Computación y Sistemas Vol. 4 No.3 pp. 264 - 269
© 2001, CIC - IPN. *ISSN 1405-5546* Impreso en México

RESUMEN DE TESIS DOCTORAL

Herramientas para la Estructuración Conceptual de Espacios

Tesista: José Francisco Martínez Trinidad
Centro de Investigación en Computación, IPN
Av. Juan de Dios Bátiz s/n Unidad Profesional Adolfo López Mateos, México D.F, fmartine@cic.ipn.mx

Asesores: José Ruiz Shulcloper, ICIMAT, Cuba
Serguei Levachline, CIC-IPN, México

Resumen

En este trabajo de investigación se propone un conjunto de herramientas (definiciones, teoremas, modelos y algoritmos) para resolver el problema de estructurar o clasificar conceptualmente una muestra de datos.

Se introducen dos modelos (duro y difuso) que permiten resolver el problema de la estructuración conceptual de espacios en un marco teórico en el que las variables utilizadas para describir a los objetos de estudio pueden ser de naturaleza cuantitativa, cualitativa, o ambos y se permite ausencia de información. Se describe cómo pueden obtenerse los algoritmos conceptuales a partir de los modelo propuestos.

También se presenta una nueva formulación del concepto de objeto simbólico de forma tal que los modelos conceptuales propuestos quedan como caso particular de esta nueva formulación. De hecho esta nueva propuesta extiende el marco teórico de tal forma que las herramientas de análisis de datos clásicas también quedan como caso particular de la nueva formulación.

De manera general se puede apreciar que los resultados obtenidos no agotan el desarrollo de esta temática, sino que el trabajo realizado abre nuevas líneas de investigación, que por demás, son de gran importancia en la actualidad para resolver problemas de análisis inteligente de datos.

1 Introducción

En muchas ciencias aplicadas está presente el problema de revelar la estructura subyacente en una colección de objetos (situaciones, medidas, observaciones, fenómenos, etc.). Esta información de los objetos típicamente se encuentra almacenada en archivos planos, bases de datos u algún otro medio electrónico en forma estructurada. El problema de clasificar o estructurar esta información ha sido estudiado intensamente en el área del Reconocimiento de Patrones no supervisado ("cluster analysis"). Los métodos desarrollados en esta área, forman agrupamientos sobre la base de parejas de objetos muy parecidos (o próximos) e ignoran la utilidad del *significado* de los agrupamientos obtenidos, siendo este último un elemento más natural para los especialistas.

Todas las técnicas tradicionales del Reconocimiento de Patrones no supervisado tienen la desventaja de formar agrupamientos los cuales no tienen una interpretación conceptual. El problema del significado de los agrupamientos obtenidos es dejado al especialista. Esta desventaja es significativa ya que el especialista para los fines de su investigación o labor requiere no sólo los agrupamientos, sino además quiere una explicación de ellos en términos humanos.

El agrupamiento conceptual surge a partir de los trabajos de Michalski (1980), en este enfoque se propone encontrar a partir de un conjunto de datos, no sólo los agrupamientos en

los que éstos se estructuran sino además conformar la explicación de tales agrupamientos. El agrupamiento conceptual está compuesto de dos tareas fundamentales: el agrupamiento de entidades en el que se determinan subconjuntos útiles de una muestra de objetos, y la caracterización, la cual determina un concepto para cada subconjunto descubierto por el agrupamiento.

De esta forma, por un problema de *estructuración conceptual de espacios*, debemos entender que es aquél en donde se parte de un conjunto de objetos descritos en términos de un conjunto de rasgos o atributos (la muestra de estudio), y el problema consiste en encontrar, de manera general un cubrimiento de este conjunto de objetos, así como las propiedades o conceptos asociados a los agrupamientos o estructuras encontradas.

En las últimas dos décadas han sido propuestos diferentes algoritmos de agrupamiento conceptual (ver Briscoe and Caelli, 1996), entre los más representativos podemos citar a: EPAM (Feigenbaum, 1963), CLUSTER/2 (Michalski, 1983), UNIMEM (Lebowitz, 1985), COBWEB (Fisher, 1990), CLASSIT (Gennari et al., 1990), COBWEB/3 (McKusick and Thompson, 1990), WITT (Hanson, 1990), LINNEO$^+$ (Béjar y Cortés,1992), K-MEANS CONCEPTUAL (Ralambondrainy, 1995). En esta investigación doctoral se realizó un estudio crítico de los mismos, en donde se muestran sus alcances y limitaciones prácticas (ver, Martínez-Trinidad, 2000a).

Este trabajo está organizado de la siguiente manera: Como primer paso se platea formalmente el problema de la estructuración conceptual de espacios. Posteriormente se presenta un nuevo modelo de algoritmos de agrupamiento conceptual y su generalización al caso difuso. También se aborda el problema de la estructuración conceptual desde el formalismo de la Teoría de los Objetos Simbólicos.

Finalmente se enumera el trabajo que a futuro será desarrollado.

2 Planteamiento Formal del Problema

El problema de la estructuración conceptual de espacios (ver, Martínez-Trinidad, 2000b y Martínez-Trinidad and Guzmán-Arenas, 2001) se formula en los siguientes términos: Sea M un conjunto de objetos. Una *descripción* $I(O)$ es definida para cada objeto $O \in M$ y ésta es representada por una secuencia finita $x_1(O), x_2(O)..., x_m(O)$ de valores de m variables del conjunto $R = \{x_1, x_2..., x_m\}$, con $x_i(O) \in M_i$, siendo M_i el conjunto de valores admisibles de la variable x_i.

Además asumiremos que en M_i hay un símbolo $*$ el cual denota *ausencia de información*, $i=1,\ldots,m$. En otras palabras, la descripción de un objeto puede estar incompleta. Esto es, para al menos una variable no se conoce el valor.

Consideraremos que $I(O) \in M_1 \times M_2 \times,...,\times M_m$ (este producto cartesiano es el espacio de representación inicial *ERI* de los objetos). La naturaleza de las variables o atributos por consecuencia puede ser, simultáneamente, cualquiera (cualitativa: Booleana, multi-valuada, difusa, lingüística y otras o cuantitativas: enteras, reales). Por lo tanto, sobre el *ERI* no supondremos ninguna estructura algebraica o topológica. Consideremos una función $C_i: M_i \times M_i \rightarrow L_i$ tal que

a) $C_i(x_i(O), x_i(O)) = \min_{O' \in M} \{C_i(x_i(O), x_i(O'))\}$ si C_i es un criterio de comparación de disimilaridad entre valores de la variable x_i o

b) $C_i(x_i(O), x_i(O)) = \max_{O' \in M} \{C_i(x_i(O), x_i(O'))\}$ si C_i es un criterio de comparación de similaridad entre valores de la variable x_i para $i=1,\ldots,m$.

C_i es una evaluación del grado de similaridad (o disimilaridad) entre cualquiera dos valores de la variable x_i donde L_i es un conjunto totalmente ordenado, $i=1,\ldots,m$.

Sea una función $\Gamma: (M_1 \times ... \times M_m)^2 \rightarrow L$, donde L es un conjunto totalmente ordenado; Γ será denominada *función de similaridad* y es una evaluación del grado de similaridad entre dos descripciones cualquiera pertenecientes a MI.

Usualmente, la información acerca de los objetos (sus descripciones) está dada en forma de una tabla o matriz $MI = |x_i(O_j)|_{n \times m}$ con n renglones (descripciones de objetos) y m columnas (valores de cada variable en los objetos seleccionados).

El *problema de la estructuración conceptual de espacios sobre M* consiste en determinar el conjunto cubrimiento $\{K_1, K_2,...,K_c\}$ $c>1$, así como el conjunto de conceptos asociados a cada K_i $i=1,\ldots,c.$, en principio K_i podrían ser subconjuntos duros o difusos de M y estos podrían ser disjuntos o no.

3 Modelo Duro de Algoritmos de Agrupamiento Conceptual

A partir de la formulación previa de un problema de estructuración conceptual de universos, fue introducido un nuevo modelo de algoritmos de agrupamiento conceptual para el caso duro, es decir, cuando los agrupamientos encontrados son subconjuntos duros del espacio a estructurar (en el sentido de la teoría clásica de conjuntos).

En todo problema relacionado con la Teoría de Conjuntos, los conjuntos pueden ser determinados de manera extensional o intencional. En el problema de la estructuración conceptual de espacios también surge esta doble situación. De ahí que el modelo propuesto responda a esta idea. Entonces, dado un conjunto de descripciones de objetos, el objetivo es encontrar una estructuración conceptual *natural* (agrupamientos y los conceptos asociados a los mismos) de estos objetos en el espacio de representación inicial. Esta estructuración conceptual debe ser lograda usando alguna medida de similaridad entre objetos y atendiendo a una cierta propiedad o criterio de agrupamiento Π para generar los agrupamientos. Los

agrupamientos así generados tendrán asociada una o más propiedades que caractericen a los objetos clasificados en los mismos. De ahí que deban seleccionarse conjuntos de variables apropiadas para caracterizar a los agrupamientos que conforman la estructuración.

De esta forma el nuevo modelo (ver Martínez-Trinidad y Ruiz-Shulcloper, 1996) está conformado en dos etapas, la primera denominada *estructuración extensional*, en donde los agrupamientos son creados sobre la base de la similaridad entre los mismos y un cierto criterio Π que indica cómo usar esta similaridad. Un resultado importante en esta dirección fue el estudio realizado sobre las relaciones de conjunto entre los agrupamientos generados por diferentes criterios de agrupamiento (ver Martínez et al., 2000b). Esto se resume en la figura 1, la jerarquía mostrada impone un orden entre los agrupamientos, según su generalidad, que puede resultar de utilidad en la práctica. Los niveles inferiores contienen categorías muy especificas (o incluso objetos aislados destacados) mientras que los niveles superiores están formados por un conjunto de estructuraciones más generales. Esta disposición de las estructuraciones puede proporcionar, para un mismo universo de objetos, diferentes visiones del mismo, consistentes en estructuraciones con diferentes niveles de abstracción.

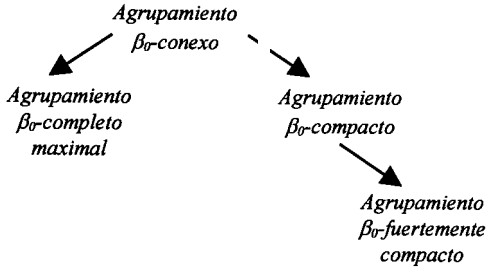

Figura 1. Grafo de inclusión entre los núcleos de los criterios agrupacionales

La segunda etapa del modelo es llamada *estructuración intencional*, donde se construyen los conceptos asociados a cada agrupamiento.

El nuevo modelo de agrupamiento conceptual propuesto (ver, Martínez-Trinidad and Ruiz-Shulcloper, 1997a), parte de un conjunto de descripciones de objetos a estructurar. Las funciones de similaridad entre variables y objetos deben ser definidas acorde a la forma en que el experto práctico lo hace. Posteriormente las estructuras o agrupamientos $K_1,...,K_c$ son encontrados mediante la aplicación de un criterio agrupacional. Los objetos son arreglados en forma matricial para seleccionar subconjuntos apropiados de características $t_1,...,t_r$, que serán útiles para caracterizar a los agrupamientos encontrados. Se propusieron dos alternativas, a saber, utilizando testores clásicos (ver Martínez-Trinidad and Ruiz-Shulcloper, 1999a) y testores por clase (Martínez-Trinidad and Ruiz-Shulcloper, 1999b). La construcción de las propiedades o conceptos asociados a cada agrupamiento K_i $i=1,...,c$. es en base a los conjuntos de características seleccionadas, la aplicación del operador REFUNION en t, que construye los conceptos (RU) y la aplicación de reglas de generalización (operador GEN, ver Martínez-Trinidad et al., 1999c) para simplificar los conceptos. Un algoritmo del nuevo modelo propuesto puede ser obtenido siguiendo el diagrama de flujo de la figura 2.

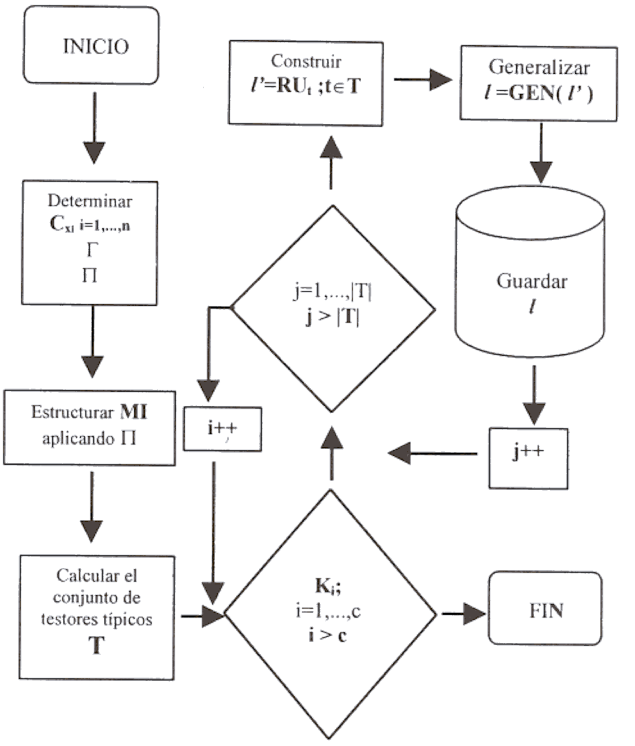

Figura 2. Diagrama de flujo para obtener un algoritmo conceptual duro del nuevo modelo

4 Modelo Difuso de Algoritmos de Agrupamiento Conceptual

En la práctica profesional de especialistas en ciencias poco formalizadas (soft-sciences), frecuentemente aparece el problema de la clasificación no supervisada conceptual en los que resulta de mucho interés saber no sólo qué objetos pertenecen a los agrupamientos, sino además en qué medida o grado los objetos pertenecen a los agrupamientos encontrados o en qué medida cumplen la propiedad de cada agrupamiento (ver, Zadeh, 1965). Dos objetivos principales en este sentido son resueltos con el modelo de estructuración difuso: a) obtener una estructuración extensional difusa –un conjunto de agrupamientos difusos; b) obtener una estructuración intencional, es decir, obtener una caracterización conceptual de los agrupamientos difusos en términos de conjuntos de características apropiadas.

En lo que concierne al primer objetivo, se obtiene una estructuración del espacio de representación de los objetos

(duros y difusos). De aquí que los dos modelos de agrupamiento conceptual antes mencionados quedan como un caso particular del modelo de objetos simbólicos aquí propuesto. Con la salvedad de que el nuevo planteamiento permite que se puedan abordar situaciones más complejas (objetos más complejos). Claro que deben ser definidos, para tal fin, los criterios de comparación apropiados para variables simbólicas más complejas.

El nuevo planteamiento de objetos simbólicos permite representar observaciones de objetos que el marco clásico no permite; así, en este nuevo planteamiento es posible representar objetos cuyas variables dependan de más de un universo de objetos, variables denominadas relacionales.

Los objetos no necesariamente están descritos por variables que toman un solo valor sino un conjunto de valores. Estos objetos dan la posibilidad de introducir en su definición, información más compleja como probabilidades, posibilidades y creencia. Además los objetos simbólicos permiten describir a los objetos de manera intencional, dando flexibilidad para expresar variación en los valores que toman las variables ([color=[rojo,blanco]]) y también se pueden expresar restricciones entre los valores de la variable ([peso≤350]).

Hablando de manera general podemos decir que existen cuatro tipos de análisis de datos dependiendo de la entrada y salida de los datos: a)análisis numérico de datos clásicos; b) análisis numérico de objetos simbólicos (por ejemplo, calculado propiedades estadísticas de las determinaciones extensionales de los objetos simbólicos); c) análisis simbólico de datos clásicos, por ejemplo el que se realiza con agrupamiento conceptual; y d)análisis simbólico de objetos simbólicos donde la entrada y la salida de los métodos son objetos simbólicos. La nueva formulación de objeto simbólico da pie a los 4 tipos de análisis antes mencionados, lo cual hace que el análisis clásico de datos quede como caso particular de la nueva Teoría de Objetos Simbólicos.

El objetivo a largo plazo es reducir el trecho entre el análisis de datos clásico (donde los objetos son vistos como puntos en un espacio n-dimensional) y la inteligencia artificial (donde el énfasis mayor es en la representación del conocimiento, razonamiento y aprendizaje). Con la definición de los objetos simbólicos, pretendemos que el análisis de datos se convierta en análisis de conocimiento.

6 Conclusiones y Trabajo Futuro

Se logró el objetivo de proponer herramientas para resolver el problema de la estructuración conceptual de espacios. Por herramientas debe entenderse: definiciones, teoremas, modelos y algoritmos.

Se introdujeron dos modelos de agrupamiento conceptual, el primero constituye una alternativa distinta para la solución de problemas de estructuración conceptual de espacios, donde los objetos de estudio están descritos por atributos cualitativos y cuantitativos simultáneamente, y donde pueden existir descripciones incompletas o ausencia de información de las variables.

El modelo difuso propuesto es el primero reportado en la literatura, la importancia de este modelo estriba en la originalidad y la trascendencia que esta línea de investigación tiene en la solución de problemas de Minería de Datos y Descubrimiento de Conocimiento.

Finalmente, el nuevo planteamiento de la definición de objetos simbólico en términos de variables simbólicas permite un marco de trabajo más amplio para abordar problemas de estructuración conceptual de espacios, en esa investigación se mostró cómo esta nueva formulación abre nuevas líneas de investigación como son: el análisis numérico de datos clásicos; el análisis numérico de objetos simbólicos; el análisis simbólico de datos clásicos; y el análisis simbólico de objetos simbólicos. De aquí podemos observar que el análisis clásico de datos queda como caso particular de la nueva Teoría de Objetos Simbólicos introducida en esta investigación.

Como trabajo futuro a corto plazo se pretenden desarrollar mejoras en la etapa intencional de los modelos de agrupamiento conceptual propuestos. También resulta de gran importancia desarrollar algoritmos conceptuales de tipo restringido, es decir, en donde a priori se pueda especificar el número de agrupamientos a formar.

Finalmente se pretende estudiar cada uno de los cuatro posibles análisis de datos que surgen a raíz de la nueva Teoría propuesta sobre OS.

Referencias

Béjar J. y Cortés U., "LINNEO+: Herramienta para la adquisición de conocimiento y generación de reglas de clasificación en dominios poco estructurados". En las memorias del *3er congreso Iberoamericano de Inteligencia Artificial*. La Habana Cuba, 1992, pp. 471-481.

Briscoe B. and Caelli T., *A compendium of Machine Learning Volume 1: Symbolic Machine Learning*. Ablex Publishing Corporation. Norwood, New Jersey. 1996.

Diday E., "Probabilist, possibilist and belief objects for knowledge analysis". *Annals of Operations Research* 55, 1995, pp. 227-276.

Feigenbaum, E. A., "The simulation of verbal learning behavior", in: E. A. Feigenbaum and J. Feldman (Eds), *Computers and Thought* McGraw-Hill, New York. 1963.

Fisher D., "Knowledge Acquisition Via Incremental Conceptual Clustering". *Readings in Machine Learning*, Shavlik and Dietterich, editors, 1990, pp. 267-283.

Gennari J. H., Langley and Fisher D., "Model of incremental Concept formation". In Jaime Carbonell, MIT/Elsevier *Machine Learning. Paradigms and Methods*, 1990, pp11-61.

Gowda K. C. and Diday E., "Unsupervised Learning through Symbolic Clustering". *Pattern Recognition Letters* 12, North-Holland, 1991, pp. 259-264.

Gowda K. C. and Diday E., "Symbolic Clustering Using a New Similarity Measure". *IEEE Transactions on Systems, Man, and Cybernetics*, Vol 22, No. 2, 1992a, pp. 368-378.

Gowda K. C. and Diday E., "Symbolic Clustering Using a New Dissimilarity Measure". *IEEE Transactions on Systems, Man, and Cybernetics*, Vol 22, No. 2, 1992b, pp. 567-578.

Gowda K. C. and Ravi T.V., "Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity". *Pattern Recognition Letters* 16, 1995a, pp. 647-652.

Gowda K. C. and Ravi T.V., "Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity". *Pattern Recognition*, Vol. 28, No. 8, 1995b, pp. 1277-1282.

Hanson S. J., "Conceptual clustering and categorization: bridging the gap between induction and causal models". In Y Kodratoff and R. S. Michalski, editors, *Machine Learning: An Artificial Intelligence Approach*, volume 3, 1990, pp.235-268. Morgan Kaufmann, Los altos, CA.

Lebowitz. M., "Categorizing numeric information for generalization", *Cognitive Sci.* 9. 1985, pp. 285-309.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., "Fuzzy semantic clustering", Proceedings of the *4th European Congress on Intelligent Techniques and Soft Computing* (Aachen, Germany), 1996, pp. 1397-1401.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., Un modelo de estructuración conceptual, II Taller Iberoamericano de Reconocimiento de Patrones. La Habana, Cuba, 1997a, pp 113-123.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., Fuzzy Conceptual Clustering, 5th European Congr. on Fuzzy and Intelligent Technologies Proceedings. Aachen, Alemania, 1997b, pp 1852-1857.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., Algoritmo LC-conceptual Duro. IV Simposio Iberoamericano de Reconocimiento de Patrones. La Habana, Cuba, 1999a, pp 195-206.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., LC-conceptual algorithm: Characterization using typical testors by class, 7th European Congr. on Fuzzy and Intelligent Technologies Proceedings (On CD). Aachen, Alemania,1999b.

Martínez-Trinidad J.Fco. and Ruiz-Shulcloper J., Pons-Porrata A., Algoritmo LC-Conceptual: Una mejora de la etapa Intencional Utilizando Reglas de Generalización. Taller de Inteligencia Artificial TAINA'99. México D.F. 1999c, pp.179-188.

Martínez-Trinidad J.Fco. Herramientas para la estructuración conceptual de espacios. Tesis en opción de grado de Doctor en Ciencias de la Computación, CIC,IPN, 2000a.

Martínez-Trinidad J.Fco., Ruiz-Shulcloper J., Lazo Cortés M., "Structuralization of universes". *Fuzzy Sets & Systems* 112/3, 2000b, pp 485-500.

Martínez-Trinidad J.Fco., A. Guzman-Arenas. The logical combinatorial approach to pattern recognition an overview through selected works, Pattern Recognition, 2001, 34/4 1-11.

McKusick K. and Thompson K., "Cobweb/3: A portable implementation". Technical report FIA-90-6-18-2, NASA Ames Research Center. 1990.

Michalski R. S., "Knowledge adquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjuntive concepts", (Special issue on knowledge acquisition and induction), *Policy Analysis and Information Systems*, No 3, 1980, pp. 219-244.

Michalski R.S., "Automated construction of classifications: conceptual clustering versus numerical taxonomy". *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. PAMI-5, No. 4, July. 1983.

Ralambondrainy H., "A conceptual version of the K-means algorithm". *Pattern Recognition Letters* volume 16, 1995, pp. 1147-1157.

Ruiz-Shulcloper J. and Montellano-Ballesteros J.J., "A new model of fuzzy clustering algorithms", Proceedings of the *3th European Congress on Fuzzy and Intelligent Technologies and Soft Computing* (Aachen, Germany), 1995, pp. 1484-1488.

Ruiz-Shulcloper J., Chac-Kantun M.G., and Martínez-Trinidad J.Fco. Bases conceptuales para una teoría de objetos simbólicos, Computación y Sistemas, Vol. 1 No.1,1997, pp.13-20.

Ruspini, E., "A new approach to clustering", *Information and Control* 15, 1969, pp. 22-32.

Zadeh L., "Fuzzy sets". *Inf. Control.* 8, 1965, pp. 338-353.