Selection of Typical Objects in Classes for Recognition Problems[1]

E. V. Djukova and N. V. Peskov

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

Abstract—Discrete recognition procedures based on a search for sets of feature values that are not encountered in the feature descriptions of training objects are considered. The constructed recognition procedures are compared with classical procedures for real-life applied problems. An approach to improving the performance of recognition algorithms based on selecting training objects typical for each class is examined. A fast method for calculating estimates in voting over representative sets for the cross-validation procedure is suggested.

INTRODUCTION

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 [1]. 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 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 what classes they belong. The training objects are represented by their descriptions. Given a set of feature values (i.e., the description of some object S from M; generally, it is not known to what class it belongs), it is required to determine the class containing this set.

The discrete approach to recognition problems is based on a combinatorial analysis of training object descriptions aimed at determining the most informative subdescriptions of objects. Usually, subdescriptions are considered informative if they make it possible to distinguish between classes. The initial object descriptions are specified in the form of sets of integer-valued features. The discrete methods have given rise to a number of heuristics, which are called discrete, or logic, recognition procedures (they include test, KORA-type, and representative voting algorithms) [1–9].

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

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

In [10, 11], two new models of discrete recognition algorithms based on constructing σ-coverings of classes were suggested. The notion of σ-covering was introduced in [7] in relation to the development of efficient implementations of procedures for voting over irreducible representative sets. The use of the models mentioned above makes it possible to somewhat reduce computational expense in the cases where $|K| < |\bar{K}|$ (for instance, if the number of classes is large). This paper shows that, in solving certain applied problems, voting over class coverings is more effective than voting over representative sets.

In [11–13], we suggested an approach that substantially increased the effectiveness of recognition algorithms when the training sample contained many objects lying on boundaries between classes. This approach was based on partitioning the training sample into a base subsample (which contained objects “typical” of their classes) and a test subsample (containing “atypical” objects). The first was used to construct the set of elementary classifiers, and the second determined their weights [12, 13]. The selection of typical objects was based on estimating the informativeness of separate feature values. In this paper, we show that typical objects can also be constructed with the use of the cross-validation procedure. We suggest a fast method for computing estimates in voting over representative and irredundant representative sets for the cross-validation procedure.

The methods suggested in this paper have been tested on real-life problems of medical prediction.

1. DISCRETE RECOGNITION PROCEDURES BASED ON CONSTRUCTING COVERINGS OF CLASSES

Let $H \in \{x_{j_1}, \ldots, x_{j_r}\}$ be a set of features, and let $S' = \{a_1, \ldots, a_n\}$ be an object from the training sample.

[1] This work was supported by the Russian Foundation for Basic Research, project nos. 01-01-00575 and 00-15-96064, and by the Presidium of Russian Academy of Sciences, program “Intelligent Computer Systems,” project no. 1.1.

Received March 28, 2002

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 three 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 (S', H), where $S' \in K$, coincide with $(\sigma_1, \ldots, \sigma_r)$;

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

The first situation is least often, thus, it is hardly possible to process the sets of feature values with property (1). The essential difference in the informativeness of the other two properties is that property (2) characterizes only some subset of training objects from K, while property (3) characterizes all objects from K. Therefore, if it is important to consider the class K separately from the other classes, then, apparently, sets of feature values with property (3) are more informative. In this case, it is more natural to refer an object S to the class K if the set of feature values under consideration describes none of the objects in the class K, nor does it describe the object S to be recognized.

Suppose that two objects $S' = (a'_1, a'_2, \ldots, a'_n)$ and $S'' = (a''_1, a''_2, \ldots, a''_n)$ are given. We estimate the closeness of the objects S' and S'' with respect to a feature set $H = \{x_{j_1}, \ldots, x_{j_r}\}$ by the value

$$B(S', S'', H) = \begin{cases} 0 & \text{if } a'_{j_t} = a''_{j_t} \text{ for } t = 1, 2, \ldots, r; \\ 1, & \text{otherwise.} \end{cases}$$

The closeness of the object S' to an elementary classifier $\sigma = (\sigma_1, \ldots, \sigma_r)$ generated by a feature set $H = \{x_{j_1}, \ldots, x_{j_r}\}$ is estimated by

$$B(\sigma, S', H) = \begin{cases} 0 & \text{if } a'_{j_t} = \sigma_t \text{ for } t = 1, 2, \ldots, r; \\ 1, & \text{otherwise.} \end{cases}$$

For a class $K \in \{K_1, \ldots, K_l\}$, we set $\bar{K} = \{K_1, \ldots, K_l\} \backslash K$.

A particular recognition model A is determined by the principle for constructing the set C^A and the estimate $\Gamma(S, K)$ of the membership of an object S in a class K, which is evaluated by voting over elementary classifiers from C^A(K). For example, we can assume that an elementary classifier σ from $C^A(K)$ generated by a feature set H votes for the membership of S in K if $B(\sigma, S, H) = 0$. Then, the object S is referred to the class with the largest membership estimate Γ (if there are several such classes, recognition is rejected).

We say that a fragment (S', H) of the description of an object S' from a class K is a representative set for K if $B(S', S'', H) = 1$ for any training object S'' not belonging to the class K. A fragment (S', H) of the description of an object S' from a class K is a irredundant representative set for K if (1) $B(S', S'', H) = 1$ for any training object S'' from $\bar{K}$ and (2) for any set $H' \subset H$, there exists a training object S'' in $\bar{K}$ such that $B(S', S'', H') = 0$.

In the classical model of an algorithm for voting over (irredundant) representative sets, the set $C^A(K)$ comprises the (irredundant) representative sets for K. In the simplest modification of this model, the membership of an object S in a class K is estimated by the value

$$\Gamma_1(S, K) = \frac{1}{|C^A(K)|} \sum_{(S', H) \in C^A(K)} (1 - B(S, S', H)),$$

here and in what follows, $|N|$ denotes the cardinality of N.

Short representative sets are believed to be more informative; for this reason, in applied problems, only short representative sets are usually considered in order to improve recognition quality and reduce computational expenses. These representative sets may have bounded length or be irredundant.

Let us introduce the following notation: M^k_{mn}, where $k \geq 2$, is the set of all $m \times n$ matrices with elements from $\{0, 1, \ldots, k - 1\}$, and E is the set of all k-ary r-tuples.

Take $L \in M^k_{mn}$ and $\sigma \in E^r_k$. We say that a set H of r different columns of the matrix L is a σ-covering if the submatrix L^H of L formed by the columns from H does not contain the row σ. A set H of r different columns of the matrix L is an irredundant σ-covering if (1) the submatrix L^H does not contain the row $\sigma = (\sigma_1, \ldots, \sigma_n)$ and (2) for any $p \in \{1, 2, \ldots, r\}$, L^H contains at least one row of the form $(\sigma_1, \ldots, \sigma_{p-1}, \beta_p, \sigma_{p+1}, \ldots, \sigma_r)$, where $\beta_p \neq \sigma_p$.

Consider a class $K \in \{K_1, \ldots, K_l\}$. A training table T can be treated as a pair of matrices L_1 and L_2, where L_1 is the matrix formed by the descriptions of the training objects from the class K and L_2 is the matrix formed by the descriptions of the remaining training objects. Then, obviously, an elementary classifier $(\sigma_1, \ldots, \sigma_r)$ generated by a pair (S_i, H), where $S_i \in K$ and $H = \{x_{j_1}, \ldots, x_{j_r}\}$, is $a(n)$ (irredundant) representative set for K if and only if the set of columns with the numbers $j_1, \ldots, j_r$ in L_1 is not a $(\sigma_1, \ldots, \sigma_r)$-covering and the set of columns with numbers $j_1, \ldots, j_r$ in L_2 is a(n) (irredundant) $(\sigma_1, \ldots, \sigma_r)$-covering.

Now, consider discrete-type models based on constructing σ-coverings of matrices formed by descriptions of training objects from every class, namely, the models of voting over coverings of a class and over antirepresentative sets for a class [10, 11]. The use of these models makes it possible to somewhat reduce computational resources if $|K| < |\bar{K}|$ (for instance, if the number of classes is large). Below, we describe these models.

In the model of voting over (irredundant) coverings of a class, the set $C^A(K)$ comprises the elementary classifiers generated by (irredundant) coverings of the matrix formed by the descriptions of the training objects from the class K. The membership of an object S in the class K is estimated (in the simplest modification of the model) by the value

$$\Gamma_2(S, K) = \frac{1}{|C^A(K)|} \sum_{\sigma \in C^A(K)} B(\sigma, S, H).$$

Now, consider the model with antirepresentative sets. An elementary classifier σ generated by a (irredundant) σ-covering of a class K is a (irredundant) antirepresentative set if it coincides with at least one fragment of the form (S', H), where S' is a training object from $\bar{K}$. The membership of an object S in the class K is estimated (in the simplest modification) by the value

$$\Gamma_3(S, K) = \frac{1}{|C^A(K)|} \sum_{\sigma \in C^A(K)} B(\sigma, S, H).$$

Note that a set representative for a class K is antirepresentative for $\bar{K}$. It is easy to show that, if $l = 2$, both models refer an object S to the same class. Indeed, let A_1 be a representative voting algorithm, and let A_2 be an antirepresentative voting algorithm. As was mentioned, $C^{A_1}(K_1) = C^{A_2}(K_2) = \mathrm{C}_1$ and $C^{A_1}(K_2) = C^{A_2}(K_1) = \mathrm{C}_2$. Suppose that the object S contains q_1 fragments coinciding with representative sets for the class K_1 and q_2 fragments coinciding with representative sets for the class K_2. Then,

$$\Gamma_1(S, K_1) = \frac{q_1}{C^A(K_1)} = \frac{q_1}{C_1},$$

$$\Gamma_1(S, K_2) = \frac{q_2}{C^{A_1}(K_2)} = \frac{q_2}{C_2},$$

$$\Gamma_3(S, K_1) = \frac{C^{A_2}(K_1) - q_2}{C^{A_2}(K_1)} = \frac{C_2 - q_2}{C_2} = 1 - \frac{q_2}{C_2},$$

$$\Gamma_3(S, K_2) = \frac{C^{A_2}(K_2) - q_1}{C^{A_2}(K_2)} = 1 - \frac{q_1}{C_1}.$$

Thus, $\Gamma_1(S, K_1) > \Gamma_1(S, K_2)$ if and only if $\Gamma_3(S, K_1) > \Gamma_3(S, K_2)$.

2. SELECTION OF TYPICAL OBJECTS IN CLASSES AND CONSTRUCTION OF "WEIGHTIEST" ELEMENTARY CLASSIFIERS

In solving an applied recognition problem, it is interesting to try to estimate how effectively the constructed algorithm recognizes objects not included in the training sample. For this purpose, the well-known cross-validation method can be applied. Unfortunately, in some applied problems, the algorithms described in Section 1 not always exhibit high effectiveness. This happens if the classes are poorly separated from each other (i.e., each class contains many objects whose descriptions are similar to those of objects not belonging to this class). In this case, algorithms often well recognize the objects "known" to them (those used in constructing the algorithms) but poorly recognize "new" objects. This section suggests an approach that makes it possible to improve the quality of recognizing algorithms. This approach is exemplified by a representative voting model.

In applied problems, serious difficulties are caused by the presence of the objects lying on the boundary between classes (their descriptions are similar to the descriptions of objects not belonging to the class under consideration). Naturally, such objects are hard to recognize, and, apparently, they do not admit short representative sets. Suppose that the description of a training object not belonging to a class K is similar to the descriptions of some objects from K. Then this object "deprives" the class K of some short representative sets, which substantially deteriorates the effectiveness of the algorithm. To overcome this difficulty, we suggest to divide the training sample into two subsamples; the first (base) is used to construct representative sets and the second (test), to evaluate their weights. The sample should be divided in such a way that all objects lying on boundaries between classes are contained in the test subsample and all remaining (typical) objects, in the base subsample. Practical experiments based on applied problems show that such a division increases the number of short representative sets and, thereby, makes it possible to improve the performance of the recognition algorithm.

For selecting typical objects, we suggest to apply the well-known cross-validation method. We include the training objects correctly recognized under the cross-validation test in the base subsample and all remaining objects in the test subsample. This approach is fairly effective, but it is very burdensome when applied to problems of high dimensions.

To reduce computational expenses, we suggest the procedure for selecting typical objects based on computing the informativeness of separate feature values [11–13].

Suppose that $S' \in K_i$, where $i \in \{1, 2, \ldots, l\}$ and $j \in \{1, 2, \ldots, n\}$. We set

$$\bar{K}_i = \bigcup_{q=1}^{l} K_q \backslash K_i,$$

$$\mu_{ij}^{(1)}(S') = \frac{1}{K_i} \sum_{S'' \in K_i} (1 - B(S', S'', \{x_j\})),$$

$$\mu_{ij}^{(2)}(S') = \frac{1}{\bar{K}_i} \sum_{S'' \in \bar{K}_i} (1 - B(S', S'', \{x_j\})),$$

The quantities $\mu_{ij}^{(1)}(S')$ and $\mu_{ij}^{(2)}(S')$ characterize the closeness of the object S' to its class and to the other classes, respectively. The value

$$\mu_{ij}(S') = \mu_{ij}^{(1)} - \mu_{ij}^{(2)}$$

is called the weight of the value of the feature x_j for the object S'. We say that the value of the feature x_j is typical of S' if $\mu_{ij}(S') > \mu$, where μ is the minimum informativeness threshold for feature values.

Take an integer p such that $1 \le p \le n$. We consider the object S' to be typical of the class K_i with respect to the threshold p if the inequality $\mu_{ij}(S_i) > \mu$ holds for at least p features.

The thresholds μ and p can be chosen from heuristic considerations; for example, we can take $\mu = 0$ and $p = [n/2]$. Then, the value of the feature x_j for S' is typical of the class K_i if it is encountered in K_i more frequently than in $\bar{K}_i$, and the object S' is typical of K_i if at least half the feature values in its description are typical of K_i.

Suppose that the training sample is divided into base and test subsamples by one of the methods described above. We use the base subsample to construct representative sets. To each constructed representative set, we assign a weight, which is calculated with the use of the test sample.

Let ω be a representative set for a class $K \in \{K_1, \ldots, K_l\}$ generated by a pair (S', H), where S' is an object from the base subsample; by $\delta(K, \omega)$, we denote the number of objects in the test sample for which the representative set votes "correctly," and by $\delta(\bar{K}, \omega)$, the number of objects in the test sample for which it votes "incorrectly." As $\nu_{(S', H)}$, we can take the functions

$$1)\ \nu_{(S', H)} = \delta(K, \omega);$$

$$2)\ \nu_{(S', H)}$$

$$= \begin{cases} \delta(K, \omega) - \delta(\bar{K}, \omega) & \text{if } \delta(K, \omega) > \delta(\bar{K}, \omega), \\ 0 \text{ if } \delta(K, \omega) < \delta(\bar{K}, \omega); \end{cases}$$

$$3)\ \nu_{(S', H)} = \frac{1 + \delta(K, \omega)}{1 + \delta(\bar{K}, \omega)}.$$

We estimate the membership of an object S in the class K by the value

$$\Gamma_4(S, K) = \frac{1}{|C^A(K)|}$$

$$\times \sum_{(S', H) \in C^A(K)} \nu_{(S', H)}(1 - B(S, S', H)),$$

As the informative weight of a feature x_j, we take

$$I_j = \frac{\sum_{(S', H) \in C^A,\, x_j \in H} \nu_{(S', H)}}{\sum_{(S', H) \in C^A} \nu_{(S', H)}}.$$

3. A FAST METHOD FOR CALCULATING ESTIMATES IN REPRESENTATIVE VOTING FOR THE CROSS-VALIDATION PROCEDURE

When the cross-validation test is used in the representative voting algorithm, the estimates are usually calculated by the following procedure. At each $i \in \{1, \ldots, m\}$, representative sets for the sample $\{S_1, \ldots, S_m\} \backslash \{S_i\}$ are constructed, and for these sets, $\Gamma(S_i, K)$ and $\Gamma(S_i, \bar{K})$ are evaluated. Clearly, this procedure involves substantial computational resource for large problems. Below, we suggest a method reducing the computation time approximately m-fold.

Suppose that $K \in \{K_1, \ldots, K_l\}$, $S \in K$, and $S \in \{S_1, \ldots, S_m\}$. For simplicity, we consider the case of $l = 2$.

Let us introduce the following notation:

$Q_1(S, K)$ is the family of all sets of features $H \subseteq \{x_1, \ldots, x_n\}$ such that none of the fragments (S', H), where $S' \in \{S_1, \ldots, S_m\}$ and $S' \notin K$, coincides with the fragment (S, H) (thus, $Q_1(S, K)$ is the family of all feature sets H such that the fragment (S, H) is a representative set for the class K);

$N_1(S, H)$, where $H \in Q_1(S, K)$, is the number of the fragments (S', H), where $S' \in \{S_1, \ldots, S_m\}$ and $S' \in K$, that coincide with the fragment (S, H), including the fragment (S, H);

$Q_2(S, K)$ is the family of all feature sets $H \subseteq \{x_1, \ldots, x_n\}$ such that (1) none of the fragments (S', H), where $S' \in \{S_1, \ldots, S_m\}$, $S' \in K$, and $S' \in S$, coincides with (S, H) and (2) at least one fragment (S'', H), where $S'' \in \{S_1, \ldots, S_m\}$ and $S'' \notin K$, coincides with (S, H);

$N_2(S, H)$, where $H \in Q_2(S, K)$, is the number of fragments (S'', H), where $S'' \in \{S_1, \ldots, S_m\}$ and $S'' \notin K$, coinciding with (S, H);

$$\lambda_{11}(S, K) = \sum_{H \in Q_1(S, K)} [N_1(S, H) - 1];$$

$$\lambda_{12}(S, K) = \sum_{H \in Q_1(S, K)} [N_1(S, H) - 1]$$

$$+ \sum_{\substack{S' \in K \\ S' \in S}} \sum_{H \in Q_1(S, K)} N_1(S', H);$$

$$\lambda_{21}(S, K) = \sum_{H \in Q_2(S, K)} N_2(S', H);$$

and

$$\lambda_{22}(S, K) = \sum_{H \in Q_2(S, K)} N_2(S, H)$$

$$+ \sum_{S' \in \bar{K}} \sum_{H \in Q_1(S, \bar{K})} N_1(S', H).$$

It is obvious that

$$\Gamma(S, K) = \lambda_{11}(S, K)/\lambda_{12}(S, K) \qquad (1)$$

and

$$\Gamma(S, \bar{K}) = \lambda_{21}(S, K)/\lambda_{22}(S, K) \qquad (2)$$

Thus, to evaluate the required estimates, we must find, first, representative sets for the class K (these sets are involved in constructing the families $Q_1(S, K)$ for $S \in K$) and, second, so-called $(1, q)$-representative sets for K (see [6]), i.e., fragments such that, for each of them, the given set of the feature values is encountered precisely once in K and precisely q times in the other class.

The values $\lambda_{11}(S, K)$, $\lambda_{12}(S, K)$, $\lambda_{21}(S, K)$, and $\lambda_{22}(S, K)$ for $S \in \{S_1, \ldots, S_m\}$ and $K \in \{K_1, \ldots, K_l\}$ can be calculated as follows. Initially, we set these values to zero. Suppose that, at the current step, a representative set for K generated by objects $S_{i_1}, \ldots, S_{i_p}$ is found. Then, we increase the values $\lambda_{11}(S_{i_t}, K)$ and $\lambda_{12}(S_{i_t}, K)$ with $t = 1, 2, \ldots, p$ by $p - 1$ and $\lambda_{12}(S_j, K)$ with $j \notin \{i_1, \ldots, i_p\}$ by p. If, at the current step, a $(1, q)$-representative set for K generated by the object S is found, then we increase the values $\lambda_{21}(S, K)$ and $\lambda_{22}(S, K)$ by q.

Formulas (1) and (2) can easily be generalized to $l > 1$ and to irredundant representative sets.

Clearly, in voting over irredundant representative sets, as $Q_1(S, K)$, we should take the family of all feature sets $H \subseteq \{x_1, \ldots, x_n\}$ such that the fragment (S, H) is a(n) irredundant representative set for the class K, and as $Q_2(S, K)$, we should take the family of all feature sets $H \subseteq \{x_1, \ldots, x_n\}$ such that (1) none of the fragments (S', H), where $S' \in \{S_1, \ldots, S_m\}$, $S' \in K$, and $S' \neq S$, coincides with (S, H); (2) at least one fragment (S'', H), where $S'' \in \{S_1, \ldots, S_m\}$ and $S'' \notin K$, coincides with (S, H); and (3) for each $t \in \{1, 2, \ldots, r\}$, K contains a row $S_t' \neq S$ such that $(S_t', H^{(t)}) = (S, H^{(t)})$, where $H^{(t)} = H \backslash \{x\}$.

4. TESTING ON MEDICAL PREDICTION PROBLEMS

The new models described in this paper were compared with classical constructions for problems of medical prediction.

Osteogenic sarcoma is a cancerous disease of bones which largely attack young people (and is virtually not encountered among elderly people). Unfortunately, the probability of lethal outcome of osteogenic sarcoma is very high. Patients with sarcoma are mostly cured by chemical methods; that is, they take small portions of toxic substances during a certain period. Since the cancerous cells grow much faster than the healthy cells, they consume the poison faster. As a result, the cancerous tumor begins to decay even before the organism is poisoned.

In this field, there are two important problems, the problem of survivability (whether the patient survives one year after the treatment) and the problem of predicting pathomorphosis, i.e., the degree of tumor destruction after a course of chemotherapy. Preceding investigations show that the prediction of survivability is very difficult, because, in addition to the condition of the cancerous cells, it involves a lot of other important objective factors, such as patient’s immunity, psychic condition, environment, etc. The problem of predicting the degree of pathomorphosis is much easier to solve, because the state of the cancerous cells plays a key role in this problem, and the influence of the other factors is much less.

The training sample comprised 77 objects (patients) divided into two classes. For the survivability prediction problem, the cardinalities of the classes were 52 and 25, and for the problem of predicting the pathomorphosis degree, 47 and 30. The objects were described in a system of seven features (certain characteristics of cancerous tumor). Each feature was three-valued.

The effectiveness of the recognition procedures was estimated by the cross-validation method.

Testing showed that, in the model of representative voting, representative sets of length 3 were sufficient to solve both problems. Employing longer representative sets in the construction of the recognition algorithm did not affect the effectiveness of the algorithm. Employing shorter representative sets reduced the effectiveness.

The effectiveness of the classical model of representative voting (over representative sets of length 3) was 61% for the survivability problem and 83% for the pathomorphosis problem, while the effectiveness of the

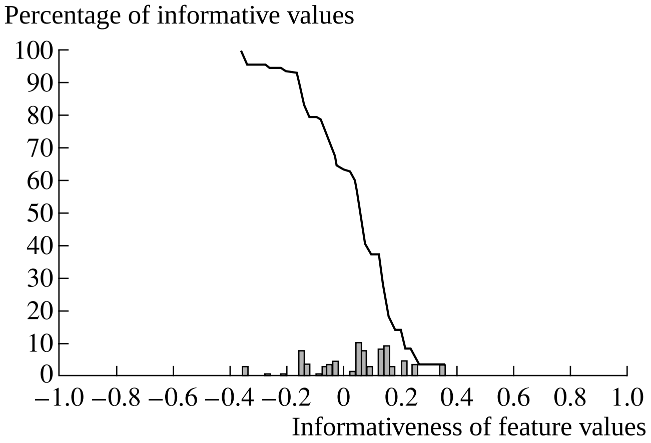

Fig. 1. Distribution of feature value typicality in the survivability problem.

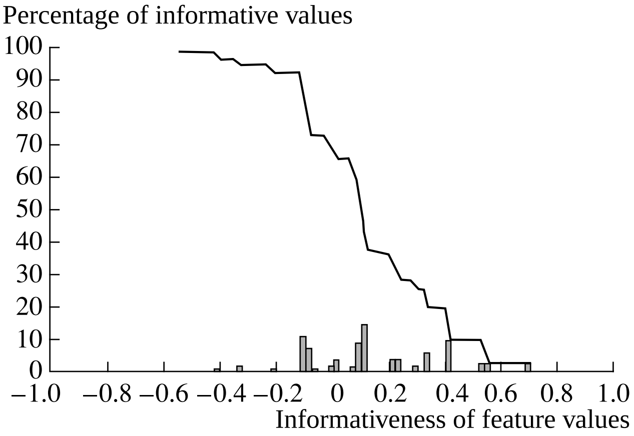

Fig. 2. Distribution of feature value typicality in the pathomorphosis problem.

algorithm of voting over coverings of classes was 75% and 92%, respectively.

To understand the reason for the low effectiveness of the classical algorithm as applied to the survivability problem, the informativeness of feature values was analyzed according to the scheme described in Section 2. The results of the analysis for the survivability and pathomorphosis problems are given in Figs. 1 and 2, respectively. These figures show the dependence of the percentage of the typical feature values on the minimum informativeness threshold. The diagrams show the distribution of the weights of feature values (the height of each column corresponds to the percentage of those feature values in the descriptions of the objects from the training sample that have weights contained in the given interval).

Figure 1 shows that the majority of feature values have weights close to zero. This means that these values are encountered in both classes with equal frequency. In other words, objects from different classes are hard to separate from each other, which causes the low effectiveness of the classical recognition algorithm.

Figure 2 shows that, in the pathomorphosis problem, a part of the feature values have weights close to zero, but there are many values with fairly large weights, i.e., very typical of one of the classes.

Table

	Survivability	Pathomorphosis
Classical model	61%	83%
Partition by the cross-validation method	75%	94%
Partition based on estimating the typicality of feature values	75%	92%

To increase the effectiveness of the recognizing algorithms, the following approach was used. The initial sample was divided into base and test samples in two ways, namely, by the cross-validation method and by the method based on estimating the typicality of feature values relative to the classes (see Section 2). The effectiveness of the constructed algorithms was estimated by the number of recognized objects with the use of the cross-validation method. More precisely, the following procedure was implemented. One object was removed from the training sample; the remaining objects were divided into base and test subsamples; then, the base subsample was used to construct representative sets and the test subsample, to calculate their weights. Voting over the representative sets with taking into account the weight was performed, and a decision on the classification of the removed object was made. This procedure was repeated for each object from the training sample. The computational results are given in the table.

Thus, the application of the methods suggested in this paper makes it possible to substantially improve the performance of recognition algorithms. The reason for this is as follows. Let us analyze the constructed representative sets, or, more precisely, their cardinalities. In the survivability problem, when the classical model is used, the numbers of the constructed representative sets for the first and second classes are 472 and 154, respectively. If the representative sets are constructed only for typical objects, and these typical objects are selected with the use of, e.g., the cross-validation method, then these numbers are 730 and 252, respectively. In the pathomorphosis prediction problem, the numbers of representative sets are 657 and 468 in the classical model and 849 and 668 in the model with division into base and test subsamples.

These data confirm the conjecture that the presence of atypical objects decreases the number of short representative sets and, thereby, deteriorates the performance of the constructed algorithm.

CONCLUSION

This paper studies discrete recognition procedures. The key stage in constructing these procedures is searching for informative fragments of feature descriptions of objects. This paper suggests new approaches to searching for such fragments.

(1) General principles for constructing discrete (logic) recognition procedures are described.

(2) Some models of discrete-type procedures are considered; these are the classical model of voting over representative sets and two new models based on constructing sets of feature values not encountered in the feature descriptions of the training objects. These models are compared for real-life applied problems.

(3) An approach to improving the performance of recognition algorithms based on selecting training objects typical of each class is examined. On the example of medical prediction, it is shown that this approach can improve the performance of recognition algorithms.

(4) A fast method for estimate calculation in voting over representative sets for the cross-validation procedure is suggested; this method makes it possible to substantially reduce the computational time in comparison with the conventional method.

REFERENCES

1. Zhuravlev, Yu.I., An Algebraic Approach to Recognition and Classification Problems, in *Problemy kibernetiki* (Problems of Cybernetics), Moscow: Nauka, 1978, issue 33, pp. 5–68.
2. Dmitriev, A.I., Zhuravlev, Yu.I., and Krendelev, F.P., On Mathematical Principles for Classification of Things and Phenomena, in *Diskretnyi analiz* (Discrete Analysis), Novosibirsk: Inst. Mat., Sib. Otd. Akad. Nauk SSSR, 1966, vol. 7, pp. 3–17.
3. Djukova, E.V., Asymptotically Optimal Test Algorithms in Recognition Problems, in *Problemy kibernetiki* (Problems of Cybernetics), Moscow: Nauka, 1982, issue 39, pp. 165–199.
4. Baskakova, L.V. and Zhuravlev, Yu.I., A Model of Recognition Algorithms with Representative Sets and Systems of Support Sets, *Kibernetika*, 1978, no. 4, pp. 131–137.
5. Vaintsvaig, M.N., Pattern Recognition Learning Algorithm KORA, in *Algoritmy obucheniya raspoznavaniyu obrazov* (Pattern Recognition Learning Algorithms), Moscow: Sovetskoe Radio, 1973, pp. 82–91.
6. Djukova, E.V., KORA-type Recognition Algorithms: Complexity of Implementation and Metric Properties, in *Raspoznavanie, klassifikatsiya, prognoz: Matematicheskie metody i ikh primenenie* (Recognition, Classification, Forecasting: Mathematical Methods and Their Application, Moscow: Nauka, 1989, ussue 2, pp. 99–125.
7. Djukova, E.V., On the Complexity of Implementation of Some Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 1987, vol. 21, no. 1, pp. 144–227.
8. 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.
9. Djukova, E.V. and Zhuravlev, Yu.I., Discrete Methods of Information Analysis in Recognition and Algorithm Synthesis, *Pattern Recogn. Image Anal.*, 1997, vol. 7, no. 2, pp. 192–207.
10. Djukova, E.V. and Peskov, N.V., On Discrete Recognition Procedures Based on Constructing Coverings of Classes, *Dokl. 10 Vserossiiskoi konferentsii po matematicheskim metodam raspoznavaniya obrazov* (Proc. 10th All-Russia Conf. on Math. Methods of Pattern Recognition), Moscow, 2001, pp. 48–51.
11. Djukova, E.V. and Peskov, N.V., A Search for Informative Fragments of Object Descriptions in Discrete Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 2002, vol. 42, no. 5, pp. 743–755.
12. Djukova, E.V. and Peskov, N.V., On Some Approaches to Evaluating Informative Characteristics of Training Samples, *Dokl. 9 Vserossiiskoi konferentsii po matematicheskim metodam raspoznavaniya obrazov* (Proc. 9th All-Russia Conf. on Mathematical Methods of Pattern Recognition), Moscow, 1999, pp. 181–183.
13. Djukova, E.V. and Peskov, N.V., Informativeness of Features, Separate Feature Values, and Fragments of Object Descriptions, *Dokl. 10 Vserossiiskoi konferentsii po matematicheskim metodam raspoznavaniya obrazov* (Proc. 10th All-Russia Conf. on Mathematical Methods of Pattern Recognition), Moscow, 2001, pp. 201–204.

Elena V. Djukova. Born 1945. Graduated from Moscow State University in 1967. Received her PhD (Kandidat Nauk) Degree in physics and mathematics in 1979 and Doctoral (Doktor Nauk) Degree in physics and mathematics in 1997. Leading Researcher at the Computer Center, Russian Academy of Sciences. Scientific interests: discrete mathematics and mathematical methods of pattern recognition and image processing. The author of about 60 publications.

Nikolai V. Peskov. Born 1978. Graduated from Moscow State University in 2000. Postgraduate student at the "Kibernetika" Scientific Council on Cybernetics, Russian Academy of Sciences. Scientific interests: discrete mathematics and mathematical methods of pattern recognition and image processing.

SPELL: Djukova, combinatorial, subdescriptions, subsample, irredundant, cardinality, -ary, -tuples, submatrix, antirepresentative, subsamples, Osteogenic, pathomorphosis, cardinalities, Kandidat, Doktor, ussue