Recognition Algorithms Based on the Cluster Structures

V.V. Krasnoproshin [1], V.G. Rodchenko [2]

1) Belarusian State University, 220030 Minsk, Nezavisimosti av. 4, krasnoproshin@bsu.by

2) Yanka Kupala State University of Grodno, 230023 Grodno, Ozheshko st. 22, rovar@grsu.by

Abstract: *On the basis using of cluster structures can be automatically carried out analysis of data from the training sample and then can construct decision rules to address recognition problems. The cluster structure is specially constructed geometric body, with using which is formally described the class pattern in recognition algorithms. The problem of construction cluster structures and their application in recognition algorithms is considered. Results of numerical experiments are presented.*

***Keywords*:** Pattern Recognition, Cluster Structure, Training Sample.

1. INTRODUCTION

Currently, one of the most important in practical terms of computer problems is to extract useful knowledge from previously accumulated in electronic data form.

The use of cluster structures allows realizing the process of finding and formal presentation of class patterns and hidden knowledge. Construction of class patterns and search for hidden knowledge is done automatically based on the analysis of training sample data.

The training sample contains information about each of the classes in the form of a set of objects [1, 2]. On the basis of the corresponding set of objects is proposed to construct class patterns in the form of cluster structures.

If cluster structures do not overlap one another, the training procedure was successful. This means that class patterns are located compactly and do not overlap in feature space. This is the perfect situation.

The reality may be the intersection of cluster structures. It must not exceed a permissible threshold. To get an estimate of the intersection of cluster structures is necessary to know the value of their volumes.

In general, the cluster structure is a geometric body in a space of dimension ***n*** [3]. The calculation of volume the cluster structure can be realized through the use of the Monte Carlo method.

The article investigates the issues related to the accuracy and speed of calculating the volume of the geometric bodies in high dimensions.

2. CLUSTER STRUCTURE CONSTRUCTION

Let there be a set of objects, which is composed of representatives of a certain class. Each object is a column vector species formally described by $y^T=(y_1,y_2,\ldots,y_n)$, where $y_i \in R$ – the value of the i-th feature. Combining objects from all classes of training sample sets, which can be formally written in the form of a matrix $Z_{n\ x\ m}$, where $m=m_1+\ldots+m_k$, and m_i – the number of objects of the i-th class, k – the number of classes.

The pattern of each class is invited to build on the basis of the relevant objects in the form of multi-dimensional spatial structures. Normalize the training sample $Y_{n\ x\ m}$ (where $m=m_1+\ldots+m_k$, and m_i the number of objects of the i-th class) and obtain $X_{n\ x\ m}$, where $x_{ij}=(y_{ij} - y_{min}) / (y_{max} - y_{min})$ and combine all vector of i-th class in a separate matrix of the form:

$$X^i_{n\ x\ m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \ldots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \ldots & x^i_{2m_i} \\ \ldots & \ldots & \ldots & \ldots \\ x^i_{n1} & x^i_{n2} & \ldots & x^i_{nm_i} \end{pmatrix}, \text{ where } i=\overline{1,k}\ ;\ j=\overline{1,m_i}\ ,$$

The class object in space R^n is described by the coordinates of the vertices vector $(x_1,x_2,\ldots,x_n)$, where $x_i \in R$ - value of i-th feature.

Building a spatial structure begins with the construction of the skeleton in the form of a minimum spanning tree graph. Construction of the skeleton is carried out through the use of a plurality of vertices of the vectors i-th class. You can use the Prim's or Kruskal's algorithm.

The result is a skeleton that contains m_i vertices and $C^2_{m_i}$ edges. The weight of each edge is its length. We assume that the thus constructed structure determines in n-dimensional feature space formalized pattern class.

For each vertex edges and the midpoint edges define a region of space centered at this point and limited hypersphere of radius r, where the radius is equal to half the length of the rib. As a result, based on the volume of each fin element is formed, and the entire structure becomes volumetric space.

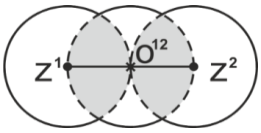

Fig.1 – The base element of the cluster structure.

The proposed approach will allow obtaining numerical estimates of mutual accommodation formalized classes of patterns in a multidimensional space of decision-making.

The process of building the spatial structure is completed when all of its basic elements are combined. The structure, built in the actions described above will be called cluster structure.

Here is the definition of cluster structure, the use of which will allow building practically useful models of information processing.

The cluster structure ***Cs*** can formally ask the Quartet <***M, Z, G, A***>, where ***M*** - linear space of dimension ***n***, ***Z*** - a non-empty set of elements of an arbitrary nature, called vectors, the ***G*** - minimum spanning tree constructed on the set ***Z***, ***A*** - the set algorithms for constructing ***n***-dimensional geometric bodies based on the edges of the minimum spanning tree ***G***.

The cluster structure ***Cs*** = <***M, Z, G, A***> is the union of overlapping linear hypersphere in n-dimensional space:

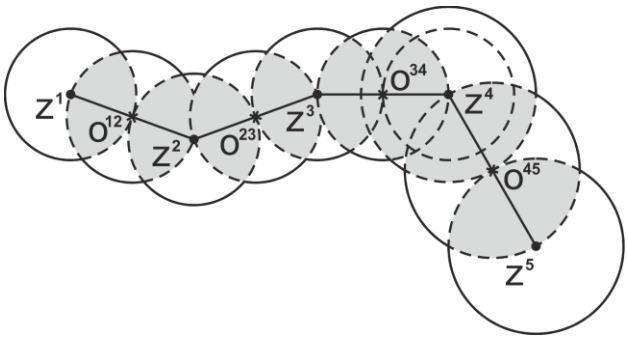

Fig.2 – Example of a cluster structure.

The cluster structure has a dimension of the space in which it is built. It is characterized by: the number of core nodes, volume, a mean density value.

Formally cluster structure constructed on the basis of ***k*** vectors can be represented in tabular form. The first column of the table contains the serial numbers of the hypersphere, the second - the coordinates of the hypersphere centres, and the third - the values of the radii of the hypersphere.

Table 1. The tabular form to describe the cluster structure

№	Centre	Radius
1	z^1	r^1
2	o^{12}	r^1
3	z^2	$\max(r^1, r^2)$
….	….	….
2k-3	z^{k-1}	$\max(r^{k-1}, r^{k-2})$
2k-2	o^{k-1k}	r^{k-1}
2k-1	z^k	r^{k-1}

For each vertex z^2,..., z^{k-1} in the table recorded the maximum value of the radius of the respective two intersecting hypersphere.

3. CLUSTER STRUCTURES EVALUATION

Let us make a numerical experiment to reveal the pattern of influence of the space dimension and the number of tests on the accuracy of calculating the volume of a hypersphere.

In Table 2 shows the results of an experiment on the calculation by Monte Carlo method of hyperspheres volumes in the space of dimension ***n***.

Table 2. Results of numerical experiment (calculation volume of a hypersphere /radius = 1/)

Quantity random points	***n***=2		***n***=3		***n***=4	
	avg (%)	max (%)	avg (%)	max (%)	avg (%)	max (%)
25x25	0.053	0.47	0.079	1.86	0.173	3.23
50x50	0.061	1.10	0.052	0.46	0.167	1.82
75x75	0.026	0.20	0.051	0.29	0.125	2.14
100x100	0.050	0.64	0.066	0.87	0.159	1.40

In Table 2: avg - average value of the error in percent of 50 tests, max – maximum value of the error in percent of 50 tests.

Based on these experimental results, we can conclude that the calculation error is acceptable, if the dependence of the number of tests on the space dimension calculated by the formula: ***Quantity random points*** $\approx (75/(n-1))^n$.

In order to calculate the volume of the cluster structure we use the Monte Carlo method. First, consider the basic element of the cluster structure, which would be as follows:

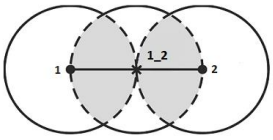

Fig.3 – The base element of the flat cluster structure.

Calculating the volume (area) of the cluster structure can be carried out through the use of analytical formulas, as well as using the Monte Carlo method.

For this cluster structure we assume that the radius is equal to one. Obviously, it is possible to analytically calculate the area of the cluster structure using the appropriate formula to calculate the circle area, and the formula for calculating the area of intersection.

The following table shows the results of numerical simulations to calculate the area of the base member.

Table 3. Results of calculation for base element of the flat cluster structure by Monte Carlo method

Quantity random points	Computing time (seconds)		The average error at 1000 tests (%)
	Formula	Monte Carlo	
1 000	0	0	0.950
5 000	0	0.001	0.440
10 000	0	0.001	0.304
20 000	0	0.002	0.218
50 000	0	0.005	0.139
100 000	0	0.011	0.097
500 000	0	0.055	0.044
1 000 000	0	0.111	0.030

In the first column of the table shows the number of random points in the Monte Carlo method. The second column contains the time in seconds for calculation using an analytical formula. The third column contains the calculation time in seconds using the Monte Carlo method. The fourth column shows the average error value at 1000 trials.

From Table 3 shows that the error in calculating the area of a flat cluster structure of the Monte Carlo method is an acceptable at 5,000 tests. In this case, the average value of the error test at 1000 was about 0.45 percent.

Let us make a numerical experiment to calculate the volume of the cluster structure and to obtain an estimate of computation time (in seconds).

Table 4 shows the results of calculations for cluster structure, the skeleton of which contains 10 vertices. Length skeleton ribs formed at random in the range of (0.0, 0.1). The cluster structure has been placed inside the unit hypercube of dimension ***n***.

In the first column of the Table 4 is determined the dimension of the space in which the cluster structure is located. From Table 4 shows that column 2 contains quantity of random points in Monte Carlo method, in column 3 - volume of cluster structure, in column 4 – time for calculation volume of cluster structure.

Table 4. Results of numerical experiment (the skeleton of the cluster structure contains 10 vertices)

n	Quantity	Volume	Time

	random points		(seconds)
2	5 000	0.0683494	0.01
	10 000	0.0676744	0.01
	50 000	0.0676744	0.101
	100 000	0.0677313	0.180
	500 000	0.067712	0.86
	1 000 000	0.067844	1.74
3	5 000	0.0071154	0.02
	10 000	0.0073235	0.03
	50 000	0.0072581	0.142
	100 000	0.0071972	0.27
	500 000	0.0072773	1.346
	1 000 000	0.0073028	2.672
4	5 000	0.0007152	0.02
	10 000	0.0007131	0.04
	50 000	0.0007434	0.19
	100 000	0.0007610	0.35
	500 000	0.0007574	1.804
	1 000 000	0.0007601	3.854
5	5 000	0.0001092	0.021
	10 000	0.0000974	0.04
	50 000	0.000098	0.21
	100 000	0.0001098	0.43
	500 000	0.0001098	2.138
	1 000 000	0.0001104	4.248

Table 5 shows the results of calculations for cluster structure, the skeleton of which contains 50 vertices. Length skeleton ribs formed at random in the range of (0.0, 0.02). The cluster structure has been placed inside the unit hypercube of dimension ***n***.

Table 5. Results of numerical experiment (the skeleton of the cluster structure contains 50 vertices)

n	Quantity random points	Volume	Time (seconds)
2	5 000	0.021286	0.01
	10 000	0.020681	0.12
	50 000	0.021002	0.56
	100 000	0.020725	1.13
	500 000	0.020764	5.63
	1 000 000	0.020649	11.19
3	5 000	0.0004264	0.122
	10 000	0.0004145	0.16
	50 000	0.0004548	0.782
	100 000	0.0005081	1.604
	500 000	0.0005014	8.139
	1 000 000	0.0004989	16.134
4	5 000	0 !!!	0.1
	10 000	0.00000544	0.19
	50 000	0.00000871	0.96
	100 000	0.00001089	2.05
	500 000	0.00001132	9.64
	1 000 000	0.00001143	19.46
	5 000 000	0.00001131	97.17
5	5 000	0 !!!	0.112
	10 000	0 !!!	0.24
	50 000	0 !!!	1.202
	100 000	0.000000277	2.53
	500 000	0.000000332	11.93
	1 000 000	0.000000249	25.39
	5 000 000	0.000000266	119.38
	10 000 000	0.000000260	251.07

Analysis of the data of Tables 4 and 5 shows that, firstly, to calculate the volume of cluster structures by Monte Carlo the quantity random points can define by function: $f(n) \approx (75 / (n-1))^n$, where ***n*** - space dimension.

Second, with increasing ***n*** (space dimension) the computation time increases practically exponentially.

4. CONCLUSION

The use of cluster structures allows realizing the training procedure in recognition algorithms automatically.

The article examines the problem of using the Monte Carlo method to calculate the volume of cluster structures. Estimations of error when using the Monte Carlo method to calculate the volume of cluster structures were obtained.

Based on the results of numerical experiments, the formula for calculating the number of random points in Monte Carlo method has been proposed.

It is demonstrated that the calculation time significantly increases with the dimension of the space, and so urgent is the search for new approaches for evaluating the mutual placement of cluster structures.

5. REFERENCES

[1] V.I.Vasil'ev. *The problem of training for pattern recognition*. Visha shkola. Golovnoe izdatel'stvo. Kiev, 1989. p. 64. (in Russian)

[2] N.G. Zagoruiko. *Applied methods of data analysis and knowledge*. Izdatel'stvo Instituta matematiki SO RAN. Novosibirsk, 1999. p. 268. (in Russian)

[3] V.V. Krasnoproshin, V.G. Rodchenko. Cluster Structures and Their Applications in Data Mining, *Informatics №2 (2016)*, p. 71-77 (in Russian)