Statistical approach to building a model of recognition operators under conditions of high dimensionality of a feature space

Sh Kh Fazilov¹, R A Lutfullaev¹, N M Mirzaev¹,³, A Sh Mukhamadiev²

¹ Scientific and innovation center of information and communication technologies, Tashkent university of information technologies named after Muhammad al-Khwarizmi, 17A, Buz-2 massif, Tashkent, 100125, Uzbekistan
² Department of Audiovisual Technologies, Tashkent university of information technologies named after Muhammad al-Khwarizmi, 108, Amir Temur street, Tashkent, 100200, Uzbekistan
³ Corresponding E-mail: nomazmirza@rambler.ru

Abstract. The task of constructing recognition algorithms (RAs) focused on the classification of objects under conditions of high dimensionality of a feature space is considered. As an initial model, a model of RAs based on the calculation of estimates is considered. A distinctive feature of the proposed model is the formation of subsets of interrelated features (SIF) and the selection of a set of representative features (RFs) in the construction of RAs. Moreover, the proximity estimates between the object and the class under consideration are calculated on the basis of the Bayesian approach. The main advantage of the proposed operators is the allocation of preferred dependency models with the subsequent calculation of the assessment of the ownership of objects and ensuring a significant reduction in the number of computational operations when recognition unknown objects. This feature is very important for real-time recognition systems. To test the performance of the proposed model, experimental studies were carried out in solving a model problem. This model can be used in the compilation of various programs aimed at solving problems of forecasting and classifying objects defined in the space of features of large dimension.

1. Introduction
The problem of pattern recognition (PR) is one of the central problems of theoretical informatics. For a long time, this problem has been in the focus of attention of specialists in the field of applied mathematics and theoretical informatics. As indicated in [1, 2], among the first works on PR, the works of R. Fisher, A.N. Kolmogorov and A.Ya. Khinchin were made in the first half of the last century. Today, a number of, a model of recognition [3–14] have been deeply developed and studied in detail, such as models based on separating functions; models based on mathematical statistics and probability theory; models built on the principle of potentials; models based on the calculation of estimates. Analysis of these models shows that they are mainly focused on solving problems when objects are described in the space of independent features (or the relationship between features is rather weak).

In practice, often the applied tasks of PR are given in the space of features of a large dimension. When solving such problems, the assumption of independence of signs is often not fulfilled. Most
models of RAs, in such conditions, require the use of enormous computing power, which can be provided only by high-performance computer equipment. Consequently, there are still insufficiently studied questions on the creation of RAs that can be applied to solve applied recognition problems for large dimensions of a feature space and the presence of interconnectedness of features. Therefore, the issues of improvement, development and research of models of RAs, focused on solving problems of diagnosing, predicting and classifying objects in conditions of high dimensionality of a feature space, are relevant.

The purpose of this work is to develop models of modified RA, aimed at adapting it to solving image recognition problems within RA, based on the calculation of estimates.

To solve the problem of PR under conditions of high dimensionality of the feature space, a statistical approach is proposed, which is based on the results of the research of scientific schools of Yu.I. Zhuravlev. On the basis of this approach, a modified model of estimation calculation algorithms (ECA) was developed, focused on the classification of objects in conditions of high dimensionality of a feature space.

It is known [3] that any RA consists of two consecutively executed operators: the recognition operator (RO) and the decision rule (see formula (5) in section 2). Therefore, in this paper, we consider the problems of constructing ROs (see on (6) in section 2), and the decision rule is assumed to be fixed (see on (7) in section 2).

2. Basic concepts and notation

Based on [3], we introduce some concepts and notation. We consider the set of valid objects $\mathcal{D} = \{S_1, S_2, \ldots \}$, which are covered by $\ell$ subsets (classes) $\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_\ell$:

$$\mathcal{D} = \bigcup_{j=1}^{\ell} \mathcal{C}_j \cap \mathcal{C}_j = \emptyset, i \neq j, i, j \in \{1, 2, \ldots, \ell\}.$$  

(1)

In this case, the partition (1) is not completely defined. There is only some initial information $I_0$ about classes $\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_\ell$.

Let objects be given $\{S_1, S_2, \ldots, S_u, \ldots, S_m\} (S_u \in \mathcal{D}, u = 1, m)$. We introduce the following notation:

$$\mathcal{S}^m = \{S_1, S_2, \ldots, S_u, \ldots, S_m\}, \mathcal{C}_j = \mathcal{S}^m \cap \mathcal{C}_j, \mathcal{R}_j = \mathcal{S}^m \setminus \mathcal{C}_j$$

(2)

Then, based on (2), you can specify the initial information $I_0$ as a set of pairs consisting of $S_u$ and $\tilde{a}(S_u)$:

$$I_0 = \{< S_1, \tilde{a}(S_1) >, \ldots, < S_u, \tilde{a}(S_u) >, \ldots, < S_m, \tilde{a}(S_m) > \}$$

(3)

where $\tilde{a}(S_u)$ - object information vector $S_u$, which is given as:

$$\tilde{a}(S_u) = (a_{u1}, \ldots, a_{uj}, \ldots, a_{u\ell}), a_{uj} = \begin{cases} \{ 1, if S_u \in \mathcal{C}_j; \\ 0, if S_u \nvin \mathcal{C}_j. \end{cases}$$

(4)

The set of information vectors corresponding to the objects $\mathcal{S}^m$, forms the information matrix $\|a_u\|_{m \times \ell}$.

It is known [3] that an arbitrary RA can be represented as a sequential execution of the operators $\mathcal{B}$ (recognition operator) and $\mathcal{C}$ (decision rule):

$$\mathcal{A} = \mathcal{B} \circ \mathcal{C}$$

(5)

From (5) it follows that each RA $\mathcal{A}$ can be divided into two successive stages. At the first stage, the discriminating operator $\mathcal{B}$ converts the admissible object $S_u$ into a numerical estimate represented by the vector $b_u$:

$$\mathcal{B}(S_u) = \bar{b}_u,$$

(6)

where $\bar{b}_u = (b_{u1}, \ldots, b_{uw}, \ldots, b_{u\ell})$. 

2
At the second stage, according to the numerical estimate $b_{uv}$ the decision rule $\mathcal{C}$ determines the belonging of the object $S_u$ to the classes $\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_l$:

$$
\mathcal{C}(b_{uv}) = \begin{cases} 
0, & \text{if } b_{uv} < c_1; \\
\Delta, & \text{if } c_1 \leq b_{uv} \leq c_2; \\
1, & \text{if } b_{uv} > c_2,
\end{cases}
$$

(7)

where $c_1, c_2$ – decision rule parameters. In this case, the $b_{uv}$ estimate is calculated using the RO (6).

3. Statement of the problem

With initial information $I_q$, presented in the form of a set (3). It is assumed that each object $S_u (S_u \in \tilde{S}^m)$ in the space $X$ corresponds to a description vector:

$$
\mathcal{J}(S_u) = (a_{u1}, \ldots, a_{ui}, \ldots, a_{un}),
$$

where $X = (x_1, \ldots, x_i, \ldots, x_n)$ – feature space.

The main task is to build such model of the RO $\mathcal{B}$ (see formula 6 in section 2), which, using the decision rule $\mathcal{C}$ (see formula 7 in section 2), approximates acceptable accuracy of the dependence between objects (observations) described by a set of features and some set of values called information vectors (see formula 4 in section 2).

4. Proposed Approach

To solve the formulated problem, a statistical approach is proposed for the construction of pattern RAs specified in a high-dimensional feature space. A distinctive feature of this approach is the definition of the statistical proximity function in the framework of the ECA. Based on this approach, a model of modified RAs based on the calculation of estimates has been developed. The task of the proposed MPR includes the following main steps. It should be noted that the considered MPR in the initial two stages does not differ from the corresponding stages of ROs based on radial functions [15]. Therefore, these steps are only listed for the sake of completeness.

1. Selecting SIF.
2. Formation of a set of RFs.
3. Defining dependency models in each feature subset for a class $\mathcal{C}_j$ ($j = \overline{1,l}$). Let $x_i$ be an arbitrary feature belonging to a subset $\Omega_q$. It is assumed that the elements $\Omega_q$ is linearly ordered by feature index (i.e. $x_i \leq x_j$, if $i < j$). Further, the zero element ($x_0$) of the subset $\Omega_q$ is $x_q$, the remaining elements are denoted by $x_i$ ($N_q = \Omega_q | i = 1, \ldots, N_q - 1$). Then the dependency model in $\Omega_q$ takes the form [14]:

$$
x_i = F(\vec{c}, x_q), \quad x_i \in \Omega_q \backslash x_q
$$

(8)

where $\vec{c}$ – vector of unknown parameters, $F$ – function from some given class $\{F\}$.

The calculated values of the vector of unknown parameters $\vec{c}$ determine the dependence model in the subset of features $\Omega_q$ for the class $\mathcal{C}_j$ ($j = \overline{1,l}$). Depending on the specification of the parametric type $F(\vec{c}, x)$ and the method of determining $\vec{c}$ we obtain various dependence models in a subset of features $\Omega_q (q = \overline{1,n})$.

We can consider a simple example. As a given set $\{F\}$ we consider linear models. Here it is assumed that the feature $x_q$ ($x_q \in \Omega_q$) is an independent variable, and the feature $x_i$ ($x_i \in \Omega_q \backslash x_q$) is a dependent variable. Then the dependency model in $\Omega_q$ is given as:

$$
x_i = c_{i0} + c_{i1}x_q,
$$

where $c_{i0}, c_{i1}$ – parameters that are determined based on the least squares test [16, 17].

4. Determination of the proximity function $B_q(\mathcal{C}_j, S)$ between $\mathcal{C}_j$ and the object $S$ according to the signs included in $\Omega_q$. At this stage, the proximity function is defined, which determines the degree of belonging of the object $S$ to the class $\mathcal{C}_j$. We suppose that dependency models (for each class $\mathcal{C}_j$) are given as (8). We introduce the distance function $d_q(\mathcal{C}_j, S)$ between the class $\mathcal{C}_j$ and the object $S$: 

3
\[ d_q(C_j, S) = |a_i - F_j(\mathcal{E}, x_q)|, \]  \hspace{1cm} (9)

where \( x_q \) – the value of the \( q \)-th RF of the object \( S \); \( a_i \) – the value of the \( i \)-th feature of the object \( S \).

The set of valid values calculated using formula (9) is denoted by \( \mathcal{D}_q(q \in \{1, 2, ..., n'\}) \):

\[ \mathcal{D}_q = \{d_q(C_j, S)|S \in S^m\}. \]

In this case, it is assumed that the elements of \( \mathcal{D}_q \) are divided into \( k \) disjoint equal or different in size intervals \( H_i^q \) \((i = 1, 2, ..., k_q)\). We define the predicate \( P_i^q(x) \) for \( x = d_q(C_j, S) \):

\[ P_i^q(d_q(C_j, S)) = \begin{cases} 
1, & \text{if } d_q(C_j, S) \in H_i^q, \\
0, & \text{if } d_q(C_j, S) \not\in H_i^q.
\end{cases} \]  \hspace{1cm} (10)

Based on (10), the proximity function \( B_q(C_j, S) \) is defined between the class \( C_j \) and the object \( S \) according to the signs \( x_i \) and \( x_q \):

\[ B_q(C_j, S) = \max_{1 \leq i \leq k_q} p(S/ d_q(C_j, S) \in H_i^q), \]

\[ p(S/ d_q(C_j, S) \in H_i^q) = \frac{1}{|\tilde{C}_j \cap H_i^q|} \sum_{S \in \tilde{c}_j} P_i^q(d_q(C_j, S)), \]

where \( |\tilde{C}_j \cap H_i^q| \) – intersection power \( \tilde{C}_j \) and \( H_i^q \).

5. We highlight preferred proximity functions \( B_q(C_j, S) \). As a result of this stage, the preferred proximity functions are determined among all \( B_q(C_j, S) \). We consider finding the preferred proximity function based on the dominance score:

\[ \mathcal{D}_q = \sum_{j=1}^{l} \left( |\tilde{C}_j| \sum_{S \in \tilde{c}_j} B_q(C_j, S) \right) \left( |\tilde{C}_j| \sum_{S \in \tilde{c}_j} B_q(C_j, S) \right) \]

The larger the value of \( \mathcal{D}_q \), the more preference is given to \( B_q(C_j, S) \). If several proximity functions receive the same preference, then any one of them is selected.

The set of preferred proximity functions is defined as a subset of cardinality \( n' \). As a result of this stage, we obtain a set of preferred proximity functions:

\[ \mathcal{N} = \{S_1, ..., S_{n'}, ..., S_1\}, S_j = \{B_1(C_j, S), ..., B_u(C_j, S), ..., B_u(C_j, S)\}. \]

At this stage, the set of preferred proximity functions is determined for each \( \tilde{C}_j \), which is denoted by \( \tilde{P}_j \). Each set of preferred proximity functions characterizes only one subset (class) of objects.

6. Evaluation for a class based on the aggregate of preferred proximity functions. The estimation of the belonging of the object \( S \) to the class \( C_j (j = 1, l) \) is calculated by the operator \( \mathfrak{B}(S) \) \([3, 4]\):

\[ \mathfrak{B}(S) = \left( \mathfrak{B}_1(S), ..., \mathfrak{B}_l(S), ..., \mathfrak{B}_l(S) \right), \]

\[ \mathfrak{B}_j(S) = \mathfrak{B}(C_j, S) / \sum_{i=1}^{n'} y_v \mathfrak{B}(C_v, S), \mathfrak{B}(C_j, S) = \prod_{u=1}^{n'} B_u(C_j, S), \]

where \( y_v \) – algorithm parameter.

Thus, we have defined a class of modified ROs based on the evaluation of estimates. Any algorithm \( \mathfrak{B} \) from this model is completely determined by specifying a set of parameters \( \tilde{\mathfrak{n}} \). The set of all RAs from the proposed model is denoted by \( \mathfrak{B}(\tilde{\mathfrak{n}}, S) \). The determination of the best RO in the framework of
the considered model is performed in the parameter space $\tilde{T}$. The best operator $\mathfrak{B}(\tilde{T}_0, S)$ is selected based on the search for the minimum value of the quality functional of the RO [3, 18]:

$$\mathfrak{R}(\tilde{T}, S^m) = \Theta \left( \mathfrak{B}(\tilde{T}, S^m) \right) / |S^m|,$$

$$\Theta \left( \mathfrak{B}(\tilde{T}, S^m) \right) = \left( \sum_{S \in S^m} \Omega(\|\tilde{a}(S) - \mathfrak{C}(\mathfrak{B}(\tilde{T}, S))\|) \right), \quad \Omega(x) = \begin{cases} 1, & \text{if } x = 0; \\ 0, & \text{if } x \neq 0, \end{cases}$$

where $\|\cdot\|$ – norm of a boolean vector.

To test the performance of the considered model of ROs, it is necessary to present experimental studies. The following is a description of the experimental studies.

5. Experiment and Results
An experimental study of the performance of the proposed model of RAs is carried out by the example of solving model and practical problems. The following models of RAs were chosen: the classical model of RAs based on the calculation of estimates ($\mathfrak{A}_1$) [3, 4]; and the model ($\mathfrak{A}_2$), proposed in this work. A comparative analysis of the listed models of RAs for solving the considered problem was carried out according to the accuracy of recognition of objects of the control sample (accuracy means the ratio of the number of correctly classified objects to the total number of objects in the sample).

In all experiments conducted in order to determine the accuracy of recognition in solving the considered problems, the initial sample (i.e., data on recognition objects) is divided into two parts (training and control samples). This division is carried out using the method of sliding control [19, 20].

5.1. Model problem
For experimental verification of the developed model of RAs, a model problem has been formed. The source data about the objects to be recognized for the model task is generated in the space of dependent features. The number of classes in this experiment is three. The size of the initial sample is 900 realizations (300 realizations for objects of each class). The number of features in the model example is 250. The number of SIF is 10.

Ten experiments were performed, i.e. Samples with a given covariance matrix $C$ and expectation $E$ were simulated 10 times. The results of solving the considered model problem using $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are given in table 1 (in the process of control).

A comparative analysis of the results of the work of these RAs shows (see table 1) that when using $\mathfrak{A}_2$, on average, 84.15 objects out of 90 are correctly recognized, which is just over 93%. The accuracy of object recognition for the problem under consideration by the proposed algorithm is 10% higher than $\mathfrak{A}_1$. When modifying estimation algorithms, the value of this indicator changed from 82.3% to 93.5%.

| RA   | The average accuracy of all experiments | The ratio of the number of correctly classified objects to the total number of objects of the control sample (%) |
|------|----------------------------------------|------------------------------------------------------------------------------------------------------------------|
| $\mathfrak{A}_1$ | 74.07 | 82.3 |
| $\mathfrak{A}_2$ | 84.15 | 93.5 |

5.2. Practical tasks
As a practical task, we consider the problem of predicting hidden mineralization. It is known that the pace of stable economic development in many countries of the world, including Uzbekistan, is directly dependent on the sources of mineral raw materials and fuels they possess. To successfully forecast the
future development of almost all sectors of the national economy, it is required, on the one hand, to
forecast the volume of demand for raw materials and fuel, and on the other hand, to have information
on which mineral resources, fuel and energy resources operating enterprises and in the newly developed
regions). Therefore, issues related to the formation of a strategy for the future development of the
mineral resource base are a vital task.

When compiling a geological forecast map of the ore area under study, a large number of factors are
taken into account. Due to the large variety of factors, a quick and objective determination of their
nature, strength of connection and importance is possible only with the use of effective methods for
analyzing geological data in large volumes.

In order to study the practical possibilities of using \( \mathcal{A}_1 \) and \( \mathcal{A}_2 \) to solve the problem of predicting
useful fields, experimental studies have been carried out. For prediction of blocks by degree of
perspectivity, uniform data was compiled, consisting of 94 objects (geological positions). The number
of classes in this experiment is two. In the first class there are 31 objects (the most promising blocks),
and in the second - 63 (unpromising blocks). Each object is described by 22 signs - the results of
chemical analysis of the composition of the soil sample of the earth. All these signs characterize the
proportion of the content of a certain set of metals in the sample. The place of sampling, the type of
terrain and the tectonic location are also indicated.

The results of solving the problem with the use of \( \mathcal{A}_1 \) and \( \mathcal{A}_2 \) on the control sample are given in
table 2. It should be noted that in this sample there are always 10 objects. The ratio of objects belonging
to different classes, the control sample after every two checks (experiments) changes, i.e. in two
experiments, the control sample consists of 3 objects of the first class and 7 objects of the second class,
and in the third experiment - of 4 objects of the first class and 6 objects of the second class.

| RA   | The average number of correctly classified test sample objects | The ratio of the number of correctly classified objects to the total number of objects of the control sample (%) |
|------|-------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| \( \mathcal{A}_1 \) | 8.20                                                        | 81.0                                                                                                       |
| \( \mathcal{A}_2 \) | 9.47                                                        | 94.7                                                                                                       |

Comparison of these results shows (see table 2) that the proposed model of RAs \( \mathcal{A}_2 \) has improved
the accuracy of recognition of objects described in the space of interrelated features (more than 10%
higher than that of \( \mathcal{A}_1 \)). This is explained by the fact that the \( \mathcal{A}_1 \) model does not take into account the
interconnectedness of features. However, for model \( \mathcal{A}_2 \) there is a slight increase in learning time due to
the implementation of an additional procedure for the formation of independent SIF.

6. Conclusion
Analysis of existing literature sources showed that many methods and algorithms are mainly focused on
solving applied problems of object recognition with independent features. However, in the majority of
applied recognition problems encountered in science, technology and production, the considered images
are characterized by interrelated features. When solving such problems, the assumption of independence
of signs is often not fulfilled. Although various algorithms for solving recognition problems for such
images were developed, they turned out to be ineffective in terms of accuracy and computational
complexity.

A statistical approach is proposed for constructing a model of ROs under the condition of
interconnectedness of features and on the basis of this approach a model of modified ROs based on the
calculation of estimates is constructed. The results of solving a model problem showed that the proposed
model of ROs improves the accuracy in recognition an unknown object. It can be used in the compilation
of various software systems focused on solving applied problems of recognition of objects specified in
the space of interrelated features.
References

[1] Zhuravlev Yu I 1989 On the application of algebraical techniques in pattern recognition and classification problems *Pattern recognition, classification, forecasting Issue 1* (Moscow: Nauka) 9-16

[2] Kononyuk A E 2012 *General theory of recognition. Vol. 1, Basics* (Kiev: Osvita Ukraine)

[3] Zhuravlev Yu I 1998 *Selected scientific works* (Moscow: Master)

[4] Zhuravlev Yu I, Kamilov M M, Tulyaganov Sh E 1974 Algorithms for calculating estimates and their applications (Tashkent: Fan)

[5] Shinmura S 2017 *New Theory of Discriminant Analysis after R. Fisher* (New York: Springer)

[6] Fomin Ya A 2012 *Pattern Recognition: Theory and Application* (Moscow: FASIS)

[7] Duda R, Hart P, Stork D 2001 *Pattern Classification* (New York: John Wiley)

[8] Dougherty G A 2013 *Pattern Recognition and Classification* (New York: Springer)

[9] Merkov A B 2011 *Pattern Recognition: An Introduction to Statistical Training* (Moscow: Editorial URSS)

[10] Lbov G S, Berikov V B 2005 Stability of crucial functions in problems of pattern recognition and analysis of heterogeneous information (Novosibirsk: Publishing house of the Institute of mathematics)

[11] Ryazanov V V 2007 Logical laws in recognition problems (parametric approach) *Journal of computational mathematics and mathematical physics* 47(10) 1793-1808

[12] Kudryavtsev V B, Andreev A E, Hasanov E E 2007 *The theory of test recognition* (Moscow: Fizmatlit)

[13] Dyakonov A G 2010 The theory of equivalence systems for describing algebraic closures of a generalized model for calculating estimates *Journal of Computational Mathematics and Mathematical Physics* 50(2) 388–400

[14] Mirzaev N and Saliev E 2018 Recognition Algorithms Based on Radial Functions *Proc. of the 3rd Russian-Pacific Conf. on Computer Technology and Applications* (Vladivostok: FEFU) 367-372 [Online]. Available: http://www.proceedings.com/41252.html [Accessed Mar. 12, 2019]

[15] Kamilov M, Fazilov S, Mirzaev N and Radjabov S 2012 Estimates calculations algorithms in condition of huge dimensions of features’ space *Proc. of The 4th Int. Conf. on Problems of Cybernetics and Informatics* (Baku: Institute of Cybernetics) 181-184 [Online]. Available: http://www.proceedings.com/17642.html [Accessed Sen. 14, 2018]

[16] Mirzaev N and Saliev E 2018 Recognition Algorithms Based on Radial Functions *Proc. of the 3rd Russian-Pacific Conf. on Computer Technology and Applications* (Vladivostok: FEFU) 367-372 [Online]. Available: http://ceur-ws.org/Vol-2098/ [Accessed Nov. 9, 2018]

[17] Dobson A J and Barnett A G 2018 *An Introduction to Generalized Linear Models* (Boca Raton: CRC Press)

[18] Weisberg S. 2014 *Applied Linear Regression* (New York: Wiley)

[19] Fazilov Sh Kh Mirzaev N M Radjabov S S and Mirzaev O N Determining of Parameters in the Construction of Recognition Operators in Conditions of Features Correlations *Proc. of 7th Int. Conf. On Optimization Problems and Their Applications (OPTA-2018)* 118-133.

[20] Braga-Neto U M and Dougherty E R 2016 *Error Estimation for Pattern Recognition* (New York: Springer)

[21] Mehryar M, Rostamizadeh A and Talwalkar A. 2018 *Foundations of Machine Learning* (Boston: The MIT Press)