One Pattern Recognition Method for Complex Geometric Clusters Configuration

Valerii FRATAVCHAN, Tonia FRATAVCHAN
Yuriy Fedkovych Chernivtsi National University
Chernivtsi, Ukraine
vgfratavchan@gmail.com, t.fratavchan@chnu.edu.ua

Abstract - The description of training and classification algorithm for a case when it is impossible to separate clusters and localize convex geometrical forms. There are also recommendations about the clustering of patterns with the use of educational sequence. An informative structure for description of all the clusters is provided, too. There is the description of the algorithm of training for forming the informative description of each class, which is realized by the genetic algorithm.

Keywords—pattern recognition, distinctions space, cluster, cluster-analise, parametrical curves, genetical algorithm, fitness function.

I. INTRODUCTION

Numerical, structural, predicative, verbal type and other signs are used for recognition of patterns. According to the type of signs there are used various methods and algorithms of clustering and classification. For recognition of patterns with numerical signs the most frequently used are probabilistic and «geometrical» methods. These methods are realized quite easily, if the locations of each pattern in numerical multidimensional space of distinctions (so-called clusters) have a simple geometrical structure and can be localized by simple geometrical figures [1, 2, 3]. If the areas of classes have a difficult structure, the convex shells of clusters intersect or one cluster is geometrically located in another cluster, the realization of procedures of recognition becomes substantially complicated. We offer the relatively simple algorithm of recognition for cases when clusters have a simple topological structure, but it is impossible to localize them and separate using the convex geometrical multidimensional figures.

II. THE GENERAL PROBLEM AND BASIC CONCEPTS

The objective is to recognise K patterns

$$C^1, C^2, ..., C^K.$$

Each pattern is described by the numerical n-measurable vector of signs:

$$X = \{ x_1, x_2, ..., x_n \}.$$

To create the adaptive system of recognition there is formed the representative training set of examples of each class – common training set $X_1, X_2, ..., X_M$,

M – size of the sample.

It is needed:

- To divide the training set into subsets, each of which contains the standards of the separate class – to execute procedure of clustering;
- To explore standards of the each cluster and to form for a class the informative structure which unequivocally differs from the analogical informative structures of other classes – to realize the process of training;
- To realize the procedure of classification – authentication of the unknown pattern as a member of only one class.

For the successive consideration we will use the following denotation:

- The distance between two vectors of signs X and Y:

$$d_1(X, Y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}; \quad (1)$$

- The distance between the vector of signs X and the set W:

$$d_2(X, W) = \min_i \{d_1(X, Y_i), Y_i \in W\}; \quad (2)$$

- The distance between two sets U and V:

$$d_3(U, W) = \min_{i,j} \{d_1(X_i, Y_j), X_i \in U, Y_j \in W\}. \quad (3)$$

978-1-5386-1495-2/18/$31.00 ©2018 IEEE

The last formula can be applied to the clusters or to some certain subsets of points belonging to clusters.

III. FEATURES OF THE PROCEDURE OF CLUSTERING

As it is known, the adaptation of the system of recognition to the specific terms of work can be realised by few methods, among which there are «training with a teacher» and autonomous studies, or «training without a teacher». In these cases the amount of clusters is beforehand known. The simplest for realization is «training with a teacher», where the user chooses the class for each sample of the training set.

«Training without a teacher» is realized as some process of optimization where appurtenance of members of the training set is beforehand unknown (fig.1):

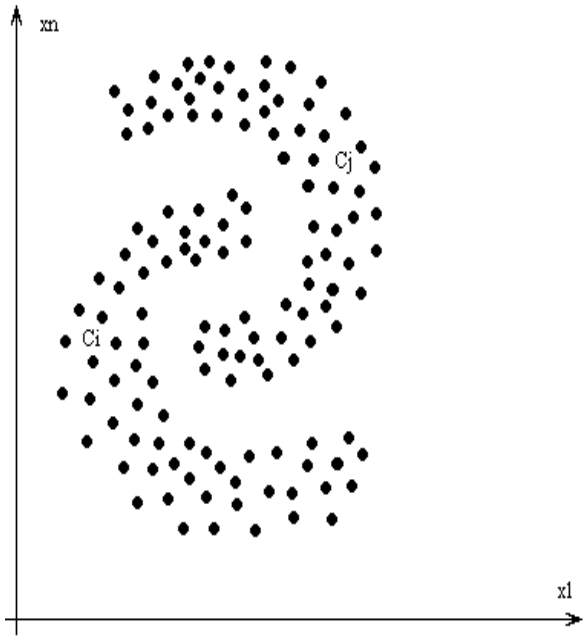

Fig.1. The training set

In our case traditional methods of search of «central points» for clusters for the autonomous clustering will be ineffective. Satisfactory results can be expected from a clustering by the method of «next-door neighbour». Also good results are achieved by application of the clustering of modification of Boruvka's algorithm [4] for the construction of the minimum spanning tree. The Boruvka`s method has got several advantages: during the construction of «branches» of the tree, each of which relates to a separate class, it is possible to calculate the additional estimations of clusters (maximal lengths of ribs of one «branch» and distance between «branches»). These estimations will demonstrate the «quality» of clustering: the successful solution of the followings tasks is only possible when the maximal length of the rib of the «branch» in a current cluster will be substantially «shorter» than distances between a current cluster and other clusters (fig.2):

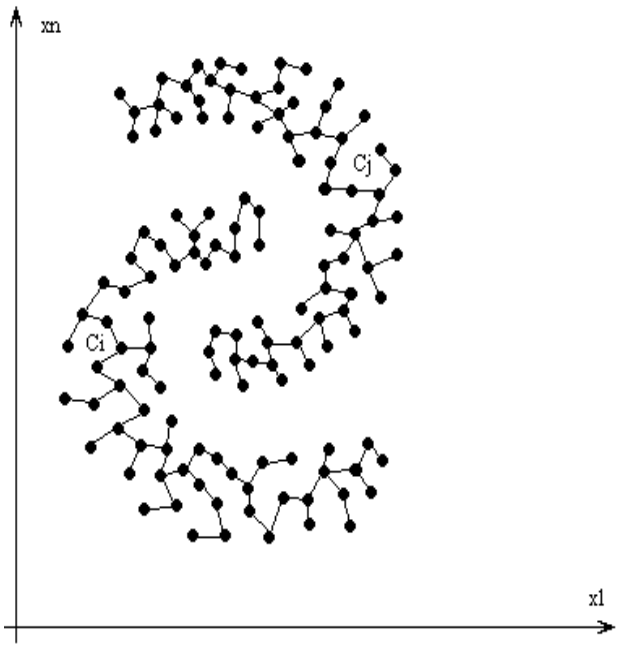

Fig.2. Result of the clustering (Boruvka`s method)

During the realization of the procedure of clustering (Boruvka`s method) there is used a metric d_1 for the calculation of distance between the two samples of the training set and metrics d_2 and d_3 for the calculation of distance between the «branches» of the tree.

IV. CREATION OF THE INFORMATIVE STRUCTURES FOR CLUSTERS DETERMINATION AND DESCRIPTION

It is difficult to determine classes that have the complicated geometrical configuration of clusters by using simple information objects, such as etalon-vectors, discriminant functions, set of inequalities, etc[3]. In this method it is suggested to define each class by some curve which is in a interior of cluster. Such curves must maximally «repeat» the topological structure of the cluster. These properties can be found for the interpolation forms of Bezier or Hermite. The characteristic of these forms is that two points in the space can be connected by the infinite amount of the crooked lines, configuration of which is determined by directions and sizes of output and entry vectors. That means that choosing the initial and final point of trajectory, as well as the initial and the final vector of «motion», it is possible to get the «inertia trajectory» of the necessary form for every cluster (fig.3).

The curves of Bezier and Hermite are determined by the parametric equations of the third order:

$$x(t) = At^3 + Bt^2 + Ct + D, \qquad (4)$$

$t \in [0,1]$.

Thus, the curved line for the cluster ***i*** in the ***n***-measurable space will be described by the system of cubic parametric equations:

$$\begin{cases} x_1^i(t) = A_1^i t^3 + B_1^i t^2 + C_1^i t + D_1^i, \\ x_2^i(t) = A_2^i t^3 + B_2^i t^2 + C_2^i t + D_2^i, \\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots \\ x_n^i(t) = A_n^i t^3 + B_n^i t^2 + C_n^i t + D_n^i, \end{cases} \quad (5)$$

$t \in [0,1].$

The form of the curve depends on the values of coefficients.

$$\{A_j^i, B_j^i, C_j^i, D_j^i\}, j = \overline{1,n}.$$

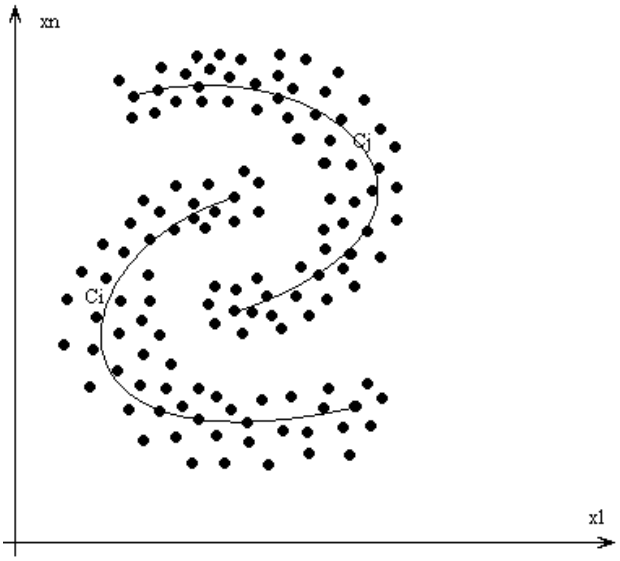

Fig.3. Parametric curve of Bezier (Hermite) for the resulting clusters

While planning the interpolation lines of Bezier and Hermite the values of coefficients of the parametric equations are calculated taking into consideration the values of co-ordinates of the initial point, the final point, start and finish vectors.

The information about the clusters doesn`t contain such data. After calculating the coefficients of the parametric equations coordinates of points on the curve lines are determined by the proper values of the parameter t. Coordinates of the initial point are used to find the value t=0, coordinates of the eventual point – for the value t=1. That means that any array of coefficients $\{A_j, B_j, C_j, D_j\}, j = \overline{1,n},$ defines in the sign space the parametric curve. Thus, the objective is to find the optimal coefficients of the parametric equations for K clusters. This task can be understood as the optimizing process, the process of searching the set of the coefficients of the parametric curves which respond to some additional conditions. The task can be solves by using the genetic algorithm [5,6].

There are two ways of its realization.

The first way is to find the parametrical coefficients for each cluster. The set of those coefficients forms the genotype of the variant of the curve. Fitness-function consists of 2 criteria with the proper weight coefficients (fig.4):

$$F = w_1 \cdot Kr_1 + w_2 \cdot Kr_2 \to \min, \quad (6)$$

where

$$Kr_1 = \sum_{i=1}^{m} d_2(X_i, L), \quad (7)$$

X_i , $i = \overline{1,m}$ – training samples of the cluster ***i***,

L – set of control points on the curve,

$$Kr_2 = \|L\| = \sum_{i=1}^{r-1} d_1(X(t_i), X(t_{i+1})). \quad (8)$$

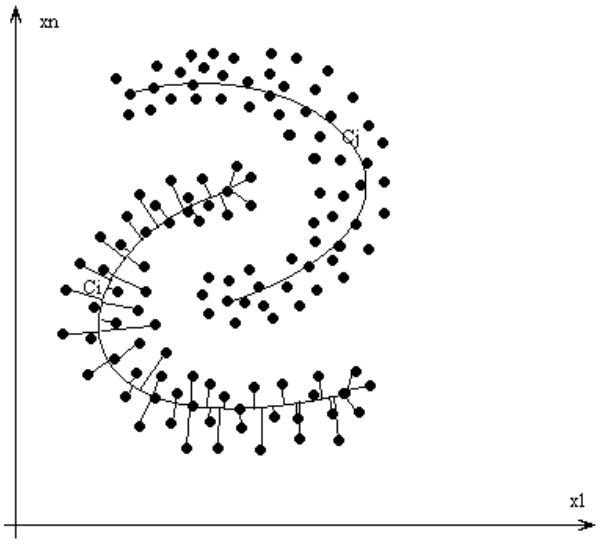

Fig. 4. Interpretation of the fitness-function for the class C^i

The second part of the fitness-function is used so that parametric curves don`t start and end in the extreme points of the cluster. In some cases there are possible situations when the distance from the unknown pattern to the extreme points of the "alien" class is smaller than the distance from this pattern to the internal points of the curve of the "native" class. The coefficients for the first and the second constituent can be chosen experimentally.

In the second case, the search of coefficients of all the parametric curves is realized simultaneously. The general genotype is formed from all the coefficients of all the

parametric equations of the system. The classification of all the samples of the training set and calculation of the number of mistakes is held to get the value of the fitness-function:

$$F = \sum_{i=1}^{M} Ind(X_i) \to \min, \tag{9}$$

where $Ind(X_i) = \begin{cases} 1, & X_i \in C^j = false, \\ 0, & X_i \in C^j = true. \end{cases}$

V. Classification of the Unknown Pattern with the use of Parametric Equation of Curves for Clusters Authentification

As each cluster is in accordance with the curve, we calculate the distance from to the vector of signs to the curves which describe each cluster (fig.5):

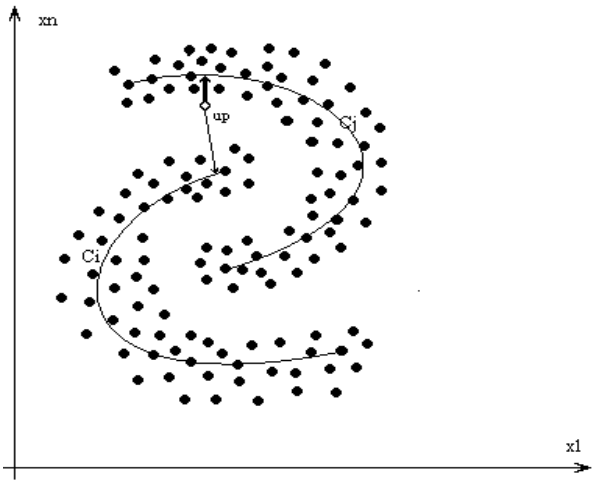

Fig.5. Estimation of the distance from the unknown pattern to each cluster (***up***-unknown pattern).

Identification of the pattern X:

$$m = \arg\min_i \ d_2(X, L^i), \tag{10}$$

$L^i = \{L^i_1, L^i_2, ...L^i_m\}$– set of points on the curve that defines the cluster i, where

$$\begin{aligned} &L^i_j = \{x^i_1(t_j), x^i_2(t_j), ..., x^i_n(t_j)\}, \\ &t_j = j * h, \quad h = 1/m, \\ &j = \overline{0, m}. \end{aligned} \tag{11}$$

So, the distance from the unknown point to the curve is estimated by the distance from the nearest control point of this curve.

VI. General Evaluation of the Method

Because the set of "etalon patterns" of each class is used to identify an unknown pattern, the efficiency of the described method exceeds the efficiency of the recognition methods based on comparison with the etalon or discriminant functions of the linear, quadratic and ellipsoidal form. Parametric equations of the third order have limited possibilities for curves simulation, therefore the offered algorithm is inferior to the Group method of data handling (Ivakhnenko method [7]). But compared to GMDH, the described algorithm has much simpler implementation.

VII. Conclusion

There is presented the method of recognition of patterns in cases when in the space of signs clusters have the complicated form and cannot be localized by the convex figures. Each cluster has the cubic parametric curve in accordance. The distance from the vector of signs of the unknown pattern to each curve is estimated for classification. The class is considered to be recognized if the estimation of distance to the unknown point is the smallest.

The given method cannot be considered universal, but it allows performing the effective analysis of patterns in clusters that have the topological form of deformed ellipsoids, cylinders, tori, etc.

References

[1] V.Fratavchan, I.Gaidaichuk, M.Rusnak, Matrix and Grammatical Methods for Pattern Recognition //Proceedings of the International Conference on Development and Application Systems, DAS 1994 (26-28 May 1994, Suceava - Romania), pp.203-208.

[2] V.Fratavchan, Using Neural Networks For Geometrical Shapes Recognition//Proceedings of the 7th International Conference on Development and Application Systems, DAS 2004 (27-29 May 2004, Suceava - Romania), pp.484-486.

[3] Фомин Я. А. Распознавание образов: теория и применения. — 2-е изд. — М.: ФАЗИС, 2012. — 429 с. — ISBN 978-5-7036-0130-4.

[4] Роберт Седжвик. Алгоритмы на C++. Фундаментальные алгоритмы и структуры данных. 2 книги в одной ! = Algorithms in C++. — М.: «Вильямс», 2011. — 1056 с. — ISBN 978-5-8459-1650-1.

[5] V.Fratavchan, The Application of Genetic Algorithm for Training "Without a Teacher", //Proceedings of the 10th International Conference on Development and Application Systems, DAS 2010 (27-29 May 2010, Suceava - Romania), pp.105-107.

[6] Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы = Sieci neuronowe, algorytmy genetyczne i systemy rozmyte. — 2-е изд. — М: Горячая линия-Телеком, 2008. — 452 с. — ISBN 5-93517-103-1.

[7] H.R. Madala, A.G. Ivakhnenko. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, Boca Raton, 1994.