The system FUZZCLASS for classification of multivariate observations

Slávka Bodjanová, Bratislava, Czechoslovakia

The goal of clustering technique is to classify a given set of data points by assigning them to a reasonably small number of natural subsets (clusters) such that all data points contained in the same cluster are more similar to each other than that they are to data points in other clusters.

The intent of this paper is to demonstrate that some of well-known methods based on fuzzy sets can be connected in practical system for classification of multivariate observations. The connection is shown in Fig. 1.

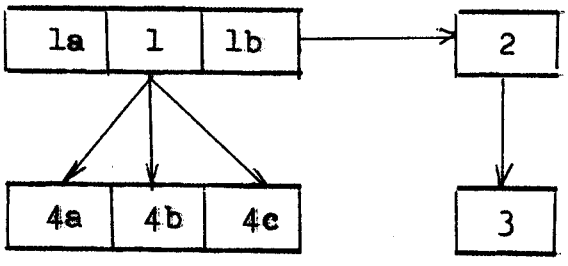

Fig.1

Comments:

1) Partitioning of a given data set X of n objects due to fuzzy k-means criterion [3] .
The number k of clusters has been decided apriori by the investigator. The result of partitioning can be considered as a matrix $W = [w_{ij}]$, $i=1,\dots,k$; $j=1,\dots,n$; where $w_{ij} \in \langle 0, 1\rangle$ for each i,j, and $\sum_i w_{ij} = 1$. The amount of fuzziness we can measure e.g. by the classification entropy of W defined as follows:

$$H(W) = \frac{1}{n} \sum_i \sum_j h(w_{ij}),$$

where $h(w_{ij}) = w_{ij} \cdot \log_k w_{ij}$ for $w_{ij} > 0$ and $h(0) = 0$.

1a) Classification (based upon the fuzzy partition W) of some new objects. The type of membership assignment is derived from fuzzy k-means algorithm.

1b) Measuring of similarity between objects of X (based upon the partition W) by a relative and symmetric fuzzy relation μ.

$$\mu: X \times X \to \langle 0,1\rangle : \mu(x_r,x_s) = 1 - \sqrt{\sum_i (w_{ir} - w_{is})^2/2} \quad \text{for each } x_r, x_s.$$

2) Construction of max-min transitive closure $\hat{\mu}$ of the fuzzy relation μ.

3) Generating a hierarchical clustering scheme of hard partitions of X (see [1]) based upon $\hat{\mu}$.
For a given decreasing sequence of hierarchical levels $\{\varepsilon_i\}_{i=1}^{s}$, $\varepsilon_i \in (0,1)$ we can obtain a sequence of hard partitions $\{U_{\varepsilon_i}\}$ according to the rule
$(x_r, x_s) \in U_{\varepsilon_i}$ iff $\hat{\mu}(x_r, x_s) \geq \varepsilon_i$ for each $x_r, x_s \in X$.

4) Approximation of fuzzy partition W.
There are a few methods available:

4a) α-membership method:
For a given parameter $\alpha \in (0,1)$ we can approximate fuzzy partition W by a hard partition Z such that
for each i,j: $z_{ij} = 1$ if $w_{ij} = \max_r \{w_{rj}\} \geq \alpha$,

else $z_{ij} = 0$.

If for some j: $\sum_i z_{ij} = 0$ then Z is called a degenerate hard partition.

4b) Sharpened membership method:
For a given parameter $\delta \in (0,1)$ we can approximate fuzzy partition W by a "sharpened" (less fuzzy) partition T as follows:
for each j: if $\max_i \{w_{ij}\} - \min_i \{w_{ij}\} \leq \delta$ then $t_{ij} = w_{ij}$,

else $t_{ij} = 0$ if $w_{ij} < \frac{1}{k}$

and $t_{ij} = w_{ij} + \sum_{w_{sj} < \frac{1}{k}} w_{sj} / (k - r_j)$ if $w_{ij} \geq \frac{1}{k}$

where $r_j = \text{card}\{w_{ij};\ w_{ij} < \frac{1}{k}\}$.

4c) Following Bezdek and Harris [4] a convex decomposition of W can be obtained;

$$W = \sum_{m=1}^{n(k-1)} c_m \cdot U_m \quad , \quad c_m > 0, \ \sum_m c_m = 1,$$

where U_m is a hard partition and c_m is its weight.
The similarity of original partition W and its approximation U can be evaluated by

$$\mathcal{L}(W,U) = 1 - \frac{1}{2n} \sum_i \sum_j (w_{ij} - u_{ij})^2$$

The system FUZZCLASS was successfuly applied in practical economic analysis.

A simple example may be helpful:

1) Let us consider the data set of Table 1 consisting of the values of six measurements for each of the objects.

1	14,390	0,168	13,091	16,423	2,854	4,956
2	25,608	0,294	16,104	17,538	4,948	7,861
3	22,912	0,035	22,874	22,220	2,355	8,005
4	27,219	25,868	24,820	24,560	2,633	7,155
5	23,231	2,557	20,873	23,292	3,118	5,155
6	23,465	11,053	19,010	25,988	1,331	6,061
7	34,059	0,000	33,952	32,393	24,274	10,700
8	16,734	0,544	12,305	18,597	-0,361	5,774
9	58,014	0,092	65,427	50,920	28,536	9,427
10	11,871	4,307	20,582	19,763	-0,565	5,384

Table 1

The initial fuzzy partition $U^{(0)}$ is given:

$$U^{(0)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0,5 & 0 & 1 & 0 \\ 0 & 0 & 0,5 & 0,7 & 0 & 0,6 & 0,5 & 0 & 0 & 0 \\ 0,6 & 0,5 & 0,3 & 0,3 & 1 & 0,4 & 0 & 0,4 & 0 & 0,2 \\ 0,4 & 0,5 & 0,2 & 0 & 0 & 0 & 0 & 0,6 & 0 & 0,8 \end{pmatrix}$$

(Also, other initial partitions have to be tried in order to examine experimentally the uniqueness of the final result.)

After some iterations we obtain by fuzzy k-means technique the final partition

$$W = \begin{pmatrix} 0,0019 & 0,0076 & 0,0023 & 0,0019 & 0,0004 & 0,0112 & 0,1609 & 0,0017 & 0,9992 & 0,0073 \\ 0,0134 & 0,0572 & 0,0175 & 0,9739 & 0,0042 & 0,2526 & 0,2567 & 0,0132 & 0,0003 & 0,0064 \\ 0,0589 & 0,6102 & 0,9063 & 0,0144 & 0,9793 & 0,5037 & 0,3776 & 0,0643 & 0,0003 & 0,2318 \\ 0,9258 & 0,3250 & 0,0739 & 0,0098 & 0,0161 & 0,2325 & 0,2048 & 0,9208 & 0,0002 & 0,7545 \end{pmatrix}$$

The amount of fuzziness evaluated by classification entropy is $H(W) = 0,33567$. That means the 4-collection of fuzzy clusters is reasonably separable and the partition W reflects the real data structure reasonably well.

1a) For new object $x_{11} = (28,719 \quad 0,067 \quad 25,882 \quad 24,817 \quad 0,109 \quad 5,615)$ the following vector of memberships is obtained:
$(0,0139 \quad 0,0949 \quad 0,7313 \quad 0,1599)$.

4a) For $\alpha = 0,5$ we obtain the following hard approximation of W:

$$W_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$\mathcal{L}(W,W_1) = 0,94742.$

We can see that object x_7 is not typical for any of 4 given clusters at the level $\alpha = 0,5$.

4b) For $\delta = 0,2$ we obtain the following sharpened approximation of W:

$$W_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0,37445 & 0,43955 & 0 & 0 & 0 \\ 0 & 0,6426 & 1 & 0 & 1 & 0,62555 & 0,56045 & 0 & 0 & 0 \\ 1 & 0,3574 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$\mathcal{H}(W_2) = 0,1442$ and $\mathcal{L}(W,W_2) = 0,981296$.

4c) The convex decomposition of W yields many hard partitions but only some of them have the significant weight. We present only U_1 with the weight $c_1 = 0,3776$ and U_2 with the weight $c_2 = 0,253$.

$$U_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$$U_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

1b) We calculate the matrix of inter-objects similarity for the subset $\{x_1, x_2, x_3, x_4, x_6, x_9, x_{10}\}$.

$$\mu = \begin{pmatrix} 1 & 0,423 & 0,151 & 0,061 & 0,393 & 0,037 & 0,794 \\ & 1 & 0,724 & 0,195 & 0,830 & 0,145 & 0,625 \\ & & 1 & 0,074 & 0,653 & 0,046 & 0,350 \\ & & & 1 & 0,363 & 0,014 & 0,179 \\ & & & & 1 & 0,178 & 0,597 \\ & & & & & 1 & 0,126 \\ & & & & & & 1 \end{pmatrix}$$

2) The transitive closure of μ is derived.

$$\hat{\mu} = \begin{pmatrix} 1 & 0,625 & 0,625 & 0,363 & 0,625 & 0,178 & 0,794 \\ & 1 & 0,724 & 0,363 & 0,830 & 0,178 & 0,625 \\ & & 1 & 0,363 & 0,724 & 0,178 & 0,625 \\ & & & 1 & 0,363 & 0,178 & 0,625 \\ & & & & 1 & 0,178 & 0,625 \\ & & & & & 1 & 0,178 \\ & & & & & & 1 \end{pmatrix}$$

We obtain the following sequence of hard partitions:

i	ε_i	U_{ε_i}
1	0,830	$\{x_2,x_6\}\ \{x_3\}\ \{x_1\}\ \{x_{10}\}\ \{x_4\}\ \{x_9\}$
2	0,794	$\{x_2,x_6\}\ \{x_1,x_{10}\}\ \{x_3\}\ \{x_4\}\ \{x_9\}$
3	0,724	$\{x_2,x_6,x_3\}\ \{x_1,x_{10}\}\ \{x_4\}\ \{x_9\}$
4	0,625	$\{x_2,x_6,x_3,x_1,x_{10}\}\ \{x_4\}\ \{x_9\}$
5	0,363	$\{x_2,x_6,x_3,x_1,x_{10},x_4\}\ \{x_9\}$
6	0,178	$\{x_2,x_6,x_3,x_1,x_{10},x_4,x_9\}$

Table 2

The hierarchy of hard clusters derived from $\hat{\mu}$ is in Fig.2.

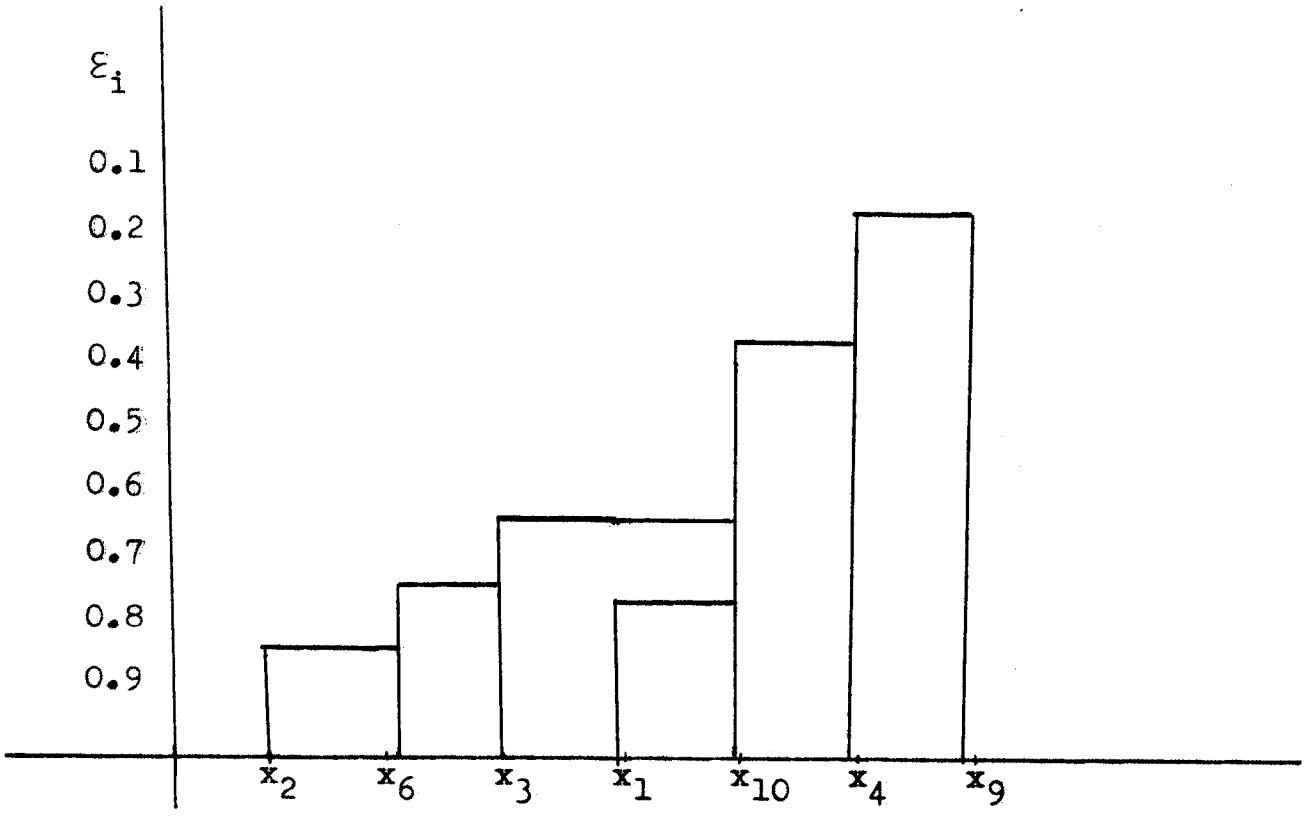

Fig. 2

References

[1] Backer,E.: Cluster analysis by optimal decomposition of induced fuzzy sets, Delftse Universitaire Pres, Delft 1978

[2] Backer,E.: Cluster analysis formalized as a process of fuzzy identification based on fuzzy relations, Report No.IT-78-15,Delft.Univ. of Technology, Delft 1978

[3] Bezdek,J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York 1981

[4] Bezdek, J.C., Harris,J.D.: Convex Decomposition of Fuzzy Partitions, Journal of Mathem.Analysis and Applications 67, /1979/, 490-512

Language - Defined Selection of Data

RNDr. Jana Talašová, Prerov Machinery

1. The basic idea

The present paper deals with a proposal of user-oriented program realization of access to database. The access is based on application of the essential means of the fuzzy-set theory to the computer representation of data selection criteria formulated in a natural language. The proposed approach could be utilized e.g. for providing information of any management subsystem state from the central company database according to the individual requirements of managers.

The paper aims at creating such a program communication with the database which would be as natural as possible for the user, and which would ensure correspondence between his requirement and the computer's response. At the same time, the proposed solution emphasizes simplicity of such a program.

If the user does not only require particular data items from the database, but on the contrary, information about the state of a certain set of objects, his requirement is not at first in the form of a set of mathematical-logical conditions (as most of the database communication systems require), but it is essentially in the form of an uncertain criterion in a natural language. For the user, his natural language is the most effective means of communication; its use in communicating with the database requires, however, that the user must first "teach" the computer his way of expressing himself.

The proposed solution is based on the idea that each user approaches the data stored in the database on the basis of a small number of relatively stable viewpoints of which each can be expressed by a data fuzzy filter (DFF), i.e. by a system of fuzzy sets and operations defined on them which represents the user's verbal requirements. It is typical for DFF's that

they have varying transmittance which reflects partial variability of the user's requirements which can necessitate a more accurate selection criterion or a change in the required detailness of data. The form of DFF is created according to the concrete requirements of the user only when he starts selecting data.

2. Realization of the solution

2.1. Definition of data fuzzy filters

The user can create DFF either with the help of a systems analyst or of a special interactive program. In both cases, the following procedure can be applied:

a) DFF is identified, i.e. its naming unit is chosen and correspondence is found between the aim of the query and the possibilities of the database.

b) The type of observed objects is delimitated (e.g. "Erections").

c) The importance is defined of information items related to the observed objects. Their grade of membership in the fuzzy set "Importance" is verbalized; the user relates the item names to qualities such as "absolutely necessary information" (Grade 1), ... "unimportant information" (Grade 0.1). Unrelated items are considered to be of zero grade of membership.

d) Simple language criteria are defined for the selected information items. These criteria are represented by means of fuzzy sets. E.g., for item "Scheduled dead-line of erection", it is possible to define a fuzzy criterion "Erection to be completed soon".

e) From simple fuzzy and interval or point-defined criteria, aggregated criteria are created with the help of logical words "and", "or", "either-or", "no". The process of aggregation is modelled by operations with fuzzy sets.

These criteria are given suitable naming units. E.g., criterion "Important erections with threatened dead-line" can represent "Erections with high costs", "rather big", those which "should be completed soon" and in which "a substantial part of costs remains to be spent".

The creation of DFF represents the most difficult part of the user's job. It is, however, more or less a single action whose realization makes routine communication with the database substantially easier.

2. 2. The selection of data

The definition of DFF contains the user's global view of the particular set of objects. The filtration of data itself is realized by means of a general program which activates the required DFF on the one hand, and on the other it serves to accurate delimitation of its transmittance. The program is run in the following way:

a) After the user logs in, he is presented a menu of DFF's defined by him, and the selected filter is activated.

b) From the menu of the aggregated criteria of the appropriate DFF, the user selects the required criterion. He can make it more accurate with the help of fuzzy criteria of a lower lever or by means of the interval or point condition for some information item.

c) Then the user defines the selection sensitivity limit, i.e. he determins the minimum degree of satisfying the complete selection criterion; when this is achieved, the object in question is selected. Also this limit can be defined using the language scale. E.g. "Prevailing fulfilling of the complete criterion" can be required.

d) In the end it is necessary to determine the degree of detail of information about the selected objects. In accordance with the importance of items defined

in the given DFF it is possible to require "only absolutely necessary information" or "also unimportant information". At the same time, it is possible to require any particular item of non-zero importance.

If the selection process is to satisfy the user as much as possible, it is necessary to find an optimum proportion between the number of aggregated criteria (the user should not lose his accurate idea of their contents) and the necessity of making the selection criteria more accurate. It is also necessary to maintain a reasonable level of detail when creating the language scales.

3. Advantages of applying DFF's to selection algorithms

A natural language is, from the user's viewpoint, an ideal means of communication with the database; from the viewpoint of the computer, however, the direct comprehension of the text represents a very difficult problem. Data selection based on the application of DFF's enables the computer to understand a verbalized query if it has once been explained to it. Thus it maintains certain advantages of language communication, with minimum claims to both the programmer and user.

On the bazis of experience, DFF can be constantly improved, perfected, and adapted to altered requirements. In fact, it represents a certain basis of knowledge of the particular user applied to evaluating information about certain objects. This can be made effective use of at the absence of the particular person or when replacing him by someone else.

DFF's need not be only created for individual managers. They can also serve to whole groups of workers of similar profession, and the individual different requirements can be applied when using the DFF itself.

References:

[1] Novák,V.: Fuzzy množiny a jejich aplikace (Fuzzy sets and their applications). SNTL, Praha, 1986