores típicos,* por mencionar algunos. Al igual que mediante una fotografía se logra reconocer los objetos tridimensionales fotografiados, con estos métodos se consigue reconocer ciertas propiedades de los conjuntos de datos originales.

Para el caso de estudio en este trabajo se utilizó la técnica de testores típicos, que es parte del enfoque lógico combinatorio.

2.1. Teoría de Testores

La teoría de testores se formuló como una de las direcciones científicas independientes de la Cibernética Matemática a mediados de los años 60, en la desaparecida Unión de Repúblicas Socialistas Soviéticas (URSS). Su origen, en 1954-1955, está vinculado a la utilización de métodos lógicos matemáticos para la localización de desperfectos en los circuitos eléctricos [9].

En 1965, se abre una línea de aplicación de la Teoría de Testores a los problemas clásicos del Reconocimiento de Patrones [9].

De acuerdo a la definición de Zhuravliov [9], un testor es un conjunto de rasgos (columnas) que permite diferenciar entre dos clases, por que ningún objeto de la clase T_0 se confunde con objeto alguno de la clase T_1. Un testor se llama irreducible (típico) si al eliminar cualquiera de dichas columnas deja de ser testor para (T_0, T_1).

La selección de variables, tiene dos usos principales:

- Reducir el número de rasgos en términos de los cuales se deben describir los objetos en modo eficiente.
- Encontrar los rasgos que inciden en el problema de manera determinante.

El término "irreducible", refleja con claridad la idea que no pueden eliminarse más columnas, el término "típico" tiene una intención más en el sentido de la modelación matemática, y refleja el hecho, que la combinación de rasgos que forman un testor típico tiene, en cierto sentido, la misma idea de la "tipicidad" para una clase de objetos, es decir, un conjunto de rasgos que en cierto sentido tipifican una clase de objetos y en otro sentido los diferencian de las demás clases [6].

3. Técnicas de Clasificación

Los algoritmos de clasificación o *clustering* se refieren a la capacidad de creación de *clusters*, es decir, clases o patrones. La única información que requieren los algoritmos de clasificación es la definición previa del vector de características. Algunos de estos algoritmos, precisan conocer también el número de clases [7].

Las técnicas de clasificación se utilizan cuando no existe un conocimiento suficiente acerca de las clases en que se pueden distribuir los objetos de interés, es decir, en donde no se encuentran bien definidas las clases. A estas técnicas se les conoce como no supervisadas o autoorganizadas, como son: el algoritmo de las distancias encadenadas, el algoritmo Max-Min, etc.

No obstante, también son de gran interés práctico en algunas áreas en las que existe un conocimiento completo de las clases y por tanto, se pueden aplicar los métodos supervisados, como pueden ser: K-Medias e ISODATA, por mencionar algunos.

En este trabajo se utilizó la técnica de clasificación k-medias, con el fin de determinar la pertinencia al aplicar la técnica de testores típicos a una base de datos.

3.1. Algoritmo K-Medias

Este algoritmo hace referencia a que existen *k* clases o patrones, siendo necesario, por tanto, conocer *a priori* el número de clases existentes.

Es un algoritmo sencillo y muy eficiente, siempre que el número de clases se conozca *a priori* con exactitud. Es decir, es muy sensible al parámetro *k*. Un valor de *k* superior al número real de clases dará lugar a clases ficticias, mientras que un *k* inferior producirá menos clases de las reales [7].

El método k-medias permite procesar un número ilimitado de objetos, pero sólo permite utilizar un método de agrupación y requiere que se conozca previamente el número de clases que se desea obtener.

El método k-medias ha mostrado ser muy eficiente en sus resultados al realizar clasificaciones adecuadas en muchas aplicaciones prácticas. Sin embrago, el algoritmo del método k-medias requiere tiempo proporcional al producto del número de objetos y al número de grupos por iteración [1].

Está dado por un conjunto de *n* puntos de datos en un espacio *d*-dimensional de tipo real, R^d, y un *k* número entero, el problema es determinar el conjunto de *k* puntos en R^d, llamados *centroides*, para así minimizar la distancia media cuadrada de cada punto a su centro más cercano [4].

4. Implementación de la Teoría de Testores

Para desarrollar esta investigación se implemento la teoría de testores, la cual se encarga de obtener el menor número de variables con las cuales se pueda describir un objeto sin que exista pérdida de información. Si resulta importante disponer de un concepto que se adecue a las condiciones específicas del problema que queremos resolver, es más importante aún el hecho de contar con un algoritmo que nos permita calcular eficientemente todos los testores (típicos) de una matriz dada y para ello se utiliza el algoritmo denominado BT [6].

5. Algoritmo BT para el Cálculo de los Testores Típicos

En términos generales podríamos decir que encontrar todos los testores (típicos) de una matriz dada es equivalente a realizar una búsqueda exhaustiva entre todos los subconjuntos de rasgos que son $2^{|R|}$, siendo R el conjunto de los rasgos. Con el aumento del número de filas y columnas o del número de clases en dicha matriz, el costo en tiempo de este procedimiento puede elevarse hasta el punto de ser prácticamente imposible. Como consecuencia se han realizado estudios en la búsqueda de algoritmos eficientes para el cálculo de todos los testores (típicos), de hecho este problema por si sólo constituye una línea de trabajo en el marco de la Teoría de Testores [6].

A continuación se muestra el algoritmo utilizado en dicha implementación:

1. Se genera la primera lista α no nula de longitud n.
2. Se determina (condición (*)) si la lista generada es una lista testor en la Matriz Básica (MB).
3. Si es lista testor, se aplica la proposición 1.4*. Si es lista testor típico, se imprime α y se aplica la proposición de salto 2^{n-k}-1. Si no es lista testor, se determina la fila de la MB que provoca este hecho (de no ser única se toma la que tenga el último 1 más a la izquierda) y se aplica la proposición correspondiente.
4. Se genera la lista siguiente a las descartadas en virtud del paso 3 y se regresa al paso 2, en caso de que la lista resultante del paso 3 no sea posterior a (1,1,...,1,1).

* Sea α una lista testor (típico) y *k* el subíndice del último 1 en α, entonces los siguientes 2^{n-k}-1 n-uplos son listas testores pero no son típicos.

Note que antes de emplear el algoritmo BT, se deben obtener tanto la matriz de diferencia (MD), como la matriz básica, ya que dicho algoritmo trabaja con esta última. Para conocer la forma de obtener la MD y la MB vea [9].

6. Ejemplo de aplicación

A continuación se presenta un ejemplo aplicando la teoría de testores:

Sea T una matriz (tabla) formada por la descripción de 5 objetos distribuidos en dos clases, los 3 primeros objetos pertenecen a la primera clase (T_0) y los restantes a la segunda (T_1). Esta matriz será conocida como matriz de aprendizaje.

T	X_1	X_2	X_3	X_4	X_5
O_1	1	1	0	1	1

O_2	0	1	0	1	1
O_3	1	0	1	0	1
O_4	1	1	1	1	1
O_5	0	0	0	1	1

Al formar la matriz de diferencia, esta queda de la siguiente manera:

MD	X_1	X_2	X_3	X_4	X_5
O_{14}	0	0	1	0	0
O_{15}	1	1	0	0	0
O_{24}	1	0	1	0	0
O_{25}	0	1	0	0	0
O_{34}	0	1	0	1	0
O_{35}	1	0	1	1	0

Ahora se presenta la matriz básica, la cual se utiliza en el algoritmo BT para determinar los testores típicos:

MB	X_1	X_2	X_3	X_4	X_5
a_1	0	0	1	0	0
a_2	0	1	0	0	0

Como se puede apreciar el conjunto de variables $\{X_2, X_3\}$ es un testor típico para (T_0, T_1). En efecto, después de la eliminación en T de las variables X_1, X_4 y X_5 se puede observar que no existen filas en T_0 iguales a las filas de T_1.

T_0	X_2	X_3
O_1	1	0
O_2	1	0
O_3	0	1

T_1	X_2	X_3
O_4	1	1
O_5	0	0

Como se puede observar la tipicidad de $\{X_2, X_3\}$ es evidente. Sin embargo, $\{X_1, X_5\}$ no es un testor de T.

T_0	X_1	X_5
O_1	1	1
O_2	0	1
O_3	1	1

T_1	X_1	X_5
O_4	1	1

O_5	0	1

Este hecho se puede verificar fácilmente, vea como la fila de O_1 y O_4 u O3 y O_4 son idénticas, al igual que O_2 y O_5.

7. Aporte de Resultados

Para observar y analizar los resultados proporcionados por el algoritmo de testores típicos y compararlos con los resultados que se obtuvieron manualmente, se utilizó el ejemplo anterior, proporcionándole al algoritmo la matriz de aprendizaje. Lo cual nos permitió comprobar la eficiencia de la teoría de testores y la fiabilidad al emplear este tipo de técnicas para la reducción de dimensiones. Por otro lado, se utilizó el algoritmo de las k-medias para realizar la clasificación de los datos utilizando únicamente las variables que el algoritmo de testores típicos consideró serían las máximas variables que describirían de manera correcta los objetos de las diferentes clases.

Los resultados que se obtuvieron después de aplicar ambos algoritmos para todas las variables, así como para las proporcionadas por los testores típicos se presentan en la Fig.1.

Como se puede observar el resultado de la clasificación en ambos casos es exactamente la misma (objetivo de esta investigación), para los dos reportes del k-medias la clasificación se realizó de manera correcta en un 60% y por ende el error fue del 40%. Esto se debe a las características propias de los objetos utilizados para la prueba. Sin embargo, esto no indica ineficiencia por parte del algoritmo de clasificación, así como tampoco del algoritmo de testores típicos. Simplemente indican la existencia de objetos que a pesar de su pertenencia a una clase comparten la mayoría de sus características con las de objetos de otras clases, motivo por el cual son clasificados en una clase diferente de acuerdo a los criterios de clasificación del propio algoritmo.

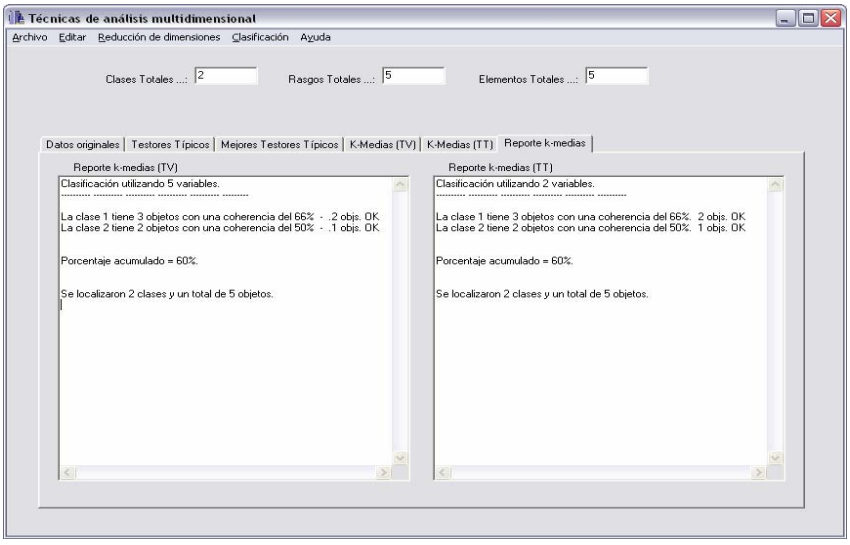

Fig. 1. Pantalla comparativa de resultados en la clasificación empleando todas las variables, así como al usar sólo las variables propuestas por el algoritmo de testores típicos.

8. Testores Típicos Aplicados a la Reducción de Dimensiones en Imágenes

Para evaluar el desempeño y funcionalidad de la teoría de testores con de bases de datos de un mayor número de objetos, variables y clases, se está trabajando con una base de imágenes. Esta base de imágenes contiene 123 imágenes representadas por 21 variables o características de tipo binario. Los datos de las imágenes se dividen en tres categorías o clases.

Las clases de las imágenes son: SpringFlowers; imágenes de flores de verano, SwissMountains; imágenes aéreas de montañas suizas y YellowStone; imágenes del parque nacional de Estados Unidos.

Algunas de las variables utilizadas son: cielo, flores, arbustos, agua, montañas, nieve, etc.

A continuación se muestra el resultado del sistema utilizando la base de imágenes antes mencionada, Fig. 2.

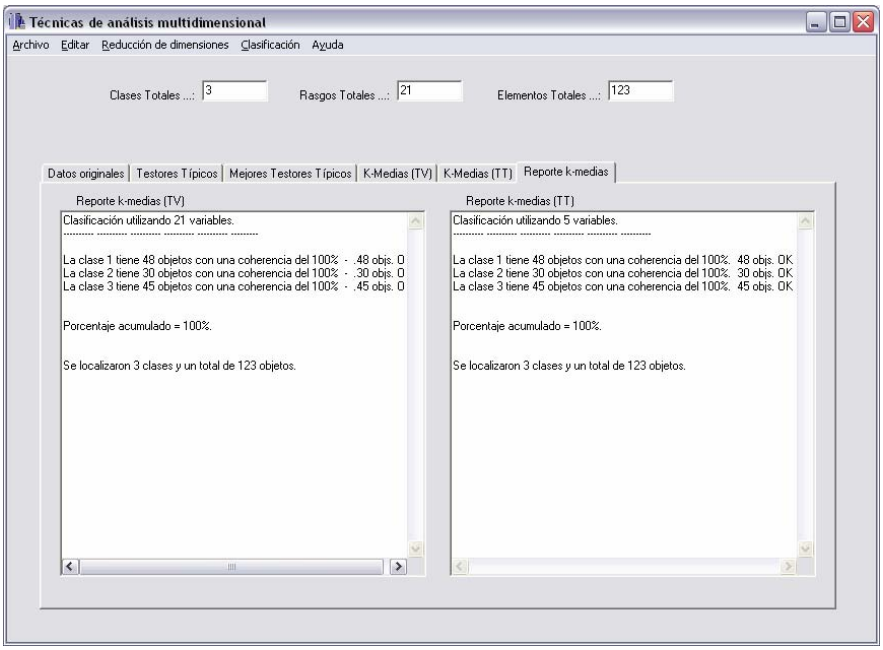

Fig. 2. Pantalla comparativa de resultados en la clasificación de la base de 123 imágenes.

9. Trabajos Futuros

Con el objetivo de ampliar esta investigación se plantea la implementación de la técnica denominada Análisis de Componentes Curvilíneos (CCA, por sus siglas en inglés, *Component Curvilinear Analysis*) [3], para realizar un estudio comparativo de ambas técnicas en la reducción de dimensiones y así determinar la ventaja y desventaja de cada una de ellas.

10. Conclusiones

El interés por facilitar el manejo de los datos para la manipulación de la información se ha convertido en una tarea ardua, en la cual se han realizado un gran número de investigaciones a partir de las cuales se han derivado varias técnicas que persiguen el mismo fin, esto se consigue reduciendo el número de variables que se utilizan para llevar a cabo la descripción de un objeto, a este método se le conoce con el nombre de reducción de dimensiones.

El objetivo de esta investigación es comprobar el funcionamiento de algunas de estas técnicas. En la primera parte de este proyecto de investigación se implementó la teoría de testores con la cual se obtuvieron buenos resultados en la reducción del espacio de representación para el caso de aplicación que se utilizó, esto se comprobó al realizar la clasificación de los objetos con el algoritmo k-medias, el cual presentó un error despreciable, este se dio por las propias características de los objetos utilizados en la prueba, sin embargo no se responsabiliza de esto al algoritmo de testores típicos, ya que aún utilizando todas las variables, el algoritmo de las k-medias muestra los mismos resultados. Finalmente se cumplió el objetivo de esta primera etapa de la investigación: demostrar que se obtienen los mismos resultados en la clasificación con todas las variables al igual que con las utilizadas después de la reducción de dimensiones.

11. Bibliografía

[1] K. Alsabti, S. Ranka, V. Singh, *An Efficient K-Means Clustering Algorithm*, Syracuse University, 2003.

[2] A. I. Ganzo, *Análisis de Datos Multivariantes: Introducción al Análisis Multivariante*, 2003.

[3] J. Hérault, *Curvilinear Component Analysis for High-Dimensional Data Representation: I. Theoretical Aspects and Practical Use in the Presence of noise*, INPG-LIS, Grenoble Cedex, Francia, 2003.

[4] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, A. Y. Wu, *An Efficient K-Means Clustering Algorithm: Analysis and Implementation*, IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol 24, No. 7, 2002.

[5] M. Kendall, *Multivariate Analysis*, Griffin, 1980.

[6] M. Lazo, *Reconocimiento Lógico Combinatorio de Patrones*, Instituto de Cibernética, Matemática y Física, Cuba, 2003

[7] D. Maravall, *Reconocimiento de Formas y Visión Artificial*, Addison-Wesley Iberoamericana, 1994.

[8] J. D. Rodino, C. Batanero, *Enfoque Exploratorio en el Análisis Multivariante de los Datos Educativos*, *Épsilon*, No. 29, pp. 11-22, 1994.

[9] J. R. Shulcloper, A. Guzmán, J. F. Martínez, *Enfoque Lógico Combinatorio al Reconocimiento de Patrones*, Instituto Politécnico Nacional, México, 1999.