Discrete Recognition Procedures: The Complexity of Realization[1]

E. V. Djukova

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

Abstract—New results concerning the complexity of realization of the discrete recognition procedures that use irreducible coverings of Boolean matrices (the search for maximal conjunctions of monotonous Boolean functions) are discussed.

INTRODUCTION

In constructing discrete recognition procedures, the apparatus of discrete mathematics is used, particularly, the methods of transformation of normal forms of logic functions and the theory of coverings of the Boolean and integer-valued matrices. When the problem at hand has large dimensionality, the computations become cumbersome due to the exhaustive search at the stage of searching for informative fragments in the feature descriptions of objects [1].

The recognition problem is stated as follows [2].

Let the objects be described in a system of features $\{x_1, \ldots, x_n\}$, where $x_j \in \{0, 1\}$, $j = 1, 2, \ldots, n$, and let $f_K(x_1, \ldots, x_n)$ be a partial (not everywhere defined) two-valued function equal to one on the tuples describing the class K of learning objects and to zero on the descriptors of the remaining learning objects. We consider an arbitrary fragment of a learning-object description as an elementary classifier. The particular conjunctions f_K correspond to elementary classifiers, and each recognition algorithm is determined by a set of such conjunctions M^A. Only elementary classifiers generated by the conjunctions from M^A are considered to be informative in using algorithm A. In most typical cases, e.g., in the construction of voting algorithms by using representative sets, it is necessary to obtain the admissible and maximal conjunctions of the function f_K. Most computationally difficult is a search for maximal conjunctions. Obviously, this problem can be solved on the basis of the transformation of the normal form of the total Boolean functions.

When the informational fragments are constructed, along with different problems of the transformation of the normal form of logic functions, the problem of constructing the irreducible coverings of a Boolean matrix naturally arises. This problem can be reduced to the problem of searching for maximal conjunctions of the monotonous Boolean function that is given by the conjunctive normal form (CNF). In this work, a computationally efficient algorithm for solving the above problems is presented (the polynomial time-delay algorithm). This is a slight modification of the algorithm previously introduced by the author. It is designed for solving the same problems in certain conditions, almost always with asymptotical accuracy; all intermediate calculations are also carried out with a polynomial time delay.

MAIN RESULTS

Let E^n be a set of n-tuples of the form $(\alpha_1, \ldots, \alpha_n)$, where $\alpha_i \in \{0, 1\}$ for $i = 1, 2, \ldots, n$. Let also A_K and B_K be the sets of n-tuples from E^n, where function f_K takes values 1 and 0, respectively; B is an elementary conjunction (EC) over variables $x_1, \ldots, x_n$; and N_B is the interval where the conjunction B is valid.

EC B is called *admissible* for f_K if $N_B \cap A_K \neq \emptyset$ and $N_B \cap B_K = \emptyset$.

EC B is called *irreducible* for f_K if there is no EC B' such that $N_{B'} \supset N_B$ and $|N_{B'} \cap B_K| = |N_B \cap B_K|$.

EC B is called *maximal* for f_K if it is admissible and there is no admissible conjunction B' such that $N_{B'} \supset N_B$.

It is obvious that the admissible and irreducible conjunction for f_K is maximal for f_K.

The above definitions of the admissible, irreducible, and maximal conjunction of a partially defined Boolean function are valid for the case of a fully determined Boolean function f_K, i.e., when $A_K = E^n \backslash B_K$.

One of the most known ways of constructing maximal conjunctions of a partially defined Boolean function is the following. Consider the fully determined function $F_K(x_1, \ldots, x_n)$ which coincides with f_K on the set of zeros; for all other sets of the Boolean cube, this function is equal to 1. Setting the fully determined Boolean function by a set of zeros is equivalent to the setting of its expanded CNF. Let B_K consist of the

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

Received October 29, 2002

n-tuples $(\beta_{11}, \ldots, \beta_{1n})$, $(\beta_{21}, \ldots, \beta_{2n})$, …, $(\beta_{u1}, \ldots, \beta_{un})$. Then, it is obvious that function F_K is realized by CNF of the form D_1 & … & D_u, where

$$D_i = x_1^{\bar{\beta}_{i1}} \vee \ldots \vee x_n^{\bar{\beta}_{in}}, \; i = 1, 2, \ldots, u.$$

It is easy to see that EC $B = x_{j_1}^{\sigma_1}, \ldots, x_{j_r}^{\sigma_r}$ is admissible for F_K if and only if each disjunction D_i, $i = 1, 2, \ldots, u$, has at least one multiplier from B. Whence, if we multiply logic brackets and simplify the obtained disjunctive normal form (DNF) using the identities $x\bar{x} = 0$, $xx = x$, and $x \vee x = x$, then we obtain a DNF that consists of all admissible conjunctions of function F_K. Now let us remove the admissible not-irreducible conjunctions by using the identity $x \vee xx' = x$. We obtain the DNF that consists of all maximal conjunctions of the function F_K (or reduced DNF). To obtain a set of maximal conjunctions for f_K, it is necessary to select those in the set of the constructed set of maximal conjunctions for F_K that are admissible for f_K. This can be easily done by searching for the sets from A_K.

For the discrete approach, the most interesting is the case when the number of features is greater than the number of objects. In this case, the number of brackets is less than the number of variables. We can demonstrate that almost always (for almost all CNF of the considered type) the number of admissible conjunctions is greater in order than the number of maximal conjunctions for $n \longrightarrow \infty$. Therefore, the algorithm for constructing maximal conjunctions described above is not efficient. Theoretic and practical investigations testify that, in this case, it is reasonable to start with constructing irreducible conjunctions of function F_K and, then, to check whether each of these conjunction is admissible. The following fact was proved in [3]. If $u \leq n^{1-\varepsilon}$ ($\varepsilon > 0$), then (for almost all CNFs of the considered type) the number of irreducible conjunctions almost always asymptotically coincides with the number of maximal conjunctions of function F_K for $n \longrightarrow \infty$. This allowed us to introduce an approach to searching for maximal conjunctions of function F_K which practically reduces the search to a minimum. This was done using the following technique. The initial problem of constructing the set G_K of all maximal conjunctions of function F_K was replaced by a simpler problem of constructing the set of all irreducible conjunctions of function F_K; i.e., the problem was solved approximately. The complexity of the approximate solution was estimated from the number of conjunctive multiplications. An algorithm was proposed for searching for all irreducible conjunctions of function F_K for which the number of conjunctive multiplications in the case when $u \leq n^{1-\varepsilon}$ ($\varepsilon > 0$) almost always asymptotically coincides with the number of maximal conjunctions of function F_K for $n \longrightarrow \infty$. In this algorithm, one conjunctive multiplication requires searching for no more than Oun variables in the given CNF. The computational (time) complexity of the initial problem almost always does not asymptotically exceed $O(un)|G_K|$, $n \longrightarrow \infty$. Thus, the approach to the construction of irreducible DNF of the function F_K introduced in [3] is, in a sense, asymptotically optimal.

The later works of the author generalize these results for the case of arbitrary CNF. Here, CNF is not perfect and the function F_K set by CNF is a two-valued function determined on the k-tuple n-dimensional sets for $k \geq 2$. A particular case was considered when the initial CNF realizes a monotonous Boolean function, i.e., does not contain negative variables (this is especially important for practice).

In implementation of some discrete procedures, such as test algorithms, voting algorithms using representative samples, etc., the constructions that are based on searching for irreducible coverings of Boolean matrices are used most often.

Let L be an arbitrary Boolean matrix. A set H of columns of L is called a covering if each row in L has a common 1 with at least one of the columns in H. A covering is called irreducible (irredundant) if none of its proper subsets is a covering. The following principle is used in constructing irreducible coverings. The collection of columns H of matrix L is irreducible if and only if the following two conditions are fulfilled: (1) the submatrix of L^H of matrix L made up by the columns of H contains no rows of the form (0, 0, …, 0) and (2) L^H contains each of the rows (1, 0, …, 0), (0, 1, …, 0), …, (0, 0, …, 1); i.e., it contains a unit submatrix.

How does the problem of constructing the irreducible coverings arise in the most typical cases? A specialized Boolean matrix is constructed for finding the desired set of elementary classifiers. We designate it $\tilde{L}$. Every row of this matrix is obtained after matching the pair of objects from the learning sample that belong to different classes. We put 1 in the jth column if the descriptions of the matched objects differ by the x_j feature and 0 otherwise. Let the learning sample contain objects $S_1, \ldots, S_m$ and let the object S_i, $i \in \{1, 2, \ldots, n\}$ belong to the class K, $\tilde{L}^{(i)}$ is a submatrix of the matrix $\tilde{L}$ obtained by matching the object S_i with all learning objects that do not belong to the class K.

The set of features $\{x_{j_1}, \ldots, x_{j_r}\}$ is called an (irredundant) test if the set of columns of the matrix $\tilde{L}$ with the numbers $j_1, \ldots, j_r$ is the (irreducible) covering.

A fragment of descriptions of a learning object S_i generated by the feature sample $\{x_{j_1}, \ldots, x_{j_r}\}$ is called an (irredundant) representative sample for the class K if the set of columns of the matrix $\tilde{L}^{(i)}$ with the numbers $j_1, \ldots, j_r$ is the (irreducible) covering.

Note that the problem of constructing all irreducible coverings in the $u \times n$ matrix L can be stated as a problem of transforming the conjunctive normal form of a Boolean function into its disjunctive normal form. Indeed, let us put into correspondence the ith row and disjunction $D_i = x_{p_1} \vee \ldots \vee x_{p_q}$, where $p_1, \ldots, p_q$ are the columns that yield 1 at the intersection with jth row. Let f_L be the monotonous Boolean function realized by CNF D_1 & … & D_u.

It is easy to prove the following three statements.

—EC $B = x_{j_1}, \ldots, x_{j_r}$ is allowed for f_L if and only if the collection of columns H of matrix L with the numbers $j_1, \ldots, j_r$ is a covering.

—EC $B = x_{j_1}, \ldots, x_{j_r}$ is maximal for f_L if and only if the collection of columns of matrix L with the numbers $j_1, \ldots, j_r$ is an irreducible covering.

—EC $B = x_{j_1}, \ldots, x_{j_r}$ is irreducible for f_L if and only if the collection of columns of matrix L with the numbers $j_1, \ldots, j_r$ contains a unit submatrix.

Let S_L be the set of all unit submatrices of matrix L and P_L be the set of all irreducible coverings of matrix L.

The above reasoning implies that the algorithms for constructing maximal conjunctions of a monotonous Boolean function can be easily modified for constructing the irreducible coverings of a Boolean matrix and vice versa. In particular, the insignificant modification of the asymptotically optimal algorithm for constructing maximal conjunctions of the monotonous Boolean function set by CNF gives the algorithm for constructing irreducible coverings of matrix L based on searching for all its unity submatrices. This algorithm has been introduced in [4]; its computational complexity is no greater than $O(un)|S_L|$, and, for $u \le n^{1-\varepsilon}$ ($\varepsilon > 0$), it almost always does not exceed $O(un)|P_L|$, $n \longrightarrow \infty$.

The insignificant modification of the algorithm allows us to avoid intermediate computations associated with finding all the unit submatrices of L and, on each step, to construct only those unit submatrices that correspond to the admissible conjunctions of function f_L, i.e., to construct an element from P_L at each step. The modified algorithm has a computational complexity that does not exceed $O(u^3n)|S_L^*|$, where S_L^* is a set of all unit submatrices from S_L that generate coverings from P_L.

REFERENCES

1. 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.
2. Zhuravlev, Yu.I, On Algebraic Approach to Problems of Recognition or Classification, in *Problemy Kibernetiki* (Problems in Cybernetics), Yablonskii, S.V., Ed., Moscow: Nauka, 1978, issue 33, pp. 5–68.
3. Djukova, E.V., On the Complexity of Implementation of Some Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 1987, vol. 27, no. 1, pp. 114–127.
4. Djukova, E.V., On an Asymptotically Optimal Algorithm for Constructing Irredundant Tests, *Dokl. Akad. Nauk SSSR*, 1977, vol. 233, no. 4, pp. 527–530.