Combinatorial (Logical) Data Analysis in Pattern Recognition Problems[1]

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

**Computer Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, GSP-1, 119991 Russia*
e-mail: djukova@ccas.ru; andre_w@mail.ru; nick@motor.ru
***Moscow State Pedagogical University, Malaya Pirogovskaya ul. 1, Moscow, GSP-2, 119992 Russia*
e-mail: sahar@mail.ru

Abstract—The methods of increasing efficiency of the recognition, classification, and prediction procedures based on the combinatorial (logical) analysis of initial information are discussed.

INTRODUCTION

Combinatorial (logical) analysis of a feature space of high dimensionality measured by tens or even hundreds of features (when the number of learning objects is comparatively low) is the principal problem in pattern recognition. In this context, the fundamental problem of constructing fragments of descriptions of objects, which, in a sense, have extreme properties, is brought about. These fragments should reflect a certain regularity in descriptions of learning objects; they can play the role of elementary classifiers and allow one to classify new objects. Methods of logical analysis are especially efficient in the case of integral information when the number of the allowed values is low [4, 13].

This paper presents the results of analysis of implementation complexity of discrete recognition procedures. New approaches that allow one to enhance the quality of recognition procedures are also considered.

BASIC RESULTS

The method considered involves some difficulties related to computational complexity caused by an exhaustive search. Generally, this problem can be reduced to construction of coverings of Boolean and integer matrices, i.e., to the search for admissible and maximal conjunctions of logical functions. The process of constructing irreducible coverings (maximal conjunctions) is especially difficult.

The methods for constructing coverings of Boolean matrices have a crucial role here. These methods can be easily modified for constructing coverings of a more general type, they can be used directly, and they are more frequently demanded in practice. Also, a number of important problems in the framework of a discrete approach, such as a decrease in the modality of initial information [5] and synthesis of logical correctors on the basis of elementary classifiers [6], are directly connected with the problems of development of effective methods for constructing coverings of Boolean matrices.

An algorithm with a polynomial delay for constructing irreducible (0, …, 0) coverings of a Boolean matrix was introduced in [3, 10]. This algorithm is a modification of the algorithm with a polynomial delay developed in [1, 2] for approximate solution of the same task based on the search for all unitary submatrices of the initial matrix. Both algorithms have a common disadvantage: repeating steps. When the number of matrix columns n is greater in order than the number of rows m, the number of steps in each algorithm almost always asymptotically coincide with the number of irreducible (0, …, 0) coverings (for almost all Boolean matrices of dimensionality $m \times n$) at $n \longrightarrow \infty$. Theoretical estimates of complexities of approximate and exact algorithms were also worked out for other cases. Experiments with random matrices were performed.

Computational complexity and quality of logical procedures of recognition are traditionally connected with metric (quantitative) characteristics of the class of elementary classifiers. For example, a typical number of elementary classifiers and a typical length of an elementary classifier better allow us to evaluate the computational resources required, optimize computer memory, and, thus, lower the requirements for hardware during implementation of recognition-procedure programs. Investigations into the metric properties of the class of elementary classifiers are based on calculation of the asymptotic estimate of typical values of the number of (irreducible) σ-coverings and the length of the (irreducible) σ-covering of the integral matrix. These investigations are also based on the calculation of similar estimates for specially constructed submatri-

[1] This work was financially supported by the Russian Foundation for Basic Research, project no. 04-01-00795, and by the president of Russian Federation, grant no. 1721.2003.1.

Received October 25, 2004

Original Text Copyright © 2005 by Pattern Recognition and Image Analysis. English Translation Copyright © 2005 by MAIK "Nauka/Interperiodica" (Russia).

ces (σ-submatrices). The following results were obtained.

In [1, 2, 4, 13], the case where number of matrix rows m is less than number of columns n was considered. It was shown that the number of irreducible coverings of the integral matrix almost always coincides asymptotically with the number of σ-submatrices and is less than the number of coverings (for almost all matrices of dimensionality $m \times n$) at $n \longrightarrow \infty$. On this ground, an asymptotically optimal algorithm was designed for the search of irreducible coverings of the integral matrix.

In [8, 11], quite an opposite case was considered: the case where the number of rows in a matrix is greater than the number of its columns. The asymptotics of the typical values of the number of the σ-submatrices and the rank of the σ-submatrix were obtained similarly.

In [7], asymptotic estimates for the typical value of the number of (irreducible) σ-coverings and the length of (irreducible) σ-coverings close to minimal were obtained. It was shown that the number of (irreducible) σ-coverings of a length no greater than $\log_k m - \log_k \ln \log_k m - 1$ almost always coincide asymptotically with the number of (irreducible) σ-coverings of length $[\log_k m - \log_k \ln \log_k m - 1]$ at $n \longrightarrow \infty$.

In [9], the asymptotic estimate of a logarithm of the typical number of the irreducible σ-coverings was obtained for $n \le m$.

The estimates listed above were used for studying metric properties of disjunctive normal forms of logical functions.

The pressing problem now is improvement of the quality of pattern recognition and classification by using sets of allowed feature values, which are not included in the description of learning objects of the class. Such sets characterize the class on the whole and, thus, are more informative than sets used traditionally.

In [7, 8, 11], new models of recognition and classification algorithms based on the concept of nonoccurrence of sets of allowed feature values were constructed. Under certain conditions it was demonstrated that the algorithms described have an advantage over classical logical algorithms of recognition and classification based on calculation of the Hamming interval.

Noisy features and noisy objects lying on the class boundaries contribute to deterioration of recognition. Noisy features generate a large amount of unique elementary classifiers. These elementary classifiers are uncommon in the class, which they represent; therefore, they are hardly significant. The presence of boundary objects decreases the number of short fragments, which discriminate objects of different classes.

In [8, 11, 12], methods of improving the efficiency of discrete algorithms, based on the selection of typical feature values, typical objects, and informative zones in each class, are considered. Preliminary analysis of initial information allows us to decrease the influence of noisy features and to upgrade the recognition algorithm if the learning sample contains many objects lying on the boundary between classes. Here, by quality of recognition we mean quality of an algorithm obtained outside the learning sample. This quality is estimated by the quantity of objects correctly recognized during the cross-validation process.

Application of logical (discrete) procedures in the case where part of the information is presented by real-valued features (or integral-valued features of the high modality) can be highly difficult. To overcome this problem, correct encoding of initial information is used. By correct encoding we mean data transformation that differentiates between descriptions of learning objects from different classes. The problem can be reduced to constructing a special Boolean matrix by using the learning sample.

In [5], methods for correct encoding and calculating the estimate of encoding quality are obtained on the basis of analysis of information density (typicality) of feature values in the encoded learning sample.

REFERENCES

1. E. V. Djukova, "About Asymptotically Optimal Algorithm for Building Irreducible Tests," Dokl. Akad. Nauk SSSR **233** (4), 527–530 (1977).
2. E. V. Djukova, "The Complexity of Realization of Some Recognition Procedures," J. of Comp. Math. and Math. Phys. **27** (1), 114–127 (1987).
3. E. V. Djukova, "About Complexity of Realization of the Discreet (Logical) Procedures of Recognition," J. of Comp. Math. and Math. Phys. **44** (3), 550–572 (2004).
4. E. V. Djukova and Yu. I. Zhuravlev, "Discreet Analysis of Feature Descriptions in the Recognition Problems of High Dimensionality," J. of Comp. Math. and Math. Phys. **40** (8), 1264–1278 (2000).
5. E. V. Djukova, Yu. I. Zhuravlev, N. V. Peskov, and A. A. Sakharov, "Processing of a Real-Valued Information by the Logical Methods of Recognition," Iskusstvennyi Intellekt: J. of NAN of Ukraine, No. 2, 80–85 (2004).
6. E. V. Djukova, Yu. I. Zhuravlev, and K. V. Rudakov, "About Algebraic Synthesis of a Correct Recognition Procedures Based on Elementary Algorithms," J. of Comp. Math. and Math. Phys. **36** (8), 215–223 (1996).
7. E. V. Djukova and A. S. Inyakin, "About Classification Procedures Based on Building Class Coverings," J. of Comp. Math. and Math. Phys. **43** (12), 1910–1921 (2003).
8. E. V. Djukova and N. V. Peskov, "Discreet Procedures of Recognition: Search for Informative Fragments in Objects," J. of Comp. Math. and Math. Phys. **42** (5), 741–753 (2002).

9. E. V. Djukova and A. B. P'yanov, "Asymptotic of the Logarithm of the Number of Irreducible σ-coverings of Integral-Valued Matrixes," in *Proceedings of the 11th All-Russian Conference on Mathematical Methods of Pattern Recognition, Moscow Region, Russia, 2003*, pp. 80–82.

10. E. V. Djukova, "Discrete (Logical) Recognition Procedures: Principles of Construction, Complexity of Realization, and Basic Models," Pattern Recognition and Image Analysis **13** (3), 417–425 (2003).

11. E. V. Djukova, A. S. Inyakin, and N. V. Peskov, "Methods of Combinatorial Analysis in Synthesis of Efficient Recognition Algorithms," Pattern Recognition and Image Analysis **13** (3), 426–432 (2003).

12. E. V. Djukova and N. V. Peskov, "Selection of Typical Objects in Classes for Recognition Problems," Pattern Recognition and Image Analysis **12** (3), 243–249 (2002).

13. E. V. Djukova and Yu. I. Zhuravlev, "Discrete Methods of Information Analysis and Algorithm Synthesis" Pattern Recognition and Image Analysis **7** (2), 192–205 (1997).