REMARKS TO THE DIRECT METHOD OF FUZZY CLUSTERING BASED ON PRESENTATION OF INITIAL DATA BY FUZZY POWERFUL TOLERANCE

D.A.Viattchenin

Intelligent Information Technologies Department, Belarusian State University of Informatics and Radioelectronics, P. Brovka St. 6, Minsk 220013, BELARUS, phone: (375-17) 239 8050; fax: (375-17) 231 0979, viattchenin@mail.ru

ABSTRACT

This short paper describes rules for calculation of tolerance threshold in the direct clustering method based on a presentation of initial data by fuzzy powerful tolerances. The concept of the fuzzy representation of an initial set of elements by fuzzy clusters is considered on a basis of the short essential consideration of the problem of fuzzy cluster analysis. General plan of the clustering procedure is presented and rules for tolerance threshold calculation are proposed. Three strategies of clustering are described shortly. Some preliminary conclusions are discussed and ways of perspective investigations are outlined.

1. INTRODUCTION

1.1 Preliminaries

Cluster analysis is structural approach to solving of the problem of objects classification without training samples. Clustering methods aim at partitioning of a set of objects into subsets, called clusters, so that the objects belonging to the same cluster as similar as possible and the objects belonging to different clusters are as dissimilar as possible. Heuristic approach, hierarchical approach, optimization approach and approximation approach are main groups of methods of cluster analysis.

Since the fundamental Zadeh [9] paper was published, fuzzy sets theory has been applied to many areas such as learning, decision-making, control and classification. The idea of fuzzy approach to clustering problems was outlined by Bellman, Kalaba and Zadeh [1]. The first formal framework of fuzzy clustering methods was proposed by Ruspini [8].

The optimization approach to fuzzy clustering is most widespread approach. However, heuristic algorithms are simple and very effectual in many cases. In particular, direct clustering algorithms have a high level of an essential clearness and a low level of a complexity.

The direct clustering method based on a presentation of initial data by fuzzy powerful tolerances was proposed in [7]. Some remarks to the direct clustering method are considered in the paper. Some essential aspects of the fuzzy clustering problem are considered in the second subsection of the paper section. Basic concepts and outlines of the method are considered in the second section of the paper. Rules for tolerance threshold calculation and clustering strategies are discussed in the third section of the paper. Preliminary conclusions are discussed in the fourth section of the paper.

1.2 Epistemological and Methodological Notes

From epistemological point of view, fuzziness is a property of the structure of an objects set. The structure of an objects set is determined by similarity relation. So, the structure of an objects set can be presented by the fuzzy similarity relations. These relations are intransitive relations very often. Fuzzy similarity intransitive relations are called fuzzy tolerances and fuzzy similarity transitive relations are called fuzzy equivalence relations.

A concept of fuzzy strict feeble similarity relation S_0, a concept of fuzzy feeble similarity relation S_1 and a concept of fuzzy powerful similarity relation S_3 were introduced in [4,5]. So, fuzzy intransitive symmetrical reflexive relation was called fuzzy usual similarity relation S_2 in these papers. We will call these relations as corresponding fuzzy tolerances and these relations will be indicated as T with a corresponding index.

A concept of a fuzzy tolerance is a basis of a concept of the fuzzy cluster, because the fuzzy cluster can be understood as a fuzzy subset of an initial set of elements, which is originated by some fuzzy tolerance, that a tolerance degree of this

http://edoc.bseu.by

fuzzy subset elements is not less than some threshold value. In other words, the value of a membership function of the every element of the fuzzy cluster is the degree of association of the element with the fuzzy cluster. So, fuzzy clustering is a representation of an initial set of objects by fuzzy clusters. Thus, the problem of fuzzy cluster analysis can be defined in general as the problem of discovery of the adequate representation of an initial set of objects by fuzzy clusters. A representation kind and it adequacy are determined by a concrete problem framework. Moreover, an approach to a problem solving determines a technique of fuzzy clustering.

2. Theoretical Premises

2.1 Basic Concepts

Let's consider conceptual and methodological bases of the method. In the first place, let's consider the general definition of the fuzzy cluster concept, which was proposed in [3].

Let $X = \{x_1, \ldots, x_n\}$ is an initial set of elements. Let $T_b, b = \{0,1,2,3\}$ is a fuzzy tolerance on X and α is α-level value of $T_b, b = \{0,1,2,3\}, \alpha \in (0,1]$. Let $\{A^1, \ldots, A^n\}$ are fuzzy sets on X, which are originated by a fuzzy tolerance $T_b, b = \{0,1,2,3\}$.

Definition 1 The fuzzy subset $g^l = \{(x, \mu_{A^l}(x)) \mid x \in X, l \in [1,n]\}$ of the fuzzy set $A^l \subseteq X, l \in [1,n]$ is the fuzzy cluster if a next condition

$$\mu_T(x_i, x_j) \geq \alpha, x_i, x_j \in A^l, l \in [1,n] \quad (1)$$

is met, where $\mu_T(x_i, x_j)$ is the membership function of a fuzzy tolerance $T_b, b = \{0,1,2,3\}$. Thus, α is the tolerance threshold of x_i and x_j elements.

In other words, if a fuzzy tolerance $T_b, b = \{0,1,2,3\}$ is represented by a matrix of the tolerance, then columns or lines of that matrix are fuzzy sets $A^l \subseteq X, l \in \overline{1,n}$ and these fuzzy sets can be considered as clustering components. So, the fuzzy cluster g^l is the fuzzy subset of a fuzzy set $A^l, l \in [1,n]$ if a condition

$$\mu_{g^l}(x) \geq \alpha, x \in A^l, l \in [1,n] \quad (2)$$

is met. Thus, $\mu_{g^l}(x)$ is the degree of association of the element x with the fuzzy cluster $g^l, l \in [1,n]$.

In the second place, let's consider the concept of the fuzzy cluster typical point. Generalization of the typical point concept definition was proposed in [5].

Definition 2 If $T_b, b = \{0,1,2,3\}$ is a fuzzy tolerance on X, where X is an initial set of elements, and g^l is a fuzzy cluster, then some point $x_t \in X$, that satisfies a condition for all $x \in X$

$$\mu_{g^l}(x_t) = \max_x \mu_{g^l}(x), x \in g^l \quad (3)$$

is called a typical point of the fuzzy cluster g^l.

The fuzzy powerful tolerance concept is very important for the consideration. Let's remind the definition of the fuzzy powerful tolerance.

Definition 3 The fuzzy powerful tolerance T_3 is the fuzzy binary intransitive relation which possesses the symmetric property

$$\mu_{T_3}(x_j, x_i) = \mu_{T_3}(x_i, x_j), \forall x_i, x_j \in X \quad (4)$$

and the powerful reflexivity property. The powerful reflexivity property is defined as the condition of reflexivity

$$\mu_{T_3}(x_i, x_i) = 1, \forall x_i \in X, \quad (5)$$

together with the condition

$$\mu_{T_3}(x_i, x_j) < 1, \forall x_i, x_j \in X, x_i \neq x_j \quad (6)$$

Evidently, that for T_3 relation a condition $\mu_{g^l}(x_t) = 1, x_t \in g^l$ is met, where g^l is a fuzzy cluster and x_t is it typical point. Moreover, any fuzzy cluster, which is originated by T_3 relation, has the unique typical point, because the powerful reflexivity property is met. So,

$$\mu_{g^l}(x_l) = 1, \forall g^l, \forall x_l \in X, l = \overline{1,n} \quad (7)$$

If fuzzy cluster is originated by T_3 relation, then this fuzzy cluster is called the fuzzy powerful cluster with center. The unique typical point is the center of the cluster in this case. Thus, if initial data are represented by a matrix of T_3 relation, then any diagonal element of the matrix can be considered as a center of the some powerful cluster with center.

Let's introduce the concept of the fuzzy representation of an initial set by fuzzy clusters.

Definition 4 Let $R(X) = \{g^l \mid l = \overline{1,k}, k \le n\}$ is the family of fuzzy clusters, which are originated by the some fuzzy tolerance $T_b, b = \{0,1,2,3\}$ on an initial set of elements $X = \{x_1, \ldots, x_n\}$. If conditions

$$\sum_{l=1}^{k} \mu_{g^l}(x_i) > 0, i \in \{1, \ldots, n\} \quad (8)$$

$$card(R(X)) \ge 2 \quad (9)$$

are met for all $l = \overline{1,k}, k \le n$, where $card(R(X))$ is a cardinality of the $R(X)$ set of fuzzy clusters, then the family is the fuzzy representation of an initial set of elements $X = \{x_1, \ldots, x_n\}$ by fuzzy clusters $\{g^l\}$.

Obviously, that the concept of the fuzzy representation is more general concept, than concepts of the fuzzy partition and the fuzzy covering. The fuzzy partition and the fuzzy covering are special cases of the fuzzy representation.

A concept of the adequacy of the fuzzy representation depends on a purpose of classification problem in the every concrete case. We will use the concept of a minimal fuzzy representation for illustration of the method.

Definition 5 If $R(X) = \{g^l \mid l = \overline{1,k}, k \le n\}$ is the fuzzy representation of an initial set $X = \{x_1, \ldots, x_n\}$ by fuzzy clusters and intersection of any two different fuzzy clusters is empty set:

$$card(g^l \cap g^m) = 0, \forall g^l, g^m \in R(X), l \ne m \quad (10)$$

then the fuzzy representation $R(X)$ is the minimal fuzzy representation and it will be indicated as $R_{\min}(X)$.

In other words, if the every element $x_i \in X, i = \overline{1,n}$ is belong to only some one fuzzy cluster with the positive membership degree, then a family of these fuzzy clusters constructs some minimal fuzzy representation $R_{\min}(X)$ and the number of fuzzy clusters must be the least.

A condition of the fuzzy clusters separability was proposed in [6]. A modification of the condition will be used for constructing of the minimal fuzzy representation in the clustering procedure.

Condition 1 Fuzzy clusters g^l and g^m are fully separate fuzzy clusters, if a condition

$$h(A^l \cap A^m) < \alpha,$$

$$A^l, A^m \in \{A^1, \ldots, A^m\}, l \ne m, \alpha \in (0,1] \quad (11)$$

is met, where h symbol is designated different fuzzy clusters intersection height.

Lemma 1 If Condition 1 is met for any two different fuzzy clusters g^l and $g^m, g^l, g^m \in R(X)$ then the fuzzy representation $R(X)$ is the minimal fuzzy representation $R_{\min}(X)$ of an initial set $X = \{x_1, \ldots, x_n\}$.

Proof. The proof of the Lemma 1 is obvious. Really,

$h(A^l \cap A^m) = \sup_{x \in X} \mu_{A^l \cap A^m}(x) =$

$= \sup_{x \in X}(\mu_{A^l} \wedge \mu_{A^m})(x), \forall x \in l, l \ne m,$ and

$\sup_{x \in X}(\mu_{A^l} \wedge \mu_{A^m})(x) < \alpha, \forall x \in l, l \ne m, \alpha \in (0,1],$

because condition (11) is met. However,

$\mu_{g^l}(x) \ge \alpha,$ and $\mu_{g^m}(x) \ge \alpha, \forall x \in X,$

where $l \ne m, \alpha \in (0,1]$

because condition (2) is met. So, a condition

$\sup_{x \in X}(\min\{\mu_{g^l}(x), \mu_{g^m}(x)\}) < \alpha,$

$\forall x \in l, l \ne m, \alpha \in (0,1]$

is contradict to the definition of fuzzy cluster. Thus, an intersection of fuzzy clusters g^l and g^m is empty set. So, the condition of the Definition 5 is met for any fuzzy clusters from $R(X)$.

Q.E.D.

Proposition 1 Let $X = \{x_1, \ldots, x_n\}$ is an initial set of elements and T_3 is fuzzy powerful tolerance on X. Then some unique minimal representation $R_{\min}(X)$ is exists for some unique value of $\alpha \in (0,1]$.

Proof. The proof follows from Definition 1, *Definition 4, Definition 5 and Lemma 1* immediately.

Q.E.D.

Let the structure of initial data is satisfies to conditions of T_3 relation. Thus, the problem of classification of elements of an initial set $X = \{x_1, \ldots, x_n\}$ is the discovery of the unique minimal representation $R_{\min}(X)$ of $X = \{x_1, \ldots, x_n\}$ set by fuzzy powerful clusters with center for some $\alpha \in (0,1]$.

2.2 A Plan of the Algorithm

Let $X=\{x_1,\ldots,x_n\}$ is an initial set of elements and T_3 is the matrix of tolerance associated with coefficients of powerful tolerance $\mu_{T_3}(x_i,x_j)\in[0,1], i,j=\overline{1,n}$ measuring tolerance between elements of $X=\{x_1,\ldots,x_n\}$. Let α is some tolerance threshold, $\alpha\in(0,1]$. A symbol $R^l(X)$ designates some fuzzy representation of $X=\{x_1,\ldots,x_n\}$, where the fuzzy cluster $g^l, l=\overline{1,n}$ consists in $R^l(X)$ by first and symbol $R^l_{\min}(X)$ designates corresponding minimal representation, which is satisfies to conditions of Definition 4 and Definition 5. A symbol g^m designates any fuzzy cluster, that $g^m\in(R(X)\setminus g^l)$.

The general plan of the clustering procedure consists of four stages.

1. The initial representation $R(X)=\{g^1,\ldots,g^n\}$ is constructed on a basis of Definition 1 for some user's value $\alpha\in(0,1]$;
2. The minimal representation $R^l_{\min}(X)$ is constructing for every $l:=1$ to n on a basis of Definition 4 and Condition 1;
3. If $R^l_{\min}(X)$ is not constructed for every $l:=1$ to n then a value of α must be increased and go to 1; else if $R^l_{\min}(X)$ is constructed for different $l, l\in[1,n]$ then a value of α must be decreased and go to 1; else go to 4;
4. For some calculated value of $\alpha, \alpha\in(0,1]$ the unique minimal representation $R^l_{\min}(X)$, $l\in[1,n]$ is constructed; $R_{\min}(X)=R^l_{\min}(X)$;

However, the method can be modified. Firstly, a rule for calculation of the initial value of the α parameter can be elaborated. In the second place, a procedure for calculation of the α parameter on the third stage of the algorithm can be constructed. These problems will be considered in the next section of the paper.

3. On the Methodology of Clustering

3.1 Tolerance Threshold Calculation Rules

The initial representation $R(X)=\{g^1,\ldots,g^n\}$ can be constructed on a basis of Definition 1 for some user's value $\alpha\in(0,1]$. However, first value of the tolerance threshold can be found naturally. If T_3 is the matrix of tolerance associated with coefficients of powerful tolerance $\mu_{T_3}(x_i,x_j)\in[0,1], i,j=\overline{1,n}$ measuring tolerance between elements of $X=\{x_1,\ldots,x_n\}$, then initial value of the α parameter can be calculated as follows:

$$\alpha^1=\min_x \mu_{T_b}(x_i,x_j), \alpha^1>0$$
$$\forall x_i,x_j\in X, b=\{0,1,2,3\} \qquad (12)$$

or

$$\alpha^1=\max_x \mu_{T_d}(x_i,x_j), \alpha^1<\max_x \mu_{t_b}(x_i,x_j),$$
$$\forall x_i,x_j\in X, b=\{0,1,2,3\} \qquad (13)$$

Obviously, that for T_3 relation condition (13) must be looked as follows:

$$\alpha^1=\max_x \mu_{T_3}(x_i,x_j), \alpha^1<1, \qquad (14)$$

because conditions (5),(6) are met.

Let's consider a problem of calculation of the α parameter for third step of the procedure. Two rules can be used for found of the value of the α parameter. These rules can be described as follows:

$$\alpha^f:=\min_x \mu_{T_3}(x_i,x_j), \mu_{t_3}(x_i,x_j)\in(\alpha^c,1] \qquad (15)$$

or

$$\alpha^f:=\max_x \mu_{T_3}(x_i,x_j), \mu_{t_3}(x_i,x_j)\in(\alpha^c,0) \qquad 16)$$

where α^c is the current value of the α parameter and α^f is the following value of the α parameter. The rule (15) can be used for increase of the α parameter on third step of the clustering procedure. The rule (16) can be used for decrease of the α parameter on third step of the procedure.

3.2 A Note on Strategies of Clustering

We can determine a few strategies of clustering. Firstly, if initial value of the α parameter is calculated by formula (12), then the rule (15) can be used for calculation of following values of the α parameter until unique minimal representation will be found. Unique minimal representation will be found, because Proposition 1 is met. This clustering strategy will be called ascending strategy.

Secondly, if initial value of the α parameter is calculated by formula (13), then the rule (16) can be used for calculation of following values of the α parameter until unique minimal representation will be found. This clustering strategy will be called descending strategy.

Thirdly, we can use ascending strategy and descending strategy simultaneously. This approach can be very effective. This strategy will be called oncoming strategy. However, computer realization of this strategy can be difficult.

So, rules (12),(13) can be used on the first step of the algorithm, and rules (15),(16) can be used on the third step of the clustering procedure.

4. Concluding Remarks

4.1 Discussion

In general, the algorithm, which proposed in [7], is similar to the algorithm, which was proposed by Couturier and Fioleau [2]. However, the algorithm is more simply and clear from essential position, because the concept of fuzzy representation is very general and clear from epistemological point of view. Moreover, the rules of calculation of the α parameter, which were proposed here, make the method swift.

So, ascending strategy of clustering is determined by rules (12),(15) and descending strategy of clustering is determined by rules (13),(16). We can consider these clustering strategies as analogues of corresponding hierarchical methods of fuzzy clustering.

4.2 Perspectives

Let's consider some perspective ways of investigations. In the first place, a version of the algorithm for fixed number of clusters and a version for intersecting clusters can be elaborated. Moreover, versions of the algorithm for the fuzzy partition and for the fuzzy covering can be elaborated too. Secondly, ascending and descending algorithms of hierarchical fuzzy clustering can be constructed. Comparison of hierarchical fuzzy clustering algorithms and direct fuzzy clustering algorithms can be very interesting from theoretical point of view and very fruitful for new fuzzy clustering methods elaboration.

References

[1].Bellman R., Kalaba R.. Zadeh L.A. Abstraction and Pattern Classification. Journal of Mathematical Analysis and Applications. Vol. 13, 1966, pp.1-7.

[2].Couturier A., Fioleau B. Recognising Stable Corporate Groups: A Fuzzy Classification Method. Fuzzy Economic Review. Vol. II, 1997, pp.35-45.

[3].Viattchenin D.A. Towards Definition of Fuzzy Cluster Concept. Computer Data Analysis and Modeling: Proc. of the International Conference (September 4-8, 1995, Minsk), Vol. 2./Ed. by Prof. Yu.S.Kharin. BSU, Minsk, 1995, pp. 91-94 (in Russian).

[4].Viattchenin D.A. Some Remarks To Concept of Fuzzy Similarity Relation for Fuzzy Cluster Analysis. Pattern Recognition and Information Processing: Proc. of Fourth International Conference (20-22 May 1997, Minsk, Republic of Belarus). Vol. 1/Ed. by Prof. V.Krasnoproshin et al. Wydawnictwo Uczelniane Politechniki Szczecinskiej, Szczecin, 1997, pp. 35-38.

[5].Viattchenin D.A. On Projections of Fuzzy Similarity Relations. Computer Data Analysis and Modeling: Proc. of the Fifth International Conference (June 8-12, 1998, Minsk). Vol. 2: M-Z/Ed. by Prof. S.A.Aivazyan and Prof. Yu.S.Kharin. BSU, Minsk, 1998, pp. 150-155.

[6].Viattchenin D.A. On Fuzzy Clusters Separability Condition. Pattern Recognition and Information Processing: Proc. of Fifth International Conference (18-20 May 1999, Minsk, Republic of Belarus). Vol. 2/Ed. by Rauf Sadykhov et al. Belarussian State University of Informatics and Radio-Electronics, Minsk, 1999, pp. 47-51.

[7].Viattchenin D.A. Direct Clustering Method Based on Fuzzy Powerful Tolerance. Computer Data Analysis and Modeling: Proc. of the Sixfth International Conference (September 10-14, 2001, Minsk, Belarus). (to appear)

[8].Ruspini E.H. A New Approach To Clustering. Information and Control. Vol. 15, 1969, pp.22-32.

[9].Zadeh L.A. Fuzzy Sets. Information and Control. Vol. 8, 1965, pp.338-353.