Recent Trends in Discrete Analysis of Information in Recognition Problems[1]

E. V. Djukova*, A. S. Inyakin, and N. V. Peskov*****

**Computer Center, Russian Academy of Sciences,
ul. Vavilova 40, Moscow, GSP-1, 119991 Russia
e-mail: djukova@ccas.ru*

***Moscow State Pedagogical University,
ul. Pirogovskaya 1, Moscow, GSP-2, 119992 Russia
e-mail: andre_w@mail.ru*

****Scientific Council on Cybernetics, Russian Academy of Sciences,
ul. Vavilova 40, Moscow, GSP-1, 119991 Russia
e-mail: nick@motor.ru*

Abstract—This paper is a review of current trends of research in the field of discrete analysis of object feature descriptions for recognition and classification procedures of large dimensionality. Several new models of discrete recognition procedures are described alongside new approaches to the search for informative fragments in feature descriptions of objects. The metric (quantitative) properties of elementary classifiers are examined.

INTRODUCTION

The discrete approach to recognition problems relies on the combinatorial analysis of feature descriptions of objects. Fundamental here are the works of Yu. I. Zhuravlev, S.V. Yablonski, and M.N. Vainzvaig. The main advantage of the discrete (logical) recognition procedures is the possibility of obtaining a result when there is no information about distribution functions and when the learning samples are small. However, the use of discrete mathematics is often hampered by the computational difficulties associated with an exhaustive search. These difficulties arise at the stage of selecting informative fragments in object descriptions. As a rule, the fragments that allow us to discern objects from different classes are considered informative. The exhaustive search and the poor efficiency of early computers were the reasons why, for many years, most efforts have been focused on the development of a general complexity theory in discrete data analysis and design of asymptotically optimal search algorithms for informative fragments. The results obtained in this area made it possible to circumvent most difficulties associated with exhaustive search (see the works of V. Slepyan, V. Noskov, E. Djukova, and A. Andreev).

Fundamental to the entire methodology of discrete analysis of feature descriptions of objects is the concept of irreducible covering of the integer matrix. This is a generalization of the concept, popular in discrete mathematics, of irreducible covering of a Boolean matrix introduced in [1] for improving KORA-type algorithms and for obtaining necessary asymptotic estimates.

In designing discrete procedures, logical functions are often used. The problem of selecting informative fragments in feature descriptions of images is solved via construction of allowable and maximal conjunctions of logical functions. This is done by constructing coverings and irreducible coverings, respectively, of integer matrices of a special form.

An asymptotically optimal algorithm of searching for irreducible coverings of the integer matrix was designed in [1] for the case when the number of rows was less than the number of columns. The asymptotic optimality of this algorithm was substantiated, and the asymptotic values of typical values for the number of irreducible coverings and for the length of irreducible covering were obtained. The methodology of these estimates was first elaborated by Slepyan and Noskov during investigation of the metric (quantitative) properties of a set of irreducible tests. The most comprehensive description of the results can be found in [2].

Recently, new approaches to the design of discrete recognition procedures have arisen.

1. NEW MODELS OF DISCRETE RECOGNITION PROCEDURES AND NEW METHODS OF SEARCHING FOR INFORMATIVE FRAGMENTS IN IMAGE DESCRIPTIONS

We consider a standard statement of the recognition problem in the case where the objects are described by a set $\{x_1, \ldots, x_n\}$ of n integer-valued features [2]. For $j \in \{1, 2, \ldots, n\}$, let N_j denote a set of admissible values of the feature x_j; we assume that the set M of objects under

[1] This work was financially supported by the Russian Foundation for Basic Research, project no. 01-01-00575.

Received October 29, 2002

examination can be represented as a union of subsets (classes) $K_1, \ldots, K_l$. There is a finite set $\{S_1, \ldots, S_m\}$ of objects from M (a training sample), and we know to which classes they belong. The training objects are represented by their descriptions. Given a set of feature values (i.e., a description of some object S from M; generally, it is not known to which class it belongs), it is required to determine the class containing this set.

Suppose that H is a set of r different features of the form $\{x_{j_1}, \ldots, x_{j_r}\}$ and $\sigma = (\sigma_1, \ldots, \sigma_r)$, where $\sigma_i \in N_{j_i}$ for $i = 1, 2, \ldots, r$. We call the set σ an elementary classifier generated by the features from H.

Let $S' = (a_1, \ldots, a_n)$ be an object from the training sample. We denote the fragment $(a_{j_1}, \ldots, a_{j_r})$ of the description of S' by (S', H).

An elementary classifier $(\sigma_1, \ldots, \sigma_r)$ generated by the features from H can have one of the following properties:

(1) each fragment of the form (S', H), where $S' \in K$, coincides with $(\sigma_1, \ldots, \sigma_r)$;

(2) several (not all) fragments of the form (S', H), where $S' \in K$, coincide with $(\sigma_1, \ldots, \sigma_r)$;

(3) none of the fragments of the form (S', H), where $S' \in K$, coincides with $(\sigma_1, \ldots, \sigma_r)$.

The first situation occurs less often; thus, it is not possible to process the sets of feature values with property (1). The essential difference in the informativeness between the other two properties is that property (2) characterizes only some subset of training objects from K, whereas property (3) characterizes all objects from K. Therefore, if it is important to consider class K separately; then, obviously, sets of feature values with property (3) seem to be more informative. In this case, it is more natural to refer object S to the class K if the set of feature values under consideration describes none of the objects in the class K and does not describe the object S.

Let us denote the set of all elementary classifiers generated by the feature sets from $\{x_1, \ldots, x_n\}$ by C. Each recognition algorithm is determined by a subset C^A of the set C. In other words, for each class K, $K \in \{K_1, \ldots, K_l\}$, we construct a subset $C^A(K)$ of the set C and $C^A = \bigcup_{j=1}^{l} C^A(K_j)$.

In classical discrete recognition procedures, the elementary classifiers of class K, which possess property (2) and are not contained in descriptions of other classes, are considered informative. The elementary classifier votes for the membership of object S to the class K if the description of object S contains it. The problem of constructing the set of representative samples for K is reduced to the problem of constructing coverings of the Boolean matrix of descriptions of learning objects from $\bar{K}$ with subsequent searching for certain fragments in K.

In [3, 4], new heuristics are suggested. In the heuristics, the allowable sets of the features' values, which are not contained in all descriptions of the training samples of class K, i.e., they possess property (3) (coverings of class K), are considered to be informative sets for class K. Such sets vote for the membership of the object S to the class K if the description of object S does not contain it. The problem is reduced to the construction of coverings of a matrix of descriptions of objects from K.

The proposed models were tested on real problems in medical prognostics.

2. PRELIMINARY ANALYSIS OF LEARNING INFORMATION AND SELECTION OF TYPICAL OBJECTS IN A CLASS

In [3, 4], we suggested the approaches that efficiently analyze the learning tables in order to detect informationally important regions and estimate parameters that characterize the informativeness of features and representativeness (typical nature) of learning objects and their subdescriptions with respect to their classes. The selection of typical objects was based on the following two approaches: (1) calculation of typical values of features (the objects which contained the most typical values in their descriptions are considered typical) and (2) the use of the cross-validation procedure (the objects which are recognized by the cross-validation procedure using the algorithm of voting over representative samples are considered typical). The construction of the set of the typical objects allows us to partition the training sample into a basic subsample (containing typical objects) and a test subsample (containing the rest of the objects). The former was used to construct the set C^A, and the latter determined the weights of the elements from C^A. To calculate the weights of elementary classifiers from C^A, we use simple heuristics, e.g., the difference between the classified and misclassified objects from the test subsample.

The outlined methods were tested on the problems of analysis of sociology data and on tasks of medical prognostics.

3. METRICAL PROPERTIES OF THE SET OF COVERINGS

In [1, 2], asymptotic values of the number of irredundant coverings and the length of an irredundant covering were obtained for almost all integer $m \times n$ matrices with entries in $\{0, 1, \ldots k-1\}$, $k \geq 2$, subject to the conditions $n \longrightarrow \infty$ and $m^\alpha \leq n \leq k^{m^\beta}$, $\alpha > 1$, $\beta < 1$. For almost all such matrices, the number of irredundant coverings was shown to be asymptotically equal to the number of specific submatrices (hereinafter called σ submatrices) and smaller in order than the number of coverings. Thus, the asymptotically optimal algorithm of searching for irredundant coverings can be constructed. Note that the σ submatrix is an identity sub-

matrix in the special case of $k = 2$. In [3], we considered the opposite case of $n^\alpha \le m \le k^{n^\beta}$, where $\alpha > 1$ and $\beta < 1$. Asymptotic values of the typical number of σ submatrices and the typical order of a σ submatrix were obtained there. Additionally, asymptotic values of the typical number of coverings and the typical length of a covering were obtained in a fairly general case. Comparing these estimates, we showed that the number of σ submatrices is almost always greater in order than the number of irredundant coverings when $n^\alpha \le m \le k^{n^\beta}$ for $\alpha > 1$, $\beta < 1/2$.

4. TAXONOMY PROBLEMS AND IRREDUNDANT COVERINGS OF AN INTEGER MATRIX

In [5], the cluster analysis methods were described. These methods deal with data sets for which one cannot define a distance function. The integer data taxonomy algorithm based on σ coverings of classes was proposed. It was based on the following principles.

Consider the situation where the degree of membership of some object S to a group of objects M is estimated. If the description of S contains the set of feature values that none of the objects from M has in its description, then we can say that a conjunction of S and M violates the inner structure of the set M. By regarding different combinations of feature values that are not contained in descriptions of objects from M, we can evaluate the proximity of object S to set M. Thus, by determining the measure of proximity of the object to the set M, we consider as informative the feature values that are absent from the descriptions of all objects from the class M.

The above reasoning served as a basis for constructing an algorithm which, in a number of cases, improves the results of clustering in simulated and real-life problems as compared to the well-known algorithms based on the principles of the nearest neighbor, farthest neighbor, and central element selection, where the Hemming distance is used as a distance function.

In the proposed algorithm, the problem of estimation of the distance from object S to set M is reduced to the problem of counting irredundant coverings of the integer matrix. The integer matrix consists of feature descriptions of the objects in M. The problem of counting irredundant coverings is solved by constructing irreducible coverings of a Boolean matrix specially designed from the initial matrix.

CONCLUSIONS

This paper deals with the study of discrete recognition and classification procedures. The most important stage here is the search for informative fragments of the feature descriptions of objects. New approaches to this problem are proposed in this work.

(1) The general principles of discrete (logic) recognition procedures are considered, and new models are described which augment the application of the discrete analysis technique in recognition problems.

(2) An approach is proposed to increase the efficiency of recognition algorithms, based on the selection of the learning objects typical for each class.

(3) The results of developing methods for solving cluster analysis problems in the case of integer information represented as covering of integer matrices are presented.

REFERENCES

1. Djukova, E.V., On the Complexity of Implementation of Some Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 1987, vol. 27, no. 1, pp. 114–227.
2. Djukova, E.V. and Zhuravlev, Yu.I, Discrete Analysis of Feature Descriptions in Recognition Problems of High Dimension, *Zh. Vychisl. Mat. Mat. Fiz.*, 2000, vol. 40, no. 8, pp. 1264–1278.
3. Djukova, E.V. and Peskov, N.V., Discrete Methods of Information Analysis in Recognition and Algorithm Synthesis, *Zh. Vychisl. Mat. Mat. Fiz.*, 2002, vol. 42, no. 5, pp. 743–755.
4. Djukova, E.V. and Peskov, N.V., Selection of Typical Objects in Classes for Recognition Problems, *Pattern Recognit. Image Anal.*, 2002, vol. 12, no. 3, pp. 243–249.
5. Djukova, E.V. and Inyakin, A.S., Taxonomy Problem and Irredundant Coverings of Integer Matrix, *Applied Math. Proc.*, Moscow: Vychislitel'nyi Tsentr Ross. Akad. Nauk, 2002.