Representation Method of Formal Pattern Classes in Cluster Mode at Building of Recognition Systems

Rodchenko V.G. $^{1)}$, Rodchenko V.V. $^{2)}$, Zhukevich A.I. $^{3)}$
1) Rodchenko Vadim, Yanka Kupala State University of Grodno;
230023, Grodno, Ozheshko str., 22 - 212; rovar@mail.ru
2) Rodchenko Vladimir, Yanka Kupala State University of Grodno;
230021, Grodno, Malyshchinskaya st., 23 - 21; r-vladimir@mail.ru
3) Zhukevich Aleksandr, Yanka Kupala State University of Grodno;
230000, Grodno, Strelkovaya str., 19 - 25; san@grsu.by

Abstract: *At building of recognition systems there is a problem of formal representation of pattern classes or standard pattern classes in multidimensional feature space. In the article it is convenient to represent pattern classes in cluster mode and the original algorithm for building such clusters is described.*

Keywords: alphabet classes, feature vocabulary, classified training sample, cluster, pattern recognition.

1. INTRODUCTION

Application of the mathematical pattern recognition theory apparatus often appears as the most effective tool for studying of the phenomena and objects of the difficult nature which are characterized by quite great number of various by the nature features.

At building of pattern recognition systems it is necessary to execute two basic procedures that include training procedure and decision-making procedure (or control procedure) [1,2]. If training procedure manages to be realized qualitatively, the performance of the second one has just technical character [3].

Training procedure is the most labor-consuming from the point of view of realization, as in real systems are examined objects which are characterized usually by a considerable quantity of various features which have the difficult nature and are submitted to different laws [4,5].

Training procedure is carried out on the basis of the initial data which are primarily represented in the form of the classified training sample. As a result of this procedure performance the consistent pattern of class samples to corresponding classes should be established. Received consistent pattern is used further for building of a solving rule on which basis final procedure of decision-making is carried out, and object under consideration belongs to one of the initial classes, or is allocated in a separate independent distribution joker-class.

Building of classified training sample is carried out on the basis of a priori feature vocabulary (AFV), which represents sample of the general feature vocabulary corresponding to subject domain.

Ideally the a priori vocabulary contains only such features which provide division of classes in multidimensional feature space. However in real tasks an a priori generated sample does not guarantee full objectivity, as it is formed taking into account restrictions appropriated to available resources.

Realization experience of recognition systems testifies that it is seldom possible to define a priori a corresponding feature set, that is why in AFV appear features which do not accomplish dividing classes functions so creating "noise waves" at recognition and essentially worsening its quality [6].

For the purpose of maintenance of more authentic performance of recognition procedure in systems realization of research stage which is connected with the analysis of the training sample data is provided. As a result of this analysis the specified working feature vocabulary (WFV) on which basis such solution space, in which class standards will be compact and divided, will be formed and will be under construction [7].

After building of the working feature vocabulary the stage of reliability certification of recognition procedure in generated on the feature basis from WFV solution space is necessarily carried out.

In the article the representation method of formal pattern classes is suggested at recognition systems realization which is based on use of original algorithm of pattern class representation in form of clusters in multidimensional solution space.

2. PROBLEM DESCRIPTION

For building of recognition systems methods which are based on building of dividing classes surfaces in multidimensional feature space are quite often used. It is supposed that corresponding surfaces can be built. Corresponding assumptions are proved by that the feature vocabulary on which basis the description of objects under consideration is made, includes such features which provide division of pattern classes in corresponding feature space.

However at practical use of the mathematical pattern recognition theory apparatus absolutely other tendency, as such features, which do not accomplish dividing classes functions but only create “noise” (“noise waves”) at performance of final decision-making procedure, really get to an initial variant of the a priori vocabulary, is steadily shown. In this case a priori vocabulary features do not provide performance of compactness hypothesis, so, use of these features for performance of decision-making procedure will potentially lead to serious distortions and erroneous results.

For high-grade carrying out of all recognition process it is necessary to provide performance of an additional self-descriptiveness analysis stage of each of a priori feature vocabulary. As a result it is necessary to generate such feature vocabulary which would contain only the features providing division of all pattern classes in pairs among themselves in multidimensional feature solution space. In this case standard pattern classes in feature space will represent separately placed not crossed clusters, and the condition connected with performance of compactness hypothesis will be provided thus.

At building of recognition system the presence of the initial classified training sample gives possibility to realize training procedure by carrying out of the data analysis from this sample. Feature separation on self-descriptiveness degree from the point of view of division of pattern classes in multidimensional feature space is carried out on the basis of comparative data analysis from the classified training sample.

The specified sample is formed as a result of association of all vectors formally describing standard patterns of all classes.

Let us imagine there exist an alphabet class A={A_1,A_2,...,A_k} and is formed an a priori feature vocabulary P={P_1,P_2,...,P_n}. Every object is described by n features from a priori feature vocabulary in form of vector-column X^T=(x_1,x_2,…,x_n), where x_i –i-feature value, is unequivocally identified with one of the classes, and all samples of values of each class features have continuous distribution functions. The object set of a separate class forms the initial description of this class in a priori feature space, and association of all objects from all classes represents the classified training sample. This sample is described in the form of the table of type "object-feature" and formally is represented in the form of matrix $X_{n \text{ x } m}$, where m=m_1+m_2 +... +m_k, and m_i - quantity of i class objects.

At building of real systems on the basis of use of mathematical pattern recognition theory apparatus there is a necessity of formal representation of pattern classes in corresponding multidimensional feature space. Such space is built on the basis of a priory feature vocabulary, or on the basis of the specified feature vocabulary.

For formal representation of pattern class in multidimensional feature space building of corresponding cluster on the basis of all class standards is suggested. Each class standard represents a vector in space R^n with top coordinates (x_1, x_2, …, x_n), where x_i – i feature value. Association of all vectors of one class in cluster also will represent the formal description of a class in corresponding multidimensional feature space.

Cluster building process begins with search of such example class which is the most removed from others. Then the nearest to a found example representative of a class is defined, and it joins the cluster structure as a following element. Besides, are calculated and if necessary are remembered coordinates values of an auxiliary vector which indicates the midpoint, connecting two next class examples. As a result cluster "skeleton" for building of pattern class is formed, which contains 2*m_i-1 vectors, from which m_i vectors-examples of i class (where m_i - quantity of objects of i class) and m_i-1 auxiliary vectors.

Further each example of the "skeleton" is represented as the hypersphere center at cluster formation. As a result cluster appears as association of areas which are formed by crossed hyperspheres.

3. DESCRIPTION OF METHOD REALISATION

ALGORITHM

Let to a priori feature vocabulary be formed P={P_1,P_2,...,P_n}, and every class example is described on the basis of n features from this vocabulary. Formally each such example is represented in the form of vector-column X^T = (x_1, x_2, …, x_n), where x_i – i feature value. Association of all vectors-columns will represent a dimension matrix n x m, where m - quantity of class examples.

We will begin cluster building with search of the most remote from all class examples. We will designate it $X^{(1)}$. The given example will be the first element of formed cluster.

Then we will find for $X^{(1)}$ the nearest example $X^{(2)}$, the distance between them will be designated as $l^{(1)}$. We will build hyperspheres (further spheres) of radius $r^{(1)}$, where

$$r^{(1)} = l^{(1)}/2\,, \qquad (1)$$

with centers in $X^{(1)}$, $X^{(2)}$. We will designate the contact point of two spheres – $O^{(1)}$ with coordinates ($o_1^{(1)}$,...,$o_n^{(1)}$) and we will build radius sphere $r^{(1)}$ with center in $O^{(1)}$. Further for the example $X^{(2)}$ we will find the nearest example $X^{(3)}$ and distance between them will be designated as $l^{(2)}$, and is excluded $X^{(1)}$. We will build radius spheres $r^{(2)}$, where

$$r^{(2)} = l^{(2)}/2\,, \qquad (2)$$

with centers in $X^{(2)}$, $X^{(3)}$ and we will get the sphere contact point $O^{(2)}$=($o_1^{(2)}$,...,$o_n^{(2)}$). We will build radius sphere $r^{(2)}$ with center $O^{(2)}$. The example $X^{(2)}$ is a center of two radius spheres $r^{(1)}$ and $r^{(2)}$, the maximum radius is chosen for the cluster description.

Let us similarly continue building of corresponding spheres for other class examples. We will present the result of such building by means of the table 1 in which the first column contains serial numbers of the spheres, the second column includes coordinates of sphere center, and in the third column there is a radius value. The column “Feature” accepts value 1 – if the sphere center is $X^{(i)}$ and 0 – if intersection point $O^{(i)}$ appears as the sphere center.

Table 1. Cluster description

№	Center	Radius	Feature
1	$X^{(1)}$	$r^{(1)}$	1
2	$O^{(1)}$	$r^{(1)}$	0
3	$X^{(2)}$	$r^{(2)}$	1
….	….	….	…
2m-2	$O^{(m-1)}$	$r^{(m-1)}$	0
2m-1	$X^{(m)}$	$r^{(m-1)}$	1

The examples $X^{(1)}$ и $X^{(m)}$ are “extreme”, that is why for them in the table we write down radiuses $r^{(1)}$ and $r^{(m-1)}$. For each example of the class $X^{(2)}$..., $X^{(m-1)}$ the maximum value of the radius is written down in the table, connected with building of the corresponding spheres.

As a result we received that the area uniting all spheres will represent cluster which is subject of a formal pattern class in multidimensional feature space.

For computer realization of recognition system it is enough to provide storage of the formal description cluster in the form of the Table 1 containing only class examples, i.e. all the lines with the column "Feature" value equal 1.

For cluster characteristics estimation the possibility of cluster volume and density calculation is provided.

5. CONCLUSIONS

Originally formal definition of each separate class example represents a vector in multidimensional feature space. In turn, it is possible to present the initial formal class description in the form of a matrix received by association of corresponding vectors.

In the article the method which allows presenting a pattern class in cluster mode, received by association of hyperspheres in multidimensional feature space, is described.

The suggested method can be used at designing and building of recognition systems for procedure performance of the analysis of mutual placing of class patterns and standards in multidimensional feature space.

6. REFERENCES

[1] Yu.I. Zhuravlev. *Raspoznavanie obrazov i raspoznavanie izobrazhenij / Yu.I. Zhuravlev, I.B. Gurevich // Raspoznavanie, klassifikatsiya, prognoz. Matematicheskie metody i ikh primenenie. Vyp.2*. Nauka. Moskva, 1989. p. 302.

[2] V.I. Vasil'ev. *Problema obucheniya raspoznavaniyu obrazov / V.I. Vasil'ev*. Visha shk. Golovnoe izd-vo. Kiev, 1989. p. 64.

[3] N.G. Zagoruiko. *Prikladnye metody analiza dannykh i znaniy*. Izd-vo Instituta matematiki SO RAN. Novosibirsk, 1999. p. 268.

[4] S.A. Aivazian. *Prikladnaya statistika i osnovy ekonometriki / S.A. Aivazian, V.S. Mkhitarian*. YuNITI, Moskva, 1998. p. 1022.

[5] V.N. Vapnik. *Teoriya raspoznavaniya obrazov (statisticheskie problemy obucheniya) / V.N. Vapnik.* Nauka. Moskva, 1974. p. 415

[6] V.G. Rodchenko. The method of realization of stylometric studies on the basis of applying the apparatus of mathematical theory of pattern recognition, *Proceeding of the F.Skorina Gomel State University* 5(44) (2007). p. 58-62

[7] V.G. Rodchenko. On one method of construction of the compact patterns of classes at designing of pattern recognition systems, *Proceeding of the F.Skorina Gomel State University* 4(25) (2004). p. 114-117

[8] V.G. Rodchenko. A.I. Zhukevich. About one method of formal class pattern construction at recognition system realization, *Proceeding of the F.Skorina Gomel State University* 5(62) (2010). p. 79-83