INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 5, Issue - 8, Aug – 2019
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 20/08/2019 Acceptance Date: 30/08/2019 Publication Date: 31/08/2019

Algorithm based on sorting formed standard tables for recognition of patterns

O. K. Akhmedov
Indepedent Researcher
Scientific-innovation center
Tashkent University of Information Technologies, Tashkent, Uzbekistan
Email - o_b.brother@mail.ru

Abstract: *The article proposes some algorithms for the formation of the reference tables, which is the main task of determining symbols. Before creating a reference table, it is necessary to develop an training sample. The process of forming a reference table by removing unnecessary objects on the basis of an training sample is described. The theory and practice of determining (decoding) symbols is important in solving such tasks as studying, evaluating, classifying and analyzing the state of research objects related to production, the education system, management, social and economic activities of human society.*

Key Words: *symbols, reference table, algorithms, learning set, informative features, software, vectors, object of study.*

1. INTRODUCTION:

The theory and practice of determining (decoding) symbols is important in solving problems such as studying, evaluating, classifying and analyzing the state of objects of research related to production, the education system, management, and the socio-economic activity of human society.

The construction of algorithms for the formation of a reference table is an important task in pattern recognition.

Suppose that an X-training sample is specified, where the elements of the object are 0 and 1, the number of columns is N, and the number of rows is M.

Creation of a simple method and algorithm for solving the problem of selecting X-objects of a training set, in a segment of a set of dissimilar features, into subclasses that are closely related to each other and to subclasses that are not connected within the classes.

The task in the literature [1,2] is called clustering, taxonomy and classification, and there are many methods and algorithms for solving this problem.

2. ANALYSIS:

The article proposes a joint solution to the following tasks:

1. Formation of a reference table based on a training sample;
2. The choice of a set of informative features using reference tables;
3. Development of a method, algorithm and software for solving classification problems.

Suppose that a training set is given in the following form:

$$x_{11},...,x_{1m_1} \in X_1, x_{21},...,x_{2m_2} \in X_2,....., x_{r1},...,x_{rm_r} \in X_r$$

where, each object - x_{pi} is defined by a vector quantity with N - dimension, i.e. $x_{pi} = (x_{pi}^1,...,x_{pi}^N)$. The sign index - pi denotes the i-th class object - p. Also, N is the number of features of the object.

Let for given objects of the training set:

$$x_{p1}, x_{p2},...,x_{pm_p} \in X_p, p = \overline{1,r},$$

i.e., x_{pi} is specified in the feature space with N - dimension, the value of their components is 0 and 1, and objects of all classes are specified sequentially.

It is necessary to group the given objects according to the signs of similarity (identity) and not similarity, proximity and remoteness from each other.

In this case, the compilation of the algorithm for the formation of the reference table based on the training set is described as follows.

Usually, in order to distinguish each of M objects with N features, the initial values are entered, expressing N features of each of M objects with values 0 and 1;

1. Given objects are divided into groups according to the values of the characteristics.
2. Each object in the group is checked for belonging to this group, if the object does not belong to the group, then it is checked for belonging to the next group. If the object belongs to the checked group, it will be deleted from its group and added to the checked group. If it does not belong to this group, it will be checked for belonging to the next group, etc. If the object does not belong to any group, the object will be removed from the list of sets.
3. To check whether an object belongs to its group, it is necessary to calculate the hemming distances between all objects of the same group and determine their average value. Similarly, the average values of the hamming distance between objects in the remaining groups are calculated. If the average value of the hemming distance between objects in its group is less than the average in other groups, then the object is considered to belong to its group. Otherwise, the object does not belong to this group.
4. The set of informative features in the program is selected as follows: First of all, 1 feature is excluded from N features, then in the segment of the remaining features the objects belong to their group. If each object remains belonging to its own group, then this attribute is excluded, otherwise it will be left in its place and the next attribute from the set of attributes will be deleted, and based on the remaining attributes, the object will be checked for belonging to its group, etc. As a result of this process, a transition is made from the N feature set to the N-1 feature set.

   5. The above process will be performed many times, repeatedly if it is possible to switch from the N feature set to the N - 1 set, and the process will be stopped if this is not possible.

   As a result, a transition is made from the N set of attributes to the set ℓ = N - k. That is, from the feature space of dimension N go to the feature space of dimension ℓ. This is ℓ - a set of informative features.

At the end of the process, ℓ is determined - a set of informative features that indicates which group each object belongs to and which objects belong to this particular group.

It should be noted that there are many other methods of the extraction ℓ algorithm - a set of informative features described above.

Here are some of them. The first method is the full choice method. The main disadvantage of this method is the performance of large amounts of computation. It is known that for its implementation it is necessary to calculate $C_N^\ell = \frac{N!}{\ell!(N-\ell)!}$ - once the value of the criterion I (λ) and from them to select the one that assigns the value of the extermum to the functional. If we define a set of informative features by this full-sampling method, then the number of all calculations will be - $2C_N^\ell - 1$.

Based on the foregoing, we can say that it is advisable to use this method with a minimum (not large) value of N.

If the number of features is large, it is recommended to use suboptimal methods based on partial sampling. Although these methods cannot select the ideal set of informative features, it at least guarantees that the worst set of informative features will not be selected.

Let's look at the easiest way to select a set of informative features from the space of initial features. In this method, the procedure for selecting a set of informative features is as follows: firstly, the level of information content of each feature is evaluated; secondly, signs are placed in descending order of information level, and then ℓ signs from the beginning are highlighted.

The number of all considered options is equal $C_N^1 = N$, the number of calculations, i.e. the number of samples ℓ of informative features is equal $N + \frac{N(N-1)}{2}$. Such an approach to the selection of a set of informative features guarantees an ideal solution only when the features do not have a statistical relationship [1,2,3]. However, in most cases, in practice, the symptoms are statistically interdependent.

Below we use the vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ to select informative features. Here $\lambda_i \ (i = \overline{1, N})$ takes the value 0 or 1 and defines this attribute as belonging to or not belonging to a set of informative attributes.

In [1,2], the algorithm "sequential inverse sampling" (SIS) was proposed. The order of execution of this algorithm is as follows:

$$\sum_{j=1}^{N} \lambda_j = N - 1$$

At the beginning, λ vectors are taken for which the value I (λ) is calculated using [3] and a vector λwith the maximum value is selected, then the attribute corresponding to the zero component from the attribute space is excluded. Thus, a transition is made from the feature space with dimension N to the feature space with dimension N-1.

The same process continues second, third, etc. times, until the size of the feature space is equal to ℓ.

In this method, the number of calculations of the value of the functional I (λ) is equal to $\frac{(N+\ell+1)(N-2)}{2}$, and for a large value of N it is much less than C_N^ℓ, i.e. the amount of computation is much less. However, as shown in [1,2], this algorithm does not guarantee the integrity of the result.

Although this algorithm differs in form from the aforementioned algorithm, it is identical in content.

3. CONCLUSION:

Thus, the article proposes an algorithm for the formation of reference tables for pattern recognition.

The proposed software allows you to determine the degree of significance of the objects of the training sample and in accordance with this, using the sorting algorithm, a reference table is determined and ℓ is a set of informative features.

The proposed software has successfully passed experimental testing in the training sample of the iris flower.

REFERENCES:

1. 1. Cheponis K.A., Zhvirenayte D.A., Busygin B.S., Miroshenichenko L.V. Methods, criteria and algorithms used in the conversion, selection and selection of features in data analysis // Col. articles. - Vilnius, 1988 .-- 149s.
2. 2. Kutin G.I. Ranking methods for feature complexes. Overview // Foreign Radio Electronics, 1981, No. 9. S. 54-70.
3. 3. Lbov G.S. Methods for processing heterogeneous experimental data. - Novosibirsk: Nauka, 1981. - 160 p.