

Algorithm of Pattern Recognition with intra-class clustering

Kozlovskii V. A.¹⁾, Maksimova A. J.²⁾

- 1) Institute of Applied Mathematics and Mechanics of National Academy of Sciences of Ukraine, 74 Roza Luksemburg st., Donetsk, Ukraine, 83114, kozlovskii@iamm.ac.donetsk.ua, <http://iamm.ac.donetsk.ua/en/employees/e1146/>
- 2) Institute of Applied Mathematics and Mechanics of National Academy of Sciences of Ukraine, 74, Roza Luksemburg st., Donetsk, Ukraine
lunaplus@mail.ru, <http://iamm.ac.donetsk.ua/en/employees/e934/>

Abstract: This paper deals with a new fuzzy algorithm of pattern recognition on the basis of fuzzy portraits. The fuzzy portraits are formed by integral characteristics of pattern classes. It is determined that further division of some classes of images into clusters increases the quality of pattern recognition algorithm. The paper provides the stages of fuzzy production knowledge base construction on the basis of fuzzy portraits. The analysis of algorithm results on artificial data and data from machine learning problems repository UCI has been made.

Keywords: pattern recognition, fuzzy logic, clustering, data mining.

1. INTRODUCTION

Pattern recognition problems very often include fuzzy aspects. According to Lotfi Zadeh, the transition between classes is more gradual than intermittent. This statement is confirmed by practice of problem solving. In many empirical fields the standard approach that seeks a boundary between several classes gives no satisfactory results. There are different forms of fuzziness, described in [1]. We shall consider the following manifestations of fuzziness: diffusiveness, vagueness and relative position of classes.

In classical pattern recognition problem statement classes of images have to be strongly divided. The separating surface can be a hyperplane or have a rather complicated form. A large group of pattern recognition algorithms is based on estimates computation algorithms (ECO) concept proposed by Yu. I. Zhuravlev [2]. The basic principle of calculating if an object belongs to a class is calculating its closeness to the master object of owner class and farness to the master objects of other classes.

If classes of images have intersections the algorithm for making decisions about whether an object belongs to the class or not plays a big role. Previously, an algorithm based on fuzzy portraits of classes of images was proposed [3]. These fuzzy portraits describe all objects belonging to a certain class of images in general. Fuzzy portraits accumulate the information about elements of the class such as frequency of object popularity, degree of objects "concentration" and number of objects in every class. The suggestion is to use these portraits for representing the results of algorithm in the form of fuzzy sets. This approach to a certain degree solves the problem of ambiguous results in the case of intersecting classes. Information about an object's distance from other classes

has a substantially smaller significance in this situation and is used only for adjustment of fuzzy sets for aims of quality improvement. However the internal structure of classes is also taken into account. The classes of images can be further divided into groups of clusters that can have intersections too. Thus initial fuzzy portraits are expanded through additional structuring.

2. FEATURES OF MODIFICATION OF PATTERN RECOGNITION PROBLEM

The following problem statement will be considered. It is a fuzzy modification of classification problem. Let us suppose that there is a set of training samples $Y = \{(x^{(i)}, v^{(i)}) \mid x^{(i)} \in X, v^{(i)} \in V, i = 1, \dots, n\}$, where $x^{(i)} \in X \subset R^m$ are vectors in the n -dimensional space of images, $V = \{v_i \mid i = 1, \dots, k\}$ – set of classes of images. Pairs $(x^{(i)}, v^{(i)})$ set the correspondence between object $x^{(i)}$ and pattern class $v^{(i)}$. It is necessary to produce the algorithm of identification $a_Y(x) = \tilde{y}$. Vector $x \in X$ is the input of the algorithm, $X \supseteq X$ – the set of possible input data that depends on the problem domain. Generally, a fuzzy set \tilde{y} describing membership x of pattern classes is obtained.

The training samples set can be incomplete, i.e. it may not contain all classes of images. The classes of images may overlap which is a manifestation of a form of fuzziness – diffusiveness of classes of images. For every object that is an input for the algorithm it is necessary to determine its membership degree in every pattern class.

Let us consider a qualitative attribute of an object called "source". A pair of "sources" can produce a pair of images \bar{x}_1 and $\bar{x}_2 : \bar{x}_1 = \bar{x}_2, \bar{x}_1 \in v_i, \bar{x}_2 \in v_j, v_i \neq v_j$, i.e. patterns with same attribute values belong to different classes in the initial training set. In a situation when this attribute is a part of classification goal some classes will be nonseparable and their intersection can be very large, to a point where one class is absorbed by another. In such cases when the researcher fails to find a new attribute that would allow to divide the classes, the problem becomes unsolvable.

Such problems are common in quality control laboratories. In these cases a "source" is a product manufacturer that needs to be identified. Many problems in chemical and food industries can be reduced to this modification of the pattern recognition problem.

For such training samples a recognition algorithm based on the use of fuzzy portraits was proposed in [3].

The result of the algorithm is a fuzzy set $\tilde{y} = \sum_{i=1}^k \mu_{v_i} / v_i$,

where μ_{v_i} - object's membership degree for pattern class v_i . The notation used here and below corresponds to the one introduced in classical theory of fuzzy sets [4].

Algorithm creates a knowledge base of fuzzy productions where every pattern class has its own rule of inference. The fuzzy inference algorithm works with knowledge base.

We consider a modification of the proposed algorithm. The analysis of empirical data often shows that there are subclasses (i.e. clusters) present in pattern classes. That is why it is useful to perform clustering for objects related to class to obtain a more precise fuzzy portrait.

3. ANALYSIS OF INITIAL SET

A necessary step in solving problems of pattern recognition is a preliminary analysis of initial data which is often called exploratory analysis of data. The basic aims of this analysis were formulated by J. Tukey:

- maximum penetration of data;
- the basic structures detection;
- selection of the more important attributes;
- discovering deviations and anomalies;
- testing the basic hypotheses;
- the construction of the initial model.

The study of initial data structure allowed to choose an algorithm for solving the problem. Generally every pattern recognition algorithm works well with certain data structure. For example, linear classifiers were created for linearly separable classes. The EM-algorithm [5] works with intersecting classes and implies the existence of a hidden variable. There is a number of algorithms covering similar situation that produce the result in the form of a fuzzy set.

In practice there are samples with one or more classes that consist of several clusters. Let us consider a case of strongly intersecting clusters that have pronounced centers (typical points) but have no clear border between them. It is possible, for example, to get this kind of data by changing some parameters of industrial technological process. A number of attributes changes which creates new clusters. It is reasonable to use fuzzy clustering approach for such models [1].

Let a coverage Y_S be defined on the set of images X , that is a part of training samples set Y . Let $P = \{P_1, \dots, P_q, \dots, P_m\}$ be a named set of attributes, where m is the number of attributes. The attributes are defined on sets of values $X_{P_q} \in X$, $P_q \in P$. Similarly to the class V_j , let us denote the set of images of training samples corresponding to this class as $\overline{V_j} = \{x_j^{(i_j)} \mid i_j = 1, \dots, t_j\}$, $j = 1, \dots, k$, where k is the number of classes represented in training sample, t_j is the number of objects in the class with number j , and where $\overline{x_j^{(i_j)}} = (x_{j1}^{(i_j)}, \dots, x_{jq}^{(i_j)}, \dots, x_{jm}^{(i_j)})$, $q = 1, \dots, m$. So the

coverage is $Y_S = \{V_1, \dots, V_j, \dots, V_k\}$.

Let the set $D_j^q = pr_{P_q} V_j = \{x_{jq}^{(i_j)} \mid i_j = 1, \dots, t_j\}$ be a projection of class with number j on attribute P_q .

The basic principle of discriminant analysis is "the effect of essential multidimensionality". According to this principle it is necessary to take into account the information about all attributes and to solve the problem in the m -dimensional space [6]. There are at least two reasons to avoid multiple dimensions. Firstly, data with small training samples gives us a sparse set of points which most pattern recognition algorithms do not process correctly. In this case a clustering by one or several attributes may be not worse than general clustering. Secondly, there are difficulties in substantiating the selection of distance measure which is used in multidimensional clustering. That is why it is proposed in this paper to extract main information about dissimilarity of classes component-wise using projection on attributes.

Thus, we have two strategies for finding clusters. The first one is to perform clustering in the m -dimensional space taking into account the previously described shortcomings of this approach. The second approach is to perform fuzzy clustering for each projection D_j^q , $q = \overline{1, m}$, $j = \overline{1, k}$. Let us consider the second strategy.

Most commonly fuzzy clustering algorithms produce a fuzzy partition or fuzzy coverage as a result. Let us consider a family of fuzzy sets $R(X) = \{A^l \mid l = 1, \dots, c\}$, where $A_l = \{(x, \mu_l(x)) \mid \mu_l(x) \in [0, 1], x \in X\}$ are fuzzy sets and c is the number of clusters described by fuzzy sets A_l . $R(X)$ is a fuzzy c -partition when the following condition is satisfied:

$$\sum_{l=1}^c \mu_l(x_i) = 1, i = 1, \dots, n, \mu_l \in [0, 1]. \quad (1)$$

$R(X)$ is a fuzzy coverage when the condition of fuzzy coverage is satisfied:

$$\sum_{l=1}^c \mu_l(x_i) \geq 1, i = 1, \dots, n, \mu_l \in [0, 1]. \quad (2)$$

The fuzzy clustering procedure is performed for every set V_j , $j = 1, \dots, k$ giving a fuzzy partition or a fuzzy coverage $R(V_j)$ as a result. It establishes a correspondence between partition Y_S and set of fuzzy partitions or coverages $R_S = \{R_j(X) \mid j = 1, \dots, k\}$.

The result of such analysis of training sample is the transformed data that is used in the algorithm and it is described by the following data model:

$$M_S = \{(V, R(V)) \mid V \in Y_S, R \in R_S\}. \quad (3)$$

Thus we receive special data structure where a fuzzy coverage corresponds to every set $V \in Y_S$.

4. DATA REPRESENTATION THROUGH FUZZY PORTRAITS OF CLASSES OF IMAGES AND KNOWLEDGE BASE GENERATION

A number of pattern recognition algorithms uses the information about pattern classes structure. In these algorithms the distance function is used to calculate the typical points or build probability distributions of points in classes. [7].

It is suggested to present the proposed data structure M_S as fuzzy portraits.

Let us introduce certain notions before giving the definition of a fuzzy portrait. Let the quintuple of objects $\{name(P_q), T_q, X_{P_q}, G, M\}$ be called the linguistic variable L_q by the attribute P_q , where $name(P_q)$ denotes the name of linguistic variable, $T_q = \{\mu_q^{(w)} \in [0,1] | w = 1, \dots, r\}$ is term-set of linguistic variable values, w is the cluster number, r is the total number of clusters for all pattern classes. Syntactical rule G that generates variable names is trivial in this case, because all terms are atomic and it simply gives terms the name of a class or a subclass. Semantic rule M is represented as an algorithm for forming term-set membership functions.

Let a first-order fuzzy portrait S_j of class V_j be the group of linguistic variables corresponding to clusters R_j :

$$S_j = \{L_q^w | q = 1, \dots, n, w = \arg R_j(X)\}, \quad (4)$$

Let us construct the knowledge base using fuzzy portraits. Fuzzy knowledge base predicate are the elements of fuzzy portraits, i.e. terms of linguistic variables that describe the pattern classes or sub-classes (clusters) through their attributes. To make a decision about whether an object belongs to a class the fuzzy inference mechanism described in [3] is used.

The advantage of this approach is having a meaningful interpretation of the result which is not the case with neural networks, for example.

In pattern recognition problems it is often sufficient to have the answer: "This image does not belong to this class". In such cases it is enough to use only those rules from the knowledge base that relate to particular classes to receive the necessary answer. This approach is important for the situation when the system can not answer the question "What class does this image belong to?" either because of the intersection of several classes, or, in a situation when this class of images is not described in the knowledge base.

The algorithm of knowledge base building has the following major stages.

1. Fuzzy portraits for every class of training sample are created without clustering. A semantic rule based on the frequency of object occurrence for every attribute is used for term-set forming [3].

2. The fuzzy clustering of pattern classes is conducted using D-AFC(c) algorithm of fuzzy clustering for every attribute q which produces fuzzy coverage $R_j^q(X)$ for

every class j . The best fuzzy coverage $R_j(X)$ is determined through solving the equation $R_j(X) = \arg \max_{\forall R_j^q(X)} F(R_j^q(X))$, where $F(R_j^q(X))$ is the

criterion of fuzzy allotment quality:

$$F(R_j^q(X)) = \sum_{w=1}^{c_j^q} \frac{1}{n_w} \sum_{i=1}^n \mu_w^q(x_i) - \alpha \cdot c_j^q, \quad (5)$$

where n_w is the number of elements on the fuzzy cluster w . The D-AFC(c) algorithm is described in [8].

3. If there are adequate fuzzy coverages they are used to adjust fuzzy portraits.

The fuzzy inference is carried out using the generated knowledge base of fuzzy productions. For every fuzzy portrait S_j one fuzzy inference rule is formed in case of no clusters or several fuzzy inference rules for every cluster in case of successful clustering. The linguistic variables values are calculated with fuzzy portraits terms $T_q: \mu_q^{(w)}(\bar{x}) = \mu_q^{(w)}(x_q)$ on fuzzification stage of fuzzy inference algorithm. The aggregation stage is carried out using the operation "and" proposed in [3] that is defined by the function:

$$f(a_1, a_2, \dots, a_n) = \log_2((a_1 + 1)(a_2 + 1) \dots (a_n + 1)) / n, \quad (6)$$

where $a_i \in [0,1]$, $f(\bar{a}) \in [0,1]$. The discrete fuzzy set \tilde{y} defined on the set of classes of images is the result of fuzzy inference algorithm. On it a defuzzification procedure can be carried out to obtain a distinct result if need be.

5. ILLUSTRATIVE EXAMPLE

Let us consider the algorithm's performance on model data in Fig. 1.

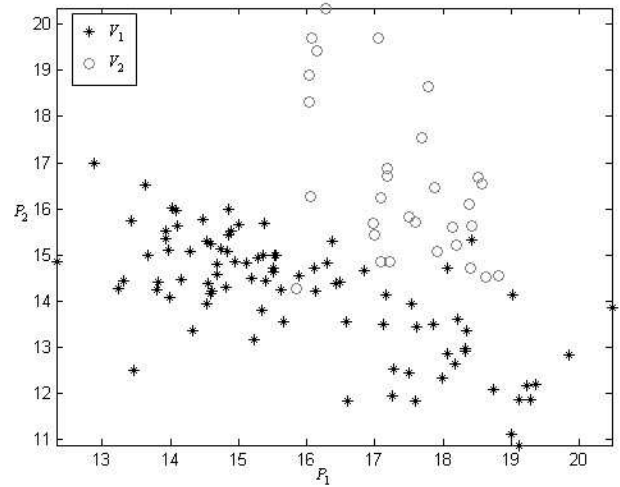


Fig1. – Model data

In this example the training set Y consists of objects from two pattern classes v_1, v_2 . The objects form two-dimensional space: $x^{(i)} \in R^2, i = 1, \dots, 120$. The initial data

is presented as points of two-dimensional space where abscissas and ordinates are attributes P_1 and P_2 correspondingly. The classes of images are linearly inseparable.

The algorithm without clustering proposed in produced about 6% of errors on the model data with distinct interpretation of result. After clustering of every class of training sample two clusters inside the first class were obtained and the modified algorithm improved the recognition rate to 98%.

6. CONCLUSION

In conclusion let us note certain features of the proposed algorithm.

The independent consideration of individual attributes on different stages of the algorithm allows to avoid preliminary normalization of training samples and introducing distance measure which is usually done when processing multidimensional data

That is why the results of the algorithm can be easily interpreted by an expert.

Algorithm is not tied to the topology of the classes layout.

The result of the procedure is given in the form of a fuzzy set which, in many cases, is more natural for the expert, who makes a final decision on whether the pattern belongs to a class and therefore can be easily integrated into decision-making support system.

The proposed variation of the algorithm using the fuzzy portraits created on the basis of individual attributes can be expanded further through using not only conventional one-dimensional projections, but also two-, three-dimensional etc. projections of original images as basic data subsets. Ideology of fuzzy portraits construction will be preserved. It is quite natural to call the one-dimensional variation of the proposed algorithm a first-order algorithm and the variation that takes into consideration not more than k -dimensional projections a k -order algorithm.

The algorithm gave good results on model data with error not more than 2% by leader defined by maximum of membership degree. Algorithm was tested on "wine" problem from UCI data repository [9] with the error rate of 4%.

7. REFERENCES

- [1] D.A. Viattchenin. *Fuzzy methods of automatic classification*. Technoprint. Minsk, 2004. p. 219. (in Russian).
- [2] Yu. I. Zhuravlev. An algebraic approach to the solution of pattern recognition and identification problems. *Probl. Kibernet.* 33 (1978). pp. 5–68. (in Russian).
- [3] V.A. Kozlovskii, A. Ju. Maksimova. Decision of pattern recognition problem with fuzzy portraits of classes. *Artificial Intelligence.* 4 (2010). pp. 221-228. (in Russian)
- [4] L.A. Zadeh The concept of a linguistic variable and its application to approximate reasoning. *Inf. Sci.* 8 (1975). pp. 199-249; *Inf. Sci.* 8 (1975). pp. 301-357. *Inf. Sci.* 9 (1975). pp. 43-80.
- [5] M. Schlesinger. V. Hlavac. *Ten Lectures on Statistical and Structural Pattern Recognition*. Springer. 2002. p. 544.
- [6] S.A. Aivasian. V.M. Buchstaber. *Applied statistics. Classification and reduction of dimensionality*. Moscow. 1989. P. 608. (in Russian).
- [7] M. Friedman. A. Kandel *Introduction to pattern recognition: statistical, structural, neural and fuzzy logic approaches*. World Scientific Publishing Company. Singapore. 1999.
- [8] D.A. Viattchenin. A direct algorithm of possibilistic clustering with partial supervision. *Journal of Automation, Mobile Robotics and Intelligent Systems* 3 (1) (2007). p. 29-38.
- [9] A. Frank. A. Asuncion. *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. 2010.