M. STOICA
I. M. STANCU-MINASIAN
E. SCARLAT

Department of Economic Cybernetics
Academy of Economic Studies

On Large-Scale Classification Problems Using Fuzzy Sets

The authors make use of concepts concerning the fuzzy sets when studying large-scale classification problems. The similarity and non-similarity concept between two elements is defined by considering the degree of their membership towards a number of known properties. Four types of classification problems are examined and for two of them the authors describe solution methods in the hypothesis of a small (up to hundreds) number of elements, as well as in the hypothesis of a large (thousands or tens of thousands) number of elements. For large dimensions the Monte-Carlo Method is suggested for use and a criterion on the achieved precision is given.

The quick growth of the information amount as a result of manysided development of modern society requires the use of some methods meant to systematize all data contributing to a correct and ready decision-making. In this context, *information classification* is one of the stages which precede decision-making. Many works dwelling on the matter have been published these latest years and a new branch known by the name of mathematic taxonomy has been developing.

Up to the beginning of this century classification problems could be solved through intuition alone, whereas at present there are a great many cases in which this is no more possible without an algorithm to enable the computer use. For this purpose it is necessary to define set M of elements to be classified, set P of properties (criteria) in relation to which classification is made, as well as the membership degree of each element towards each property. Depending on the nature of the variables by means of which the membership degree of an element (object) towards a property (criterion) is defined, there are two groups of classification proceedings:

— *deterministic proceedings*: the membership of an element i towards a property j is described by means of a boolean variable (0 or 1);

— *fuzzy proceedings*: the membership of an element i towards a property j is described by means of a membership degree $\mu_j(i)$ assuming values within the normalized interval [0, 1].

In case there are n properties, each element i can be characterized by means of a membership vector:

$$\mu_i = (\mu_1(i),\ \mu_2(i),\ ...,\ \mu_n(i)). \tag{1}$$

With the aid of these vectors we can compute the membership degree of an element i towards a class of elements C_j. We assume that both for i and for each element of class C_j the membership vectors according to properties $j(j = 1, 2, ..., n)$ can be computed.

Most computer programmed algorithms use distances between two vectors V_i and V_j. Obviously, if there are n vectors a scale matrix can be built up:

$$D = \| d_{ij} \| = \| d(V_i, V_j) \|,\ i, j = 1, ..., n. \tag{2}$$

Taking into account that these matrices are symmetrical and the distance between an element and itself is 0, it follows that the maximal number of distinct elements of the scale matrix is:

$$N = \frac{n(n-1)}{2}. \tag{3}$$

For instance, if there are $n = 1{,}000$ types of assortments of materials and raw materials to be classified in an industrial unit, then the scale matrix shall have 499,500 elements. It is obvious that there arise considerable computing difficulties for such scales, difficulties related to the time of program running and to the electronic computer store as well. If we take into account that in certain industrial units, for instance factories of electronic apparata, the number of assortments of materials, raw materials and semiproducts can be of hundreds of thousands, it follows that the concern for solving large-scale classification problems is quite seasonable.

The Monte Carlo method is used in order to avoid printing all the elements of the scale matrix; that is to say that only matrix samples are computed. It stands to reason that in this way the best classification solution is no longer obtained, but the method is assumed to be pretty good for practical purposes, because it offers the advantage of a definitely smaller amount of computation.

MATHEMATICAL STATEMENT OF CLASSIFICATION PROBLEMS

In a classification problem, a set m of elements $M = \{x_1, x_2, ..., x_m\}$ is to be divided into p disjoint classes $C_1, C_2, ..., C_p$ $(p \leqslant m)$ in relation to several properties $A_1, ...A_2, ..., A_n$, so that the similarity (from the standpoint

of the n properties) between the elements of the same class shall be maximal. In other words, we must determine $\{C_j\}$ so that

$$\text{a) } C_i \cap C_j = \emptyset,\ i \neq j;\ i, j = 1,2, ..., p$$

$$\text{b) } \bigcup_{i=1}^{p} C_i = M,\ i = 1,\ 2, ..., p. \tag{4}$$

This manner of approaching the problem already uses a fuzzy term, which is "similarity" between the elements of the same class from the standpoint of n properties. Besides, if the membership of an element towards a certain property is expressed by means of a fuzzy variable, this includes also the case in which there is a boolean variable. Therefore, we shall further describe the classification problem by making use of fuzzy concepts.

Consider A_i to be a property and x_j an element. We write $\mu_{A_i}(x_j) = \mu_{ji}$ for the membership degree of element x_j towards property A_i. Thus, we get a matrix of the membership degrees:

$$G = \begin{matrix} \\ x_1 \\ x_2 \\ \vdots \\ x_m \end{matrix} \begin{matrix} \begin{matrix} A_1, & A_2, & ..., & A_n \end{matrix} \\ \begin{bmatrix} \mu_{11}, & \mu_{12}, & ..., & \mu_{1n} \\ \mu_{21}, & \mu_{22}, & ..., & \mu_{2n} \\ \vdots & \vdots & & \vdots \\ \mu_{m1}, & \mu_{m2}, & ..., & \mu_{mn} \end{bmatrix} \end{matrix}$$

If $V_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{in})$ is the line vector of G corresponding to element x_i, we write:

$$\overline{V}_i = (1 - \mu_{i1},\ 1 - \mu_{i2} ..., 1 - \mu_{in}).$$

for the complementary vector.

The non-similarity between elements x_j and x_k is measured by means of a distance between vectors V_j and V_k. The more known distances are:

— *the absolute Hamming distance*:

$$d(V_j,\ V_k) = \sum_{i=1}^{n} |\mu_{ji} - \mu_{ki}|,\ j, k = 1, 2, ..., m \tag{5}$$

— *the relative Hamming distance*:

$$\delta(V_j\ V_k) = \frac{d(V_j,\ V_k)}{n},\quad j, k = 1, 2, ..., m \tag{6}$$

— *the absolute euclidean distance*:

$$e(V_j,\ V_k) = \sqrt{\sum_{i=1}^{n} (\mu_{ji} - \mu_{ki})^2},\quad j, k = 1,2, ..., m \tag{7}$$

— *the relative euclidean distance*:

$$\varepsilon(V_j, V_k) = \frac{e(V_j,\ V_k)}{\sqrt{n}},\quad j, k = 1, 2, ..., m \tag{8}$$

Further, an indicator to express the similarity between two elements of set M is put forward.

Definition. The similarity coefficient between two elements is defined thus:

$$\rho_{jk} = \frac{\alpha V_j V_k + \beta \overline{V}_j \overline{V}_k}{\alpha V_j V_k + \beta \overline{V}_j \overline{V}_k + \widetilde{d}(V_j, V_k)} \tag{9}$$

where:

α is an importance coefficient for the similarity of two elements by properties towards which both elements have a high membership degree;

β is an importance coefficient for the similarity of two elements by properties towards which both elements have a low membership degree;

$\widetilde{d}$ is any of the distances defined above.

Coefficients α and β meet the following conditions:

$$\alpha + \beta = 1,\ 0 \leqslant \alpha \leqslant 1,\ 0 \leqslant \beta \leqslant 1,\ \alpha \geqslant \beta.$$

In general, α is chosen close to 1 and β is chosen close to 0, because the similarity between two elements by properties towards which they have a higher membership degree is more important than the similarity by properties which the elements have not got or towards which they have a low membership degree.

Definition. The non-similarity coefficient between two elements i and j is defined by relationship:

$$\eta_{ij} = 1 - \rho_{ij}. \tag{10}$$

We notice that if $V_j = V_k$, then $\rho_{jk} = 1$, and if V_j has boolean components and $V_j = \overline{V}_k$ (the properties of the elements are complementary), then $\rho_{jk} = 0$.

As a conclusion:

$$0 \leqslant \rho_{jk} \leqslant 1.$$

If ρ_{jk} is close to 1, then element j has a high membership degree towards a class made up of elements very similar in element k. This behaviour of coefficient ρ_{jk} accounts for the name of coefficient of similarity. The non-similarity between two elements is expressed by means of the coefficient defined by relationship (10).

Consider

$$\eta = \| \eta_{ij} \|, \quad i, j = 1, 2, \ldots, m$$

to be the matrix of non-similarity (contrast) coefficients.

It stands to reason that this matrix is symmetrical, that is:

$$\eta_{ij} = \eta_{ji}, \quad i, j = 1, 2, \ldots, m$$

and

$$\eta_{ij} = 0, \quad i = 1, 2, \ldots, m.$$

Definition. The similarity coefficient between an element i and a class of elements C is defined by the relationship:

$$s(i, C) = \frac{\sum_{j \in C} \rho_{ij}}{\text{card } C}, \quad 0 \leqslant s(i, C) \leqslant 1. \tag{11}$$

Definition. The non-similarity coefficient between an element i and a class of elements C is defined thus:

$$\eta(i, C) = \frac{\sum_{j \in C} \eta_{ij}}{\text{card } C}. \tag{12}$$

It is obvious that $\eta(i, C) = 1 - s(i, C)$. Indeed

$$\eta(i, C) = \frac{\sum_{j \in C} \eta_{ij}}{\text{card } C} = \frac{\sum_{j \in C} (1 - \rho_{ij})}{\text{card } C} = 1 - \frac{\sum_{j \in C} \rho_{ij}}{\text{card } C} = 1 - s(i, C). \tag{13}$$

If set C contains very many elements (for instance, more than 100), then it is difficult to compute $s(i, C)$ by means of relationship (11). In this case $s(i, C)$ is approximated by $s(i, C^*)$:

$$s(i, C^*) = \frac{\sum_{j \in C^*} \rho_{ij}}{\text{card } C^*} \tag{14}$$

where C^* is a certain sub-set of C.

For instance, C^* can be obtained by choosing at random a sample of elements of set C.

TYPES OF CLASSIFICATION PROBLEMS

By using the notions given a little before we can build up algorithms for dividing a set M into classes in relation to n properties.

In practice this classification problem can be stated in many ways. Further, we consider the following four types of clasification problems:

I. We are given a set of classes $C_1, C_2, \ldots, C_p$ having a number of elements $n_1, n_2, \ldots, n_p$ and a set $M = \{x_1, x_2, \ldots, x_q\}$, $(M \cap C_i = \varnothing, i = 1,2,\ldots, p)$. It is required to distribute the elements of set M to these classes so that the new classes that have been obtained $C_1' C_2' \ldots, C_p'$ shall offer an optimal classification solution for set $M \cup \left(\bigcup_{i=1}^{p} C_i \right)$ in relation to all classifications where $C_i \subseteq C_i' (i = 1, \ldots, p)$.

II. We are given a set M. It is required to divide this set into p (known) classes according to n known properties, having as a criterion the maximal similarity between the elements of the same class.

III. We are given a set M. It is required to check the optimality of thiss classification in relation to properties $A_1, A_2, \ldots, A_n$. (The optimal criterion it the maximal similarity between the elements of the same class.) In case that

the division into classes is not optimal, it is required to give a better solution maintaining the number of classes.

IV. We are given a set M. It is required to divide this set into a most convenient number of classes, according to n known properties. The resulting number of classes can assume anyone of values 1, 2, ..., s ($s \leqslant$ card $M - 1$).

Further we shall dwell on the former two types of problems only.

Type I problems. The algorithm for solving this type of problem consists of the subsequent steps:

Step 1

We consider an element whatever $x_i \in M$ and compute the distances $s(x_i, C_k)$ between this element and classes $C_k (k = 1, ..., p)$. If the number of elements of set C_k is of hundreds or larger, then we consider a sub-set C_k obtained through sampling and compute distance $s(i, C_k^*)$ by making use of relationship (14).

Step 2

We determine C_r for which $s(x_i, C_r) = \min s(x_i, C_k) k = 1, ..., p$.

Step 3

We consider $C_r' = C_r \cup \{x_i\}$ and $M' = M - \{x_i\}$.

Step 4

If $M' = \phi$ we go on to step 5; otherwise, we repeat the algorithm beginning with step 1 with $C_r = C_r'$ and $M = M'$.

Step 5

The algorithm is over. The results are printed.

Type II problems. In case that the number of elements to be classified is rather small (up to 100 elements), an algorithm using exhaustively the elements of the matrix of non-similarity coefficients η_{ij} is applied.

In case that the number of p classes is larger than 3, it is necessary to define an admissible non-similarity coefficient η_{adm} within the limits of which two elements can be members of the same class. Having defined this, the algorithm mainly consists of the following two stages.

Stage 1: Building up n non-void classes.

Stage 2: Distributing the remaining elements into the n classes.

Thus, at stage 2, the type I problem shall be solved up by means of the algorithm described before.

The two stages for solving type II problem can be viewed in Fig. 1.

In case that the number of elements to be classified is large of hundreds or even thousands, the contrast matrix can no longer be built up. In such case, pairs of random numbers are generated (i_h, j_h), $j = 1, 2, ..., N$ in the interval $[1, n]$ observing condition $i_h \neq j_h$ and $(i_h, j_h) \neq (i_k, j_k)$, $(\forall h)$, $(\forall k)$, $k \neq h$, $h = 1, 2, ..., N$.

For each pair (i_h, j_h), contrast $\eta_{j_h j_h}$ is computed. To the rest, the algorithm remains unaltered.

An important problem is the checking of sample N. This number must be large enough so as the maximum estimated on the basis of this sample shall have a small probability of differing significantly from the real maximum.

If a selection of non-similarity coefficients is arranged in an increasing order

$$\eta_{u_1, v_1} \leqslant \eta_{u_2, v_2} \leqslant \ldots \leqslant \eta_{u_N, v_N}.$$

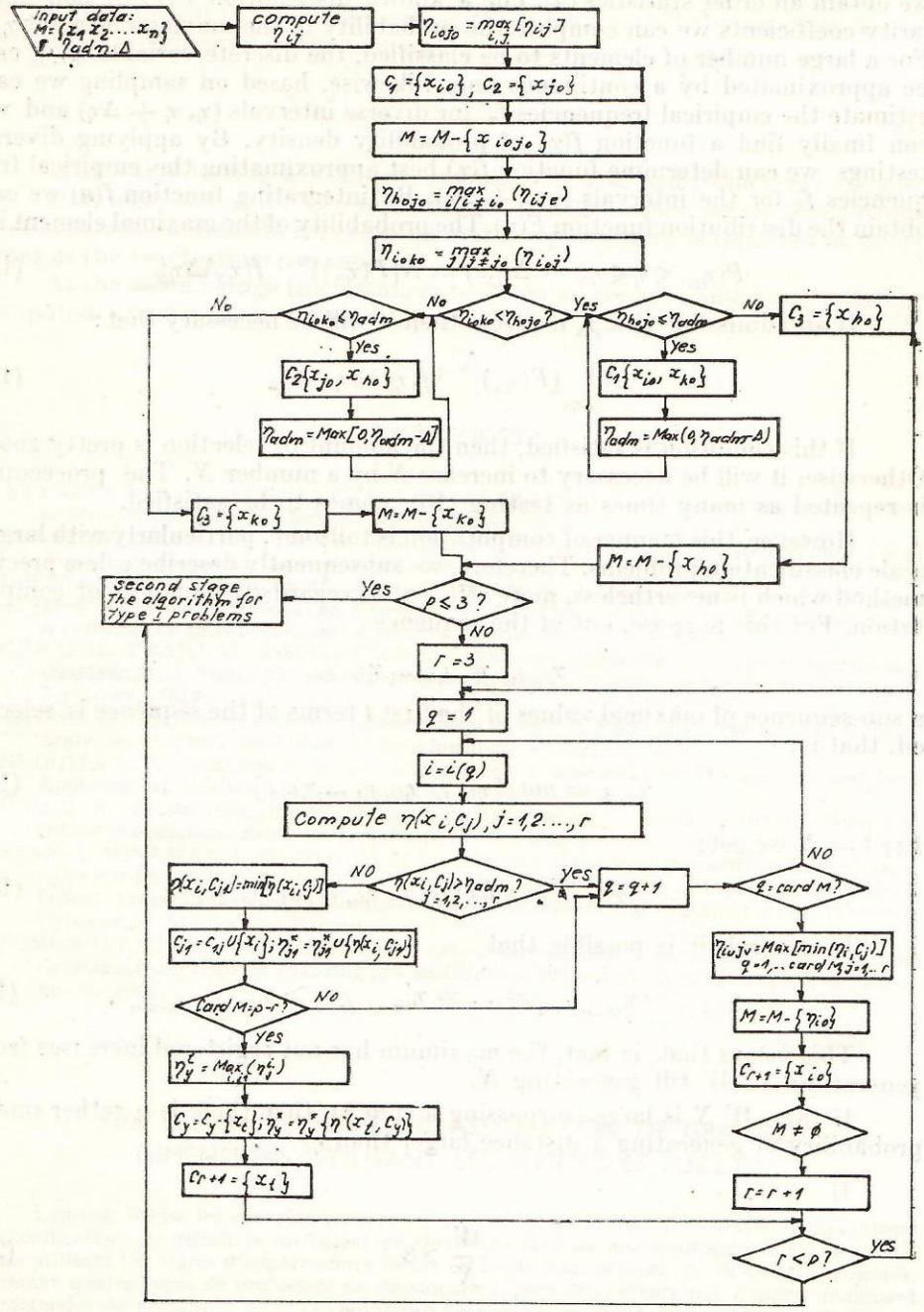


Fig. 1 – Flowchart of classification algorithm for type II problems.

we obtain an order statistics [4]. For a known distribution $F(\eta)$ of non-similarity coefficients we can compute the probability of the maximal element η_{iN}. For a large number of elements to be classified, the discrete variable η_{i_h, j_h} can be approximated by a continuous one. Likewise, based on sampling we can estimate the empirical frequencies f_η for diverse intervals $(\eta, \eta + \Delta\eta)$ and we can finally find a function $f(\eta)$ of probability density. By applying diverse testings we can determine function $f(\eta)$ best approximating the empirical frequencies f_η for the intervals $(\eta, \eta + \Delta\eta)$. By integrating function $f(n)$ we can obtain the distribution function $F(\eta)$. The probability of the maximal element is:

$$P(\eta_N, \leqslant \eta \leqslant \eta_N + \Delta\eta_N) = N[F(\eta_N)]^{N-1} f(\eta_N)\Delta\eta_N \tag{15}$$

If an admissible risk β_α is given, then it will be necessary that:

$$N\int_{\eta_N}^{\infty} [F(\eta_N)]^{N-1} f(\eta_N)\, d\eta_N \leqslant \beta_\alpha \tag{16}$$

If this condition is satisfied, then the amount of selection is pretty good. Otherwise, it will be necessary to increase N by a number N. The proceeding is repeated as many times as testing (16) comes to be satisfied.

However, this manner of computation is toilsome, particularly with large-scale classification problems. Therefore, we subsequently describe a less precise method which is nevertheless, more efficient as regards the amount of computation. For this purpose, out of the sequence

$$\eta_{i_1, j_1},\ \eta_{i_2, j_2}, \ldots, \eta_{i_N, j_N}$$

a sub-sequence of maximal values of the first t terms of the sequence is selected, that is:

$$\eta'_{i_t, j_t} = max[\eta_{i_1, j_1},\ \eta_{i_2, j_2}, \ldots, \eta_{i_t, j_t}] \tag{17}$$

For $t = N$ we get:

$$\eta'_{i_N, j_N} = max[\eta_{i_1, j_1},\ \eta_{i_2, j_2}, \ldots, \eta_{i_N, j_N}] \tag{18}$$

In general, it is possible that

$$\eta'_{i_N, j_N} = \eta'_{i_{N-1}, j_{N-1}} = \ldots = \eta'_{i_{N-w}, j_{N-w}} > \eta'_{i_{N-w-1}, j_{N-w-1}}\ . \tag{19}$$

This means that, in fact, the maximum has not registered increases from generating $N-W$ till generating N.

If ratio W/N is large (surpassing a step β), then there is a rather small probability of generating a distance larger than η'_{i_N, j_N}.

If

$$\frac{W}{N} \geqslant \beta \tag{20}$$

the number of samples is assumed to be satisfactory from the standpoint of the probability of a maximal closeness and we can go on to another

testing. To this end the relative increase of the maximum between steps $N-W-1$ and N is compared to the admissible precision ε_a:

$$\frac{\eta'_{N,jN} - \eta'_{i_{N-w-1},\, j_{N-w-1}}}{W \cdot \eta'_{i_N,\, j_N}} \leqslant \varepsilon_\alpha \quad . \tag{21}$$

If also this testing is satisfied, sample N is assumed to be large enough.

If at least one of the two testings (20) and (21) is not satisfied, a surplus N of contrast coefficients is generated and the proceeding is repeated as many times as the two testings are satisfied.

At the second stage the distances between an element and a class shall be computed by means of relationship (12).

REFERENCES

1. AYVAZYAN, S. A., BEZHAYEVA Z. Y. STAROVEROV O. V.; *Klassifikatsiya mnogomernykh nablyudenii*. Moskow, Izd. Statistika, 1974.
2. CONSTANTINESCU P.; *The Classification of a Set of Elements with Respect to a Set of Properties*. Computer Journal, Jan. 1966.
3. CONSTANTINESCU P.: *Clasificarea unei mulțimi de elemente în raport cu o mulțime de proprietăți* (Classification of an Elements Set with Respect to a Set of Properties). Studii și cercetări de calcul economic și cibernetică economică, no. 3, 1967.
4. CIUCU G., CRAIU V., ȘTEFĂNESCU A.: *Statistica matematică și cercetare operațională* (Mathematical Statistics and Operations Research). Editura didactică și pedagogică, Bucharest, 1974.
5. MOISIL G.: *Lecții despre logica raționamentului nuanțat* (Lessons about the Logic of Gradual Sentence). Editura Științifică și Enciclopedică, Bucharest, 1975.
6. NEGOIȚĂ C. V., RALESCU D. A.: *Mulțimi vagi și aplicațiile lor* (Fuzzy Sets and their Applications). Editura tehnică, Bucharest, 1974.
7. RAO M. R.: *Cluster Analysis and Mathematical Programming*. Journal of the American Statistical Association, Sept. 1971, Volume 66, no. 335, pp. 622–626.
8. STANCU MINASIAN I. M., STOICA M.: *Aplicație a taxonomiei în stabilirea liniilor tehnologice într-un proces de fabricație* (An Application of Taxonomy in Determining the Technological Lines in a Manufacturing Process). Economic Computation and Economic Cybernetics Studies and Research no. 4, 1969.
9. TOMESCU I.: *Asupra unor probleme combinatorii care intervin în teoria clasificării* (On some Combinatoric Problems Occuring in Classification Theory) Studii și Cercetări Matematice, no. 9, 1967.

6.07.1976

SUR LES PROBLEMES DE CLASSIFICATION DE GRANDES DIMENSIONS UTILISANT LES ENSEMBLES FUZZY

L'auteur utilise les concepts concernant les ensembles fuzzy dans l'étude des problèmes de classification. On définit le coefficient de similitude (soit de non-similitude) entre deux éléments utilisant les degrés d'appartenance de ces éléments à un nombre de propriétés données. On donne quatre types de problèmes de classification, pour deux d'entr'eux l'auteur indiquant les méthodes de résolution dans l'hypothèse d'un nombre restreint d'éléments jusqu'à l'ordre des centaines, ainsi que dans l'hypothèse d'un grand nombre d'éléments (mille ou dizaine de mille). Pour le cas des dimensions importantes on utilise la méthode Monte Carlo et on présente un critérium sur la précision obtenue.

О ВОПРОСАХ КЛАССИФИКАЦИИ БОЛЬШИХ РАЗМЕРОВ ПРИ ИСПОЛЬЗОВАНИИ НАОПРЕДЕЛЕЧЕННЫХ МНОЖЕСТВ

В работе использованы понятия, касающиеся неопределенных множеств в исследованиях вопросов классификации. Определяется коэффициент сходства (соответственно различия) между двумя элементами при использовании степени принадлежности их к числу данных свойств. Дается четыре типа задач по классификации, для двух из них указывается метод решения с предположением органиченного числа элементов (в переделе сотен), а также и с предположением большого числа элементов (тысячи или десятки тысяч). В случае больших размеров используется метод Монте-Карло и дается критерий достигнутой точности.