APPLICATION OF FUZZY CLUSTERING AT MEDICAL DATA PROCESSING

JOZEFÍNA PETROVIČOVÁ

Abstract. The theory of fuzzy sets is used for investigation of health state risk stratification of myocardial infarction patients. They are divided into three groups: high risk group, intermediable risk group and low risk group of reinfarction or sudden cardiac death by ISODATA algorithm and by using noninvasive cardiology diagnostic data. This method would provide a tool for assesing of the degree of risk in individual myocardial infarction patients and for evaluation the defined groups regarding the prognosis.

Keywords: fuzzy clustering, ISODATA algorithm, myocardial infarction.

1. Introduction

Fuzzy sets (Zadeh [14]) have been widely applied in medicine most at discriminating pictures (Cormier [5], Mesiar [9]) where data have to be divided into determinate classes. The latest significant application of fuzzy sets (Colombet [4]) has arisen in the project of a help to myocardial infarction (MI) people. Fuzzy sets have been also utilized at the Medical Faculty, P.J. Šafárik University at dividing patients into groups with different level of reinfarction or sudden death (Petrovičová [11]).

MI patients after discharging from hospital form a risk group, because they are endangered by either a reinfarction or a sudden cardiac death. According to the clinical, demographic and anamnestic risk factors as well as by using noninvasive cardiology diagnostic data, patients are divided by computer algorithm into three groups (in accordance with Breithardt [2], Deedwania [6], Figueredo [7]):

a) high risk group, approximate 20% of patients with 20-50% annual mortality,

b) intermediate risk group, 55% of patients with 10% annual mortality, and

c) low risk group 25% of patients with 2-5% annual mortality.

Therefore it is necessary to foresee a risk of sudden cardiac death or a repeated coronary obstruction and to affect it positively by a suitable individual therapeutic

procedure. Short and long term prognosis of MI patients depends on many clinical and laboratory factors which can be presented by fuzzy sets.

2. Mathematical description of a problem

Let n be a number of patients included into monitoring, let p be a number of state factors recorded in each patient. The observed results of all patient's health condition could be arranged into a matrix of real numbers

$$\mathbb{X} = (x_{i,j}); \quad i = 1, \cdots, p \; ; \quad j = 1, \cdots, n$$

representing the collection of input state data. Health condition of each patient is described by one p-element vector $X_j = (x_{1,j}, \cdots, x_{p,j}) \quad j = 1, \cdots, n$ which is coincident with the particular j-th column of matrix $\mathbb{X}$. The individual elements $x_{i,j}$ of these vectors are composed of measured or estimated diagnostic data (pressure, temperature,...), but also of subjective patient's feeling and doctor's assumptions, and they have a fuzzy character.

The task of mathematical processing is to make the classifying of n examined patients in c risk groups (c is a natural number, $2 \leq c \leq n-1$) objective and automatic on the basis of their health condition, considering fuzzy character of a classification process. The vectors X_j of a set $\{X_1, \cdots, X_n\}$ must be sorted to c not-empty clusters $A_k \quad k = 1, \cdots, c$ so that no two cluesters have the common element, and that clusters should form the partition of this set (Riečan [12]). Clusters represent certain classification similarity classes, in which elements of one cluster, according to a certain criterion, resemble each other more than elements of another cluster.

In case that $\{X_1, \cdots, X_n\}$ is a finite set, its each partition can be characterized by a matrix with the number of rows c equal to the number of partition elements A_k, and the number of columns equal to the number of state vectors X_j:

$$\mathbb{U} = (u_{k,j}) \; ; \quad k = 1, \cdots, c \; ; \quad j = 1, \cdots, n$$

For the elements of partition matrix $\mathbb{U}$ it holds:

$$u_{k,j} = \begin{cases} 1, & \text{if} \quad X_j \in A_k \; , \\ 0, & \text{if} \quad X_j \notin A_k \; . \end{cases}$$

In this classical case, the partition unambigously classifils X_j vector in one and only one set $A_1, \cdots, A_c$, that is in the A_k set, for which $u_{k,j} = 1$. In a fuzzy case, which can be the patient's health condition, such unambigious decision is not possible. From the nature of these phenomena it follows that there exists only a certain degree of membership of X_j vectors to A_k sets. In case of fuzzy partition, for the elements $u_{k,j}$ of a matrix partition $\mathbb{U}$ it holds:

$$\begin{aligned} u_{k,j} &\in \langle 0,1\rangle \qquad \text{for all} \quad k = 1,\cdots,c; \quad j = 1,\cdots,n; \\ \sum_{k=1}^{c} u_{k,j} &= 1 \qquad \text{for all} \quad j = 1,\cdots,n; \\ 0 < \sum_{j=1}^{n} u_{k,j} &< n \qquad \text{for all} \quad k = 1,\cdots,c. \end{aligned}$$

3. Medical characterization of problem and diagnostic methods

There are lots of factors which may increase risk of reinfarction or sudden death in MI patients. However, three factors have the dominant position:

- disfunction of the left ventricle,
- residual ischaemia,
- accurrence of heart arrhytmias, or arrythmogenous substrate.

In order to assess the given risk factors the following diagnostic methods may be applied:

1. Assessment of left ventricular function - Echocardiography. This method gives systolic pressure and left ventricular ejection fraction - EF.

2. Diagnostic of residual ischaemia - Exercise stress testing (bicycle ergometry, treadmill exercise test).

3. Assessment of electrical instability and presence of arrythmogenous substrate - Heart rate variability (HRV) and baroreflex sensitivity (BRS), signal averaging electrocardiography, time analysis of late potentials, QTc dispersion is calculated from the 12-lead ECG as the difference between the maximum and minimum QTc (Breithardt [3], La Rovere [8], Moss [10]).

Each examination procedure mentioned above provides results having relatively low prognostic value, if they are interpreted independently. A more complex description of

a patient's current health condition and, consequently, higher prognosis validity of his or her health condition can be archieved by combining the results of all given methods. However, a quantity of compared data requires computer processing and software implementation of suitable methods.

For the numeric realisation of fuzzy partition (Ruspini [13]) an ISODATA algorithm (Bezdek [1]) has been modified for the purposes of a project carried out by our team, and for its personal computer implementation the program was created in Pascal (Petrovičová [11]).

The given algorithm and its computer implementation was successfully verified on the sample of the ten patients after myocardial infarction. Real data of patients at the Fourth Internal Clinic of Faculty Hospital L. Pasteur in Košice have been employed. In table of clinical factors, the first seven factors represent demographic and history risk factors, the last five factors correspond to non-invasive examinations.

Tab.1. Risk factors of 10 tested patients.

Factor\Patient	1	2	3	4	5	6	7	8	9	10
1. Sex	m	m	f	m	m	f	m	m	f	m
2. Age [year]	73	56	59	49	78	64	58	53	64	72
3. Hypertension	+	+	+	+	−	+	−	−	+	+
4. Hyperlipidemia	−	−	+	+	+	+	+	+	+	+
5. Smoking	+	−	−	+	+	+	+	−	+	+
6. Diabetes mellitus	−	−	+	−	−	+	+	+	+	+
7. Obesity	+	+	+	−	−	+	−	−	−	+
8. Ejection fraction [%]	30	60	30	35	34	30	35	55	30	35
9. Late potentials	+	−	+	+	+	−	+	+	+	+
10. Localisation of MI	1	1	1	3	1	1	1	2	1, 2	2
11. Exercise stress testing	+	−	+	+	−	+	+	−	+	−
12. QTc dispersion [ms]	85	40	90	85	90	70	86	50	85	87

4. Data processing and discussion

Medical setting described by the Tab.1 can be transformed to input data matrix $\mathbf{X}$. With respect to the number of patients $n = 10$, and a number of examined factors

$p = 12$, $\mathbb{X}$ matrix is composed of 10 column vectors, each having 12 components:

$$\mathbb{X} = \begin{pmatrix} 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 0.0 & 1.0 \\ 73.0 & 56.0 & 59.0 & 49.0 & 78.0 & 64.0 & 58.0 & 53.0 & 64.0 & 72.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 0.0 & 0.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 30.0 & 60.0 & 30.0 & 35.0 & 34.0 & 30.0 & 35.0 & 55.0 & 30.0 & 35.0 \\ 1.0 & 0.0 & 1.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 3.0 & 1.0 & 1.0 & 1.0 & 2.0 & 1.5 & 2.0 \\ 1.0 & 0.0 & 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 0.0 & 1.0 & 0.0 \\ 85.0 & 40.0 & 90.0 & 85.0 & 90.0 & 70.0 & 86.0 & 50.0 & 85.0 & 87.0 \end{pmatrix} .$$

For coding factor No. 10 - Localisation of MI, the following notation has been used: 1 - anterior wall, 2 - diaphragmal wall, 3 - other. The sign "+" denoting the presence of a given risk factor in a patient can be substituted by the digit 1, and the sign "–" denoting the absence of a given risk factor is substituted by the digit 0. Since MI is much more frequent in men than in women, the male sex is coded by 1, and the female sex by 0.

By means of our program for the given vectors $X_1, \cdots, X_n$, and for the given $c \in \langle 2, n-1 \rangle$ can be constructed a partition characterized by the $\mathbb{U}$-matrix of type $(c \times n)$ and the matrix $\mathbb{V}$ of type $(p \times c)$:

$$\mathbb{U} = (u_{k,j}) \qquad k = 1, \cdots, c; \quad j = 1, \cdots, n ,$$

$$\mathbb{V} = (v_{i,k}) \qquad i = 1, \cdots, p; \quad k = 1, \cdots, c .$$

Let us describe the columns of $\mathbb{V}$ matrix by vectors $V_1, \cdots, V_c$, whose components are for the given k real numbers $v_{i,k}$, $i = 1, \cdots, p$. These vectors themselves have a character of some prototypes of individual classes, for example, a typical representative of a certain disease. Fuzzy partition is calculated from the condition, so that the functional

$$J(\mathbb{U}, \mathbb{V}) = \sum_{k=1}^{c} \sum_{j=1}^{n} u_{k,j} \cdot \|V_k - X_j\|^2$$

should have a minimal value. In general, for the square norm of two p-element vectors V_k and X_j it holds:

$$\|V_k - X_j\|^2 = \sum_{i=1}^{p} (v_{i,k} - x_{i,j})^2 .$$

The calculation of optimal values for matrix partition is performed by an iteration method, and it is completed when the "distance" d_{s+1} of two successive approximations (in iteration cycles s and $s+1$) is smaller than formerly given small positive number $\varepsilon > 0$. The mathematical formulation of this accuracy condition is as follows:

$$d_{s+1} = \|\mathbb{U}^{(s+1)} - \mathbb{U}^{(s)}\| = \sqrt{\sum_{k=1}^{c}\sum_{j=1}^{n}\left(u_{k,j}^{(s+1)} - u_{k,j}^{(s)}\right)^2} < \varepsilon \,.$$

For optimal will be regarded the partition

$$\mathbb{U}^{(s+1)} = \left(u_{k,j}^{(s+1)}\right) \,,$$

which can be obtained by the following procedure:

First step:

a) We null the counter of a number of iteration cycles s, i. e. we put $s = 0$.

b) We can choose an arbitrary partiton $\mathbb{U}^{(s)}$ characterized by elements $u_{k,j}^{(s)}$ as zero approximation of a matrix partition.

Second step:

For the given s and for $i = 1, \cdots, p$ and $k = 1, \cdots, c$, we can calculate the matrix $\mathbf{V}^{(s)}$ elements according to the relation:

$$v_{i,k}^{(s)} = \frac{\sum_{j=1}^{n} u_{k,j}^{(s)} \cdot x_{i,j}}{\sum_{j=1}^{n} u_{k,j}^{(s)}} \,.$$

Third step:

We can construct a new fuzzy partition represented by the elements $u_{k,j}^{(s+1)}$ as a further approximation of a matrix partition $\mathbb{U}^{(s+1)}$. The following two cases appear:

a) If for all of $k = 1, \cdots, c$ and $j = 1, \cdots, n$ it holds $V_k^{(s)} \neq X_j$, then

$$u_{k,j}^{(s+1)} = \frac{1}{\|X_j - V_k^{(s)}\|^2 \cdot \sum_{l=1}^{c} \frac{1}{\|X_j - V_l^{(s)}\|^2}} \,.$$

b) If there exist such $k \in \langle 1, c\rangle$, that for at least one $j \in \langle 1, n\rangle$ holds $V_k^{(s)} = X_j$, we put $u_{k,j}^{(s)} = 1$ for the smallest natural k, for which the given equality holds, $u_{k,j}^{(s)} = 0$ for another k.

After this modification calculation is returned to the second step and we again calculate the components of vectors $V_k^{(s)}$.

Fourth step:

We perform the accuracy test of an iteration process by comparing the "distance" of the two successive approximations of the matrix of a partition $d_{s+1} = \|\mathbb{U}^{(s+1)} - \mathbb{U}^{(s)}\|$ having ε value. If the difference is smaller than ε, the calculation may be completed, if not, we continue in calculations applying the next step.

Fifth step:

We change s value of the number of iteration cycles in the counter into a value higher by 1 than it is its actual value. We carry out a test whether this new s values has reached the stated maximum value. If yes, the calculation is completed, if not, the calculation is returned to the second step. The determination of maximum number of iteration cycles guarantees the correct completed calculation in case of unsuitable ε option and a failure in finishing calculations by using an accuracy condition in the preceding step.

In the first step, based on the input diagnostic data, the zero approximation of fuzzy partition matrix can contain, for example, the following elements:

$$\mathbb{U}^{(0)} = \begin{pmatrix} 0.7 & 0.1 & 0.5 & 0.4 & 0.4 & 0.5 & 0.1 & 0.1 & 0.4 & 0.7 \\ 0.2 & 0.5 & 0.4 & 0.5 & 0.5 & 0.4 & 0.4 & 0.2 & 0.5 & 0.2 \\ 0.1 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.7 & 0.1 & 0.1 \end{pmatrix}.$$

The interpretation of this option is as follows (see Tab. 2). The first, high risk group comprises patients having ordinal numbers 1, 3, 6 and 10 (maximal elements in the 1st, 3rd, 6th and the 10th matrix columns are in the 1st row: $u_{1,1}^{(0)} = 0.7$, etc.). Into the intermediate risk group are included patients No. 2, 4, 5 and 9 (maximal elements in the 2nd, 4th, 5th and the 9th columns are in the 2nd row: $u_{2,2}^{(0)} = 0.5$, etc.). The third, low risk group is formed by patients No. 7 and 8 (maximal elements in the 7th and 8th columns of a matrix are in the 3rd row: $u_{3,7}^{(0)} = 0.5$, etc.).

It is sufficient, if the option of the zero approximation of fuzzy partition gives an approximate real state. The initial values which are for distant from reality manifest only in the higher number of iteration cycles required for converging the calculation into an optimal solution.

In the second step, we calculate elements $v_{i,k}^{(0)}$ $(i = 1, \cdots, 12,\ k = 1, \cdots, 3)$ of matrix $\mathbf{V}^{(0)}$. In the third step we calculate norms $\|X_j - V_k^{(0)}\|^2$ for each $k = 1, \cdots, 3$ and $j = 1, \cdots, 10$. Because they all are non-zero, we can approach the calculation of the first approximation of elements $u_{k,j}^{(1)}$ of fuzzy partition matrix $\mathbb{U}^{(1)}$. As the difference of both approximations $d_1 = \|\mathbb{U}^{(1)} - \mathbb{U}^{(0)}\| = 0.971$ still exceeds an acceptable error ($\varepsilon = 0.01$), and during first cycle were realised up to five patient transitions among risk groups, we can approach the calculation of the second approximation of fuzzy partition matrix $\mathbb{U}^{(2)}$.

After the eighth iteration cycle, patient transitions among risk groups are finished. However, this accuracy $d_8 = \|\mathbb{U}^{(8)} - \mathbb{U}^{(7)}\| = 0.089$ is insufficient for finishing an iteration cycles and, therefore, approximation calculation must proceed. Finally, after the nineteenth iteration cycle, when matrix fuzzy partition is

$$\mathbb{U}^{(19)} =$$

$$\begin{pmatrix} 0.854 & 0.035 & 0.245 & 0.246 & 0.718 & 0.436 & 0.109 & 0.004 & 0.403 & 0.853 \\ 0.135 & 0.037 & 0.741 & 0.697 & 0.249 & 0.470 & 0.883 & 0.005 & 0.591 & 0.138 \\ 0.011 & 0.928 & 0.014 & 0.057 & 0.033 & 0.094 & 0.008 & 0.991 & 0.006 & 0.009 \end{pmatrix}$$

the approximation accuracy

$$d_{19} = \|\mathbb{U}^{(19)} - \mathbb{U}^{(18)}\| = 0.0098 < \varepsilon = 0.01$$

is sufficient and thus the task can be finished.

In case we have no sufficient information for the option of ε as for accuracy criteria is concerned, we can finish the iteration process by assigning maximal number of iteration cycles and we can calculate partial results (see Tab. 2) of individual cycles.

Tab. 2.: Evolution of data processing (j - ordinal number of patient, s - ordinal number of approximation cycle, d_s - "distance" of two successive approximations, t_s - number of patient transitions among risk groups in individual cycle, l - belong to the low risk group, m - belong to the middle risk group, h - belong to the high risk group).

$s\backslash j$	1	2	3	4	5	6	7	8	9	10	t_s	d_s
0	h	m	h	m	m	h	l	l	m	h	-	-
1	h	l	h	m	h	m	h	l	h	h	5	0.971
2	h	l	h	m	h	m	h	l	h	h	0	0.250
3	h	l	h	m	h	m	h	l	h	h	0	0.258
4	h	l	h	m	h	m	m	l	h	h	1	0.265
5	h	l	m	m	h	m	m	l	h	h	1	0.256
6	h	l	m	m	h	m	m	l	h	h	0	0.192
7	h	l	m	m	h	m	m	l	m	h	1	0.123
8	h	l	m	m	h	m	m	l	m	h	0	0.089
9	h	l	m	m	h	m	m	l	m	h	0	0.078
10	h	l	m	m	h	m	m	l	m	h	0	0.073
11	h	l	m	m	h	m	m	l	m	h	0	0.067
12	h	l	m	m	h	m	m	l	m	h	0	0.059
13	h	l	m	m	h	m	m	l	m	h	0	0.050
14	h	l	m	m	h	m	m	l	m	h	0	0.041
15	h	l	m	m	h	m	m	l	m	h	0	0.032
16	h	l	m	m	h	m	m	l	m	h	0	0.025
17	h	l	m	m	h	m	m	l	m	h	0	0.019
18	h	l	m	m	h	m	m	l	m	h	0	0.014
19	h	l	m	m	h	m	m	l	m	h	0	0.010

5. Conclusion

On the basis of the given calculations we could objectively conclude that patients denoted by ordinal numbers 2 and 8 belong to the third group, representing a low risk of reinfarction or sudden death (the highest element values in the second and the eigth columns of $\mathbb{U}^{(19)}$ matrix are in the third row: $u_{3,2}^{(19)} = 0.928$ and $u_{3,8}^{(19)} = 0.991$).

Patients numbers 3, 4, 6, 7 and 9 belong to the second groups with intermediate risk (the highest element values of the 3rd, 4th, 6th, 7th and 9th columns of $\mathbb{U}^{(19)}$ matrix are in the second row: $u_{2,3}^{(19)} = 0.741$, etc.). Other patients (1, 5 and 10), with respect to the results of the clinical and stratification diagnostic methods as well as calculation results are included into the first, high risk groups (the highest values of elements in the first, fifth and tenth columns of $\mathbb{U}^{(19)}$ matrix are in the first row: $u_{1,1}^{(19)} = 0.854$, $u_{1,5}^{(19)} = 0.718$ and $u_{1,10}^{(19)} = 0.853$).

Based on complex character of stratification parameters the fuzzy sets could contribute to the diagnostic and therapeutic decision making process. This method would

provide a tool for assesing of the degree of risk in individual myocardial infarction patients and for evaluation the defined groups regarding the prognosis.

References

[1] Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithm*, Plenum Press, New York, 1981.

[2] Breithardt, G., et al., *Prognosis and risk stratification after myocardial infarction*, Europ. Heart J. **16** (1995), 10–19.

[3] Breithardt, G., et al., *Standards for analysis of ventricular late potentials.*, J. Amer. Coll. Cardiol. **17** (1991), 999–1006.

[4] Colombet, I., Joulent, M.-Ch., *A fuzzy prototype-based method to predict myocardial*, Proc. of the 7th Inter. Conf. IPMU, Paris, 1998, pp. 218–224.

[5] Cormier S., et al., *Fuzzy competitive clustering: a new approach*, Proc. of the 7th Inter. Conf. IPMU, Paris, 1998, pp. 1874–1879.

[6] Deedwania, P.C., et al.., *Evidence based, costeffective risk stratification and management after myocardial infarction*, Arch. Intern. Med. **157** (1997), 273–280.

[7] Figueredo, V.M., *Risk stratification after acute myocardial infarction*, Postgrad. Med. **99** (1996), 207–214.

[8] La Rovere, M.T., Bigger, J.T., Marcus, F., *Baroflex sensitivity prediction of total cardiac mortality after myocardial infarction*, Lancet **351** (1998), 478–484.

[9] Mesiar, R., Píš, P., *Fuzzy model of inexact reasoning in medicine*, Comp. meth. and progr. in biomedicine **30** (1989), 1–8.

[10] Moss, A.J., Stern, S., *Noninvasive Electrocardiology - Clinical Aspects of Holter Monitoring*, W.B. Saunders Company, 1996.

[11] Petrovičová, J., *Some Uncertainty Models and Their Applications*, PhD Thesis, Institute of Mathematics, Slovak Academy of Sciences, Bratislava 1999.

[12] Riečan, B., Neubrunn, T., *Integral, Measure, and Ordering*, Kluwer Academic Publishers, Dordrecht and Ister Science, Bratislava, 1997.

[13] Ruspini, E., *Numerical methods for fuzzy clustering*, Inf. Sci. **2** (1970), 319–350.

[14] Zadeh, L.A., *Fuzzy sets*, Inform. and Control **8** (1965), 338–358.

Address

RNDr. Jozefína Petrovičová, PhD; Department of Medical Informatics; Faculty of Medicine; P.J. Šafárik University; Trieda SNP 1; SK - 040 66 Košice; Slovakia

e-mail: jpetrov@central.medic.upjs.sk