Computación y Sistemas Vol. 4 No.3 pp. 270 - 275
© 2001, CIC - IPN. *ISSN 1405-5546* Impreso en México

PH. D. THESIS ABSTRACT

Tools for the Conceptual Structuralization of Spaces

Graduated: 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

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

Abstract

This research proposes a set of tools (definitions, theorems, models and algorithms) to solve the problem of conceptually structuring or classifying a data sample.

Two models (hard and fuzzy) are introduced. These models permit the solving of the conceptual structuring of spaces in a theoretic frame where the attributes used to describe the objects under study can be of a quantitative or qualitative nature, or both natures and where the absence of information is permitted (missing values). The obtaining of the conceptual algorithms from the proposed models are described.

A new formulation is also proposed of the concept of the symbolic object in such form that the proposed conceptual models remain as a particular case of this new formulation. In fact this new proposition extends the theoretic frame in such a manner that the tools of the classical data analysis also remain as a particular case of the new formulation.

In a general form, the obtained results can be appreciated since they do not exhaust the development of this thematic, but the work already made opens new researches, above all, at the present time it is of great importance to solve intelligent data analysis problems.

1 Introduction

In many of the applied sciences the problem of revealing the underlying structure in a collection of objects (situations, measurements, observations, phenomena, etc.) is present. This object information is typically stored in plain files, data bases or some other electronic means in a structured form. The problem with classifying or structuring this information has been studied intensively in the area of unsupervised Pattern Recognition ("cluster analysis"). The methods developed in this area, form clusters on the base of very similar object pairs and ignore the usefulness of the meaning of the obtained clusters, this last one being a more natural element for the final users.

All the traditional techniques of unsupervised Pattern Recognition have the disadvantage of forming clusters which do not have a conceptual interpretation. The problem with the meaning of the obtained clusters is left up to the specialist. This disadvantage is significant since the specialist, for purposes of investigation or work, not only requires the clusters, but in addition requires an explanation of them in humans terms.

The conceptual clustering arises from Michalski's work (1980), this approach proposes to find from a data set, not only the clusters in which these are structured but also the composed explanation of such clusters. The conceptual clustering is composed of two fundamental tasks: the cluster of object where useful subsets of an object sample are determined, and the characterization which determines concepts for each subset discovered.

In this form, we can consider that a *conceptual structuralization of spaces* problem consists of finding, in a general form, a coverage of a object set under study, as well as the properties or concepts associated to the found clusters or structures.

In the last two decades, different conceptual clustering algorithms have been proposed (see Briscoe and Caelli, 1996), within the most representative we can mention: 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), CONCEPTUAL K-MEANS (Ralambondrainy, 1995). In this doctoral research a critical study was realized from the above, where the practical findings and limitations are shown (see, Martínez-Trinidad, 2000a).

This work is organized in the following manner: As a first step we formally propose the conceptual structuralization of spaces problem. Later, a new model of conceptual clustering algorithms is presented with their generalization in the fuzzy case. The conceptual structuralization problem is also approached from the formalism of the Theory of Symbolic Objects. Finally, the work that will be developed in the future is enumerated.

2 Formal Exposition of the Problem

The conceptual structuralization of spaces problem (see; Martínez-Trinidad, 2000b y Martínez-Trinidad and Guzmán-Arenas, 2001) is formulated as follows. Let *M* be a set of objects. A *description* I(O) is defined for each object $O \in M$ and this is represented by a finite sequence $x_1(O), x_2(O)..., x_m(O)$ of values of *m* attributes of the set $R = \{x_1, x_2..., x_m\}$, with $x_i(O) \in M_i$, M_i being the set of admissible values of the attribute x_i. In addition, we will assume that in M_i there is a symbol * which denotes the *information absence, i=1,...,m*. In other words, the object description can be incomplete. This is, for at least one attribute the value is not known. Let us consider that $I(O) \in M_1 \times M_2 \times, ..., \times M_m$ (this Cartesian product is the initial representation space IRS of the objects). The nature of the features or attributes by consequence can be, simultaneously, any (qualitative: Boolean, multi-valued, fuzzy, linguistic and other; or quantitative: whole, real).

Meanwhile, we will not assume any algebraic or topological structure for the IRS. Consider the function C_i $: M_i \times M_i \rightarrow L_i$ such that

a) $C_i(x_i(O), x_i(O)) = \min_{O' \in M} \{C_i(x_i(O), x_i(O'))\}$ if C_i is a comparison criterion of dissimilarity between attribute values x_i or

b) $C_i(x_i(O), x_i(O)) = \max_{O' \in M} \{C_i(x_i(O), x_i(O'))\}$ if C_i is a comparison criterion of similarity between attribute values x_i for *i=1,...,m*.

C_i is an evaluation of the degree of similarity (or dissimilarity) between any two values of the attribute x_i where L_i is a totally ordered set, *i=l,...,m*.

Let $\Gamma: (M_1 \times ... \times M_m)^2 \rightarrow$ L be a function, where L is a totally ordered set; Γ will be denominated a *similarity function* and is an evaluation of the degree of similarity between two descriptions, any pertaining to MI.

Usually, the object information (their descriptions) is given in form of a table or matrix $MI = |x_i(O_j)|_{n \times m}$ with *n* rows (object descriptions) and *m* columns (values of each attribute in the selected objects).

The *conceptual structuralization of spaces problem* over *M* consists in determining the covered set $\{K_1, K_2, ..., K_C\}$ *c>l*, as well as the set of associated concepts for each K_i *i=1,...,c*. In principle K_i could be hard or fuzzy subsets of *M* and these could be disjoint or not.

3 Hard Model of Conceptual Clustering

From the previous formulation, a new model of conceptual clustering algorithms was introduced for the hard case, this means, when the found clusters are hard subsets of the space to be structured (in the sense of the classical theory of sets).

In all problems related to the Theory of Sets, the sets can be determined in an extensional or intentional form. In the problem of conceptual structure of spaces this double situation also arises. From there the proposed model responds to this idea. Then, given a set of object descriptions, the goal is to find a *natural* conceptual structuralization (clusters and concepts) of these objects in the initial representation space. This conceptual structure must be reached using some similarity measurement between objects and attending a certain property or clustering criterion II to generate the clusters. The clusters generated will have associated one or more properties that characterize the classified objects in them. From there, sets must be selected from appropriate attributes to characterize the clusters that compose the structure.

In this manner, the new model (see Martínez-Trinidad and Ruiz-Shulcloper, 1996) is composed of two stages, the first is denominated *extensional structuralization*, where the clusters are created over the base of similarity between the same ones and a certain criteria II that indicates how to use this similarity. An important result in this direction was the study realized on the relations between the clusters generated by different clustering criteria (see, Martínez et al.. 2000b).

This is summarized in figure 1, the shown hierarchy imposes an order between clusters, according to their generality, it can result useful in practice. The inferior levels contain very specific categories (or even outstanding isolated objects) while the superior levels are formed by a more general structure set. This structure disposition can supply, for one same object universe, different visions of the same, consistent in structures with different abstraction levels.

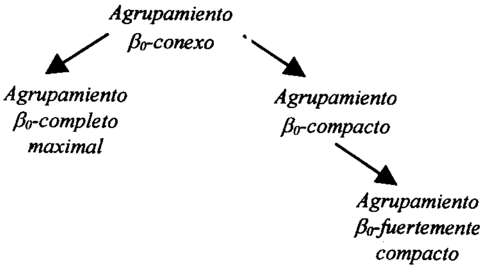
Agrupamiento β_0-conexo
Agrupamiento β_0-completo maximal
Agrupamiento β_0-compacto
Agrupamiento β_0-fuertemente compacto

Figure 1. Graph of the inclusion between the clustering criterion nucleuses

The second stage of the model is called intentional structuralization, where the associated concepts to each cluster are built.

The new conceptual clustering model proposed (see, Martínez-Trinidad and Ruiz-Shulcloper, 1997a) start with a set of object descriptions to structure. The similarity functions between attributes and objects must be defined according to the form in which the practical expert does it. Following, the structures or clusters $K_1,...,K_c$ are found through the application of a clustering criterion. The objects are fixed in a matrix form to select appropriate subsets of attributes $t_1,...,t_r$, that will be useful to characterize the found clusters. Two alternatives where proposed, using classical typical testors (see Martínez-Trinidad and Ruiz-Shulcloper, 1999a) and typical testors by class (Martínez-Trinidad and Ruiz-Shulcloper, 1999b). The construction of the associated properties or concepts to each cluster K_i $i=1,...,c$ is based on the sets of selected attributes, the application of the operator REFUNION on t, that constructs the concepts (RU) and the generalization rules application (GEN operator, see Martínez-Trinidad et al., 1999c) to simplify the concepts.

An algorithm of the new proposed model can be obtained following the flow diagram from figure 2.

4 Fuzzy Model of Conceptual Clustering

In the professional practice of the soft sciences, specialists frequently encounter the problem of conceptual unsupervised classification in what results of much interest to know not only what objects pertain to the clusters, but also to what measurement or degree the objects pertain to the clusters or in what measurement they satisfy the property of each cluster (see, Zadeh, 1965). Two principal objectives in this sense are solved with the fuzzy structure model: a) obtain a fuzzy extensional structuralization of a set of fuzzy clusters; b) obtain an intentional structure, that is to say, obtain a conceptual characterization of the fuzzy clusters in terms of appropriate attribute sets.

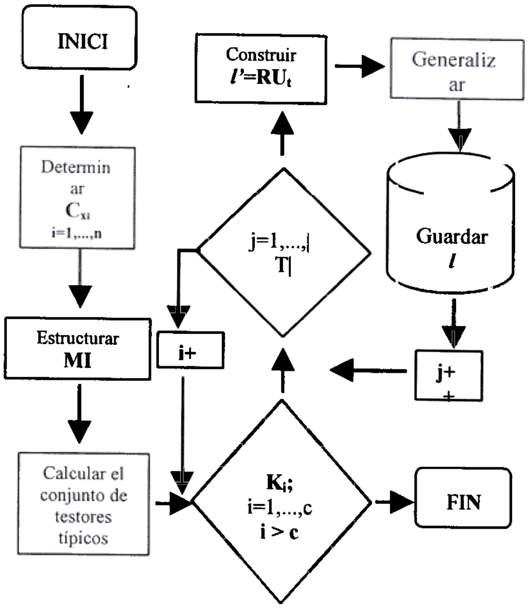

Figure 2: Flow diagram to obtain a hard conceptual algorithm of the new model

Concerning the first objective, a structuralization of the object representation space is obtained in Ruspini fuzzy partitions (Ruspini, 1969), fuzzy β-partitions, and fuzzy β-overings (Ruiz-Shulcloper and Montellano-Ballesteros, 1995). In addition, mixed information was considered, that is, quantitative and qualitative data, information absence, and the analogy criteria to compare the values of the attributes and descriptions of the objects (subdescriptions) are not necessarily distances. In the second objective, the concept or concepts that characterize each of the obtained clusters are built in the fuzzy extensional structuralization. This stage consists of selecting sets of attributes (appropriate sets) and building the properties that characterize the objects of each cluster.

In this manner, the new model is composed of two stages, the first one is denominated as an extensional structuralization, where the clusters are created over the base of similarity between them and a certain fuzzy clustering criterion Π_d, that indicates how to use this similarity. In this case in an analog manner to the hard model, the relations of the set between the generated clusters where studied for different fuzzy clustering criterion, (see Martínez-Trinidad et al., 2000). Figure 1 shows these relations.

The second stage of the model is called intentional structuralization, and is where the associated concepts to each fuzzy cluster are built.

The new fuzzy conceptual clustering model (see, Martínez-Trinidad and Ruiz-Shulcloper, 1997b) starts with a set of object descriptions to structure. The similarity functions between attributes and objects must be defined in the same manner as in the hard model. Following, the fuzzy structures or clusters $\tilde{K}_1,\dots,\tilde{K}_c$ are found through the application of some fuzzy clustering criterion Π_d, and the objects are fixed in a matrix form to select appropriate subsets of attributes that will be useful to characterize the found clusters. In this case, the *discriminant attribute* concept was introduced to select the attributes under a fuzzy context (see, Martínez-Trinidad and Ruiz-Shulcloper, 1999b). The construction of the associated properties or concepts to each cluster $\tilde{K}_i$ $i=1,\dots,c$ is made in base to the discriminant attributes sets $t_1,\dots,t_r$, and the application of the REFUNION operator on t (RU), that builds the concepts, and the application of the generalization rules (GEN operator), to simplify the concepts. An algorithm of the new proposed model can be obtained following the flow diagram from figure 3.

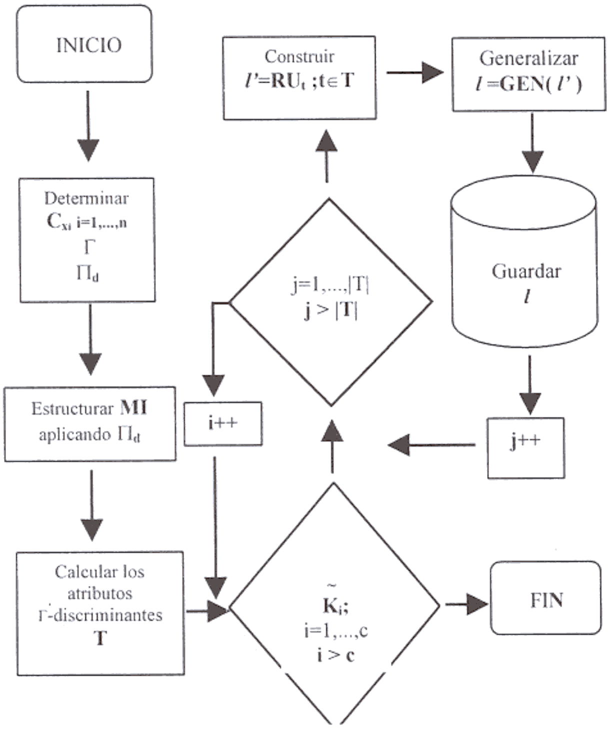

Figura 3. Flow diagram to obtain a fuzzy conceptual algorithm of the new model

5 Symbolic Objects

The concept of Symbolic Objects (SO) has its origin in the developed works on conceptual clustering by Michalski and Diday in the early 80's. Thereafter, Diday and his group developed this concept as well as a set of tools for the analysis of symbolic data (see Diday, 1995; Gowda and Diday, 1991; Gowda and Diday, 1992a; Gowda and Diday, 1992b; Gowda and Ravi, 1995a; and Gowda and Ravi, 1995b).

This research studied in a critical manner the Diday SO model, as a consequence a new SO model was proposed (see Ruiz-Shulcloper et al., 1997). For the new model definition, the symbolic attribute concept was introduced. This concept covers three kinds of attributes: heterogeneous relational attributes, homogeneous relational attributes and attributes of type set. This last type covers the classical attributes that are used in the analysis of data.

On the basis of the symbolic attribute concept, a new SO definition in terms of a Cartesian description is introduced, which can have associated an extensional determination and an intentional determination. According to the extensional determination, the SO can be hard, fuzzy or L-fuzzy. According to the intentional determination (and the selected calculus) the SO can be Boolean, K-valent, fuzzy, of belief, probabilistic, possibilistic, models, etc.

As part of the research, different comparison criteria between symbolic attributes where introduced. The concept of the partial and total similarity function between symbolic objects was also introduced (see, Martínez-Trinidad, 2000a).

Once these concepts are defined, all the resting dependent concepts of these are left analogically defined. This is the case for the clustering criteria (hard and fuzzy). From here the two conceptual clustering models mentioned before are left as a particular case of the symbolic object model here proposed. With the exception that the new proposal permits the approach of more complex situations (more complex objects). Of course, they must be defined, for that purpose, the appropriate comparison criteria for more complex symbolic attributes.

The new proposal of symbolic objects permits the representation of object observations that the classical frame does not permit; in this manner, in this new proposal it is possible to represent objects whose attributes depend more of a universe of objects, attributes denominated relational. The objects are not necessarily described by attributes that take only one value but a set of values. These objects give the possibility of introducing in its definition, more complex information like probabilities, possibilities and beliefs. In addition, the symbolic objects permit the object description in an intentional manner, giving flexibility to express variation in the values that the attributes take ([color=[red,white]]) and also expressing restrictions between the values of the attribute ([weight ≤350]).

Speaking in general terms we can say that there exist four types of data analysis depending on the data input and output: a)numerical analysis of classical data; b) numerical analysis of symbolic objects (for example, calculation of statistical properties in the extensional determinations of the symbolic objects); c) symbolic analysis of classical data, for example, the one that is made with conceptual clustering; and d) symbolic analysis of symbolic object where the input and output of the methods are symbolic objects. The new formulation of the symbolic object gives place to the 4 types of analysis mentioned before, which makes the classical data analysis remain as a particular case of the new Theory of Symbolic Objects.

The long term objective is to reduce a stretch between the classical data analysis (where the objects are seen as points in an n-dimensional space) and the artificial intelligence (where the major emphasis is in the knowledge, reasoning and learning representation). With the definition of the symbolic objects, we pretend that the data analysis be converted into knowledge analysis.

6 Conclusions and Future Work

The objective of proposing tools to solve the *conceptual structuralization of spaces* problem was met. By tools we mean: definitions, theorems, models and algorithms.

Two models of conceptual clustering were introduced, the first constitutes an alternative for the solution of conceptual clustering problems, where the objects of study are described by qualitative and quantitative attributes simultaneously, and where incomplete descriptions or information absence of the attributes can exist.

The proposed fuzzy model is the first reported in literature, the importance of this model is based in the originality and the transcendency that this line of research has in the problem solution of Data Mining and Knowledge Discovery.

Finally, the new proposal of the definition of the symbolic objects in terms of symbolic attributes permits a wider work frame to approach space conceptual structuring problems, in that research it was shown how this new formulation opens new lines of research like: numerical analysis of classic data; the numerical analysis of symbolic objects; the symbolic analysis of classical data; and the symbolic analysis of symbolic objects. We can observe from here that the classical data analysis remains as a particular case of the new Theory of Symbolic Objects introduced in this research.

As a further short term project it is intended to develop improvements in the intentional stage of the proposed conceptual clustering models. It is also of great importance to develop conceptual algorithms of restricted type, that is to say, where a priori you can specify the number of clusters to form.

Finally, it is intended to study each one of the four possible analysis of data that arise from the new proposed SO Theory.

References

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.