STUDY OF THE FEATURES OF A MULTI-ELEMENT CLASSIFIER FOR UNOBSERVABLE CLASSES BASED ON A GAUSSIAN MIXTURE DISTRIBUTION MODEL

O. Ja. Kravets (1), O. Yu. Zaslavskaya (2), P. V. Bespalov (1), Yu. V. Redkin (3), O. O. Varlamov (4)*

(1) Voronezh State Technical University, Voronezh; (2) Moscow City University, Moscow; (3) Admiral Ushakov Maritime State University, Novorossiysk; (4) Bauman Moscow State Technical University
Russian Federation

* Corresponding Author, e-mail: csit@bk.ru

Abstract: The article is devoted to the study of the features of a multi-element classifier based on a model of a Gaussian mixture of distributions. The relevance is determined by the widespread development of intelligent data processing systems based on machine learning technologies. The choice of a classifier model based on a Gaussian mixture of distributions is due to its ability to represent a large class of sample distributions. A classifier model based on a Gaussian mixture of distributions is proposed to solve the problem of multiclass image classification. The proposed model is two-level: the level of image preprocessing; the level of multi-element image classification. An algorithm for the functioning of a classifier of a multi-element classification system operating under conditions of incompleteness of classes of local classifiers has been developed.

Key words: the features, multi-element classifier, Gaussian mixture distribution model, machine learning, Federative machine learning, unobservable classes.

1. INTRODUCTION

The development of intelligent data processing systems, which currently form the basis of many information, management and control systems in various subject areas, is inextricably linked to the improvement of methods, mathematical, software, hardware and technologies of machine learning (MO) and machine learning models (MMO) [1]. In particular, MO issues are of high relevance for improving systems that implement the process of multiclass (multi-element) classification and have a distributed architecture.

In traditional classification systems, the machine learning process relies on the centralized collection, storage and distribution of data, which are training and test samples [2]. However, the ever-increasing complexity of these models requires ever-

increasing amounts of analysed data in the process of their training, which significantly complicates the organization schemes of the data centres on the basis of which they are deployed. One of the ways to solve this problem are the models and methods of the Federative machine learning (FMO) [3]. The operation of this type of distributed FMO-based systems is based on the interaction of multiple distributed computing nodes (worker nodes), each of which supports its own local MO (LMO). Thus, the complex learning process is decomposed into many simple processes (the implementation of the "weak learner" strategy), and then their aggregation is carried out, as a result of which the final (global) MO (GMO) model is formed. This model is used as the final one and is used to solve applied problems. This approach is most relevant for solving the problems of multi-element classification, when many LMO implement classification for two or more classes, and an ensembled GMO supports classification tasks for the entire set of classes [4].

The main objective of the article is to study the features of a multi-element classifier based on a Gaussian mixture distribution model.

2. INVESTIGATION OF THE FEATURES OF A MULTI-ELEMENT CLASSIFIER BASED ON A GAUSSIAN MIXTURE MODEL OF DISTRIBUTIONS

Unobservable classes of the LMO multiclass classifier model (predictor matrices $X_{j\times d}$) It can be represented by a Gaussian mixture of distributions (GMM) with R components (1):

$$f_X^{(J)}(X)=\sum_{r=1}^{R}\pi_r^j f_{x|z_j=r}(X), \tag{1}$$

where $x|z_j = r \sim N(\mu_r^{(j)}, \sum_r^{(j)})$ defines the value of the r-th Gaussian component. In this case, z_i is a hidden variable that indicates that the locally confidential data of node j belongs to the r-th Gaussian component (z_j=r), and $\pi_r^j=P(z_j=r)$.

In [5], the properties of the Gaussian mixture of distributions (GMM) model and the classifier based on it were investigated in sufficient detail. In particular, it is determined that, in the general case, it is a weighted sum of R component Gaussian densities:

$$p(x|\lambda)=\sum_{i=1}^{R} w_i g\left(x\middle|\mu_i,\Sigma_i\right), \tag{2}$$

where x is a D-dimensional continuous vector of data (in the framework of the study, classes of the LMO model), w_i, i=1…r is a mixture of weights, and $g\left(x\middle|\mu_i,\Sigma_i\right)$,i=1,...,R is a D-variant Gaussian function determining the components of the Gaussian probability density:

$$g\left(x\middle|\mu_i,\Sigma_i\right)=\frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(x-\mu_i\right)'\Sigma_i^{-1}\left(x-\mu_i\right)\right\}, \tag{3}$$

where μ_i is the average value of the vector, and Σ_i is the covariance matrix.

In this case, the weights of the GMM mixture satisfy the following constraint:

$$\sum_{i=1}^{R} w_i = 1.$$

The complete GMM model is parameterized by mean vectors, covariance matrices, and weights of a mixture of all component densities:

$$\lambda = \left\{ w_i, \mu_i, \Sigma_i \right\}, i = 1,...,R \quad (4)$$

Depending on the type of vectors, the GMM model will have a different configuration. For example, covariance matrices Σ_i can be full or bounded by a diagonal. In addition, the parameters can be shared or related between the Gaussian components, for example, have a common covariance matrix for all components. The choice of a GMM model configuration (number of components, full or diagonal covariance matrices, and parameter relationships) is often determined by the amount of data available to estimate its parameters.

The choice of the GMM classifier model is determined by its ability to represent a large class of sample distributions. It can be assumed that individual densities of GMM components can model a subset of unobserved classes of LMS.

3. THE GMM CLASSIFIER MODEL FOR SOLVING THE PROBLEM OF MULTICLASS IMAGE CLASSIFICATION

The chosen classification method based on the Gaussian mixture of distributions model (GMM-classifier) has a wide range of applications and, within the framework of the study, requires specification for the class of multi-element classification problems to be solved.

The choice of the task of multi-element classification of digital images is determined by two factors relevant from the point of view of the conducted research:

- the growing need to solve the problems of multi-element classification of digital images in various subject areas (results of tomography of human organs, results of aerial surveys of the Earth's surface, results of visual and technical quality control of products, etc.);
- the presence in many subject areas of restrictions related to the confidential nature of classified images.

The proposed GMM classifier model is summarized in Figure 1. A subset of digital images from the CIFAR100 dataset [6], grouped into 10 classes, was selected as the training and test dataset.

The proposed model is two-level:

- the level of image preprocessing;
- the level of multi-element classification of images.

Obviously, being a probabilistic model, the GMM model requires preprocessing of the digital images received at its input. The purpose of such preprocessing is to compare the classified image with a certain set of its features, relative to which the GMM model of the image class is formed.

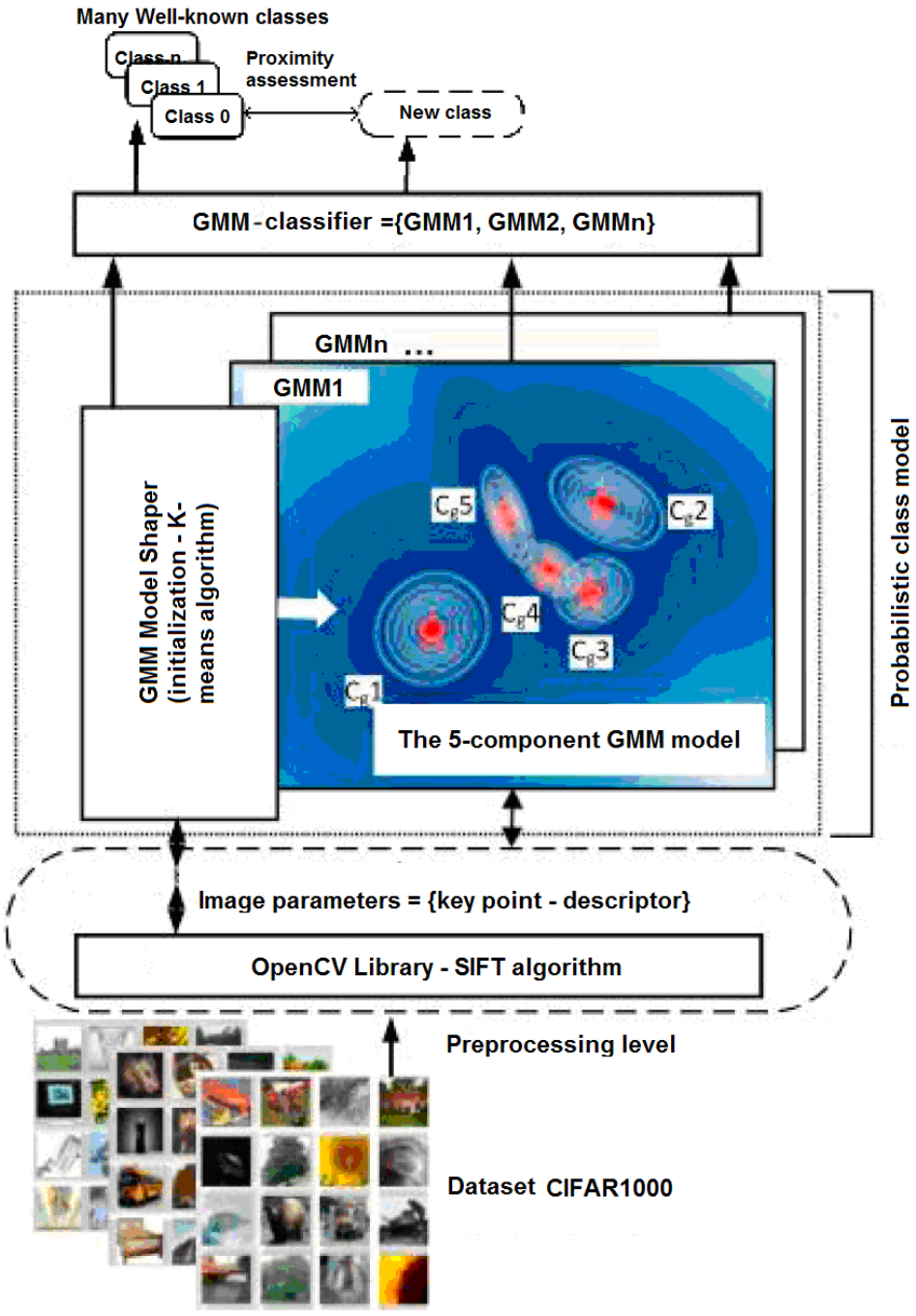


Figure 1. Generalized representation of the proposed GMM image classifier model for nodes of a multi-element classification system

Within the framework of solving the problem of image preprocessing, there are many methods for its implementation. The most famous ones include:

- convolutional autoencoders [7];
- methods for identifying key (special) points [8].

In the study, the Scale-Invariant Feature Transform method for extracting SIFT descriptors, implemented in the SIFT library of the OpenCV computer vision framework, was chosen as the basic preprocessing method [9].

The Figure 1 shows that at the preprocessing level, the SIFT method generates a parametric model for each image from the dataset – a set of parameters of the form "key_point – descriptor" (the key (special) point of the image is the descriptor of the key point). The resulting parametric space is the basis for a probabilistic model of the image class.

Obtaining this model is a two-step process:

1. **Initialization** of the model is the allocation of the centers of parameter clusters using the K-means method.

2. **Formation** of the GMM-model of the class.

It is based on the assumption that it is possible to generate all key points of a parametric space by a mixture of a finite number of Gaussian distributions.

The GMM model is trained with the following hyperparameters:

1. $C_g n$ is the number of GMM components used.

2. The type of covariance matrix of each $C_g n$ from the following set of types:

tied - all $C_g n$ have the same covariance matrix;

diagonal - each $C_g n$ has its own diagonal covariance matrix;

spherical - each $C_g n$ has a single variance;

full - each $C_g n$ has its own common covariance matrix.

3. The non-negative **regularization parameter is added** to the diagonal of the covariance matrix and guarantees the positivity of all covariance matrices.

Thus, the parametric image space is transformed into a set of parametric spaces of GMM models defining a probabilistic model of the image class – $\{GMM_1, GMM_2, \ldots, GMM_n\}$.

To reduce the dimension of this space, a GMM model representative of the class GMM_i is selected. To solve the problem of such a choice, a Bayesian information criterion (BIC) is calculated for each element of the set $\{GMM_1, GMM_2, \ldots, GMM_n\}$ [5]:

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n), \quad (5)$$

where L is the maximized value of the likelihood function of the GMM model, k is the number of parameters of the GMM model, and n is the sample size (number of observations).

The model GMM_i with the lowest BIC value is selected as a symbolic class.

A set of GMM models of class representatives form the class space of the proposed multi-element GMM classifier.

When a model GMM_j with parameters that do not match those of known classes is obtained (the principle of parameter proximity is used), it is considered a model of a new or unobserved class.

4. THE ALGORITHM OF THE CLASSIFIER FUNCTIONING OF A MULTI-ELEMENT CLASSIFICATION SYSTEM OPERATING UNDER CONDITIONS OF INCOMPLETENESS OF CLASSES OF LOCAL CLASSIFIERS

Based on the chosen EM algorithm [10], as a basic method for finding maximum likelihood estimates of the parameters of the GMM model, which represents classes that include data from multiple LMO models of the federated learning system, the study suggests the following approach to implementing a multi-element classifier algorithm that operates under conditions of incompleteness of classes of local classifiers:

1. At the level of LMO classifiers, an EM algorithm is implemented based on publicly available and proprietary data to obtain the probability of assigning unobserved data to a particular class.

2. At the level of the GMO classifier, a cycle of aggregation of parameters of a set of LMO models is implemented with the values of locally confidential data classes obtained during the first stage.

Thus, the proposed algorithm is a combination of two particular algorithms implemented on the nodes Worker (LMO model) and Master (GMO model), respectively.

The scheme of the developed generalized algorithm is shown in Figure 2.

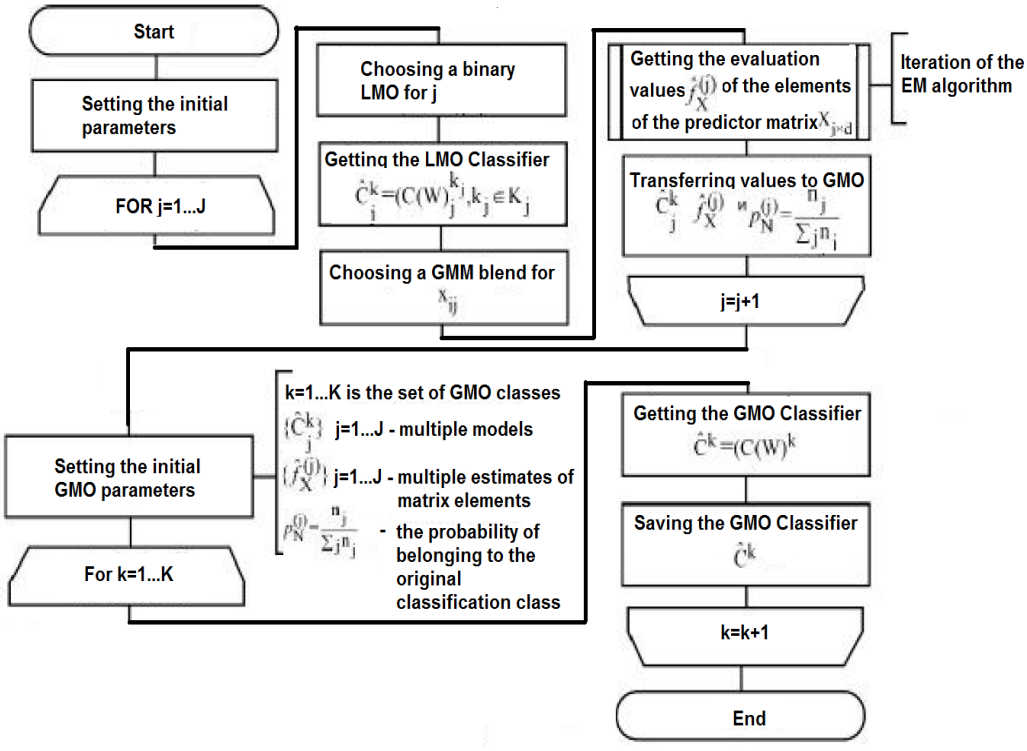


Figure 2. The scheme of the developed algorithm

It can be seen from Figure 2 that many LMO models implement the procedure for forming a local GMM classifier. At the same time, iterations of the EM algorithm are implemented based on publicly available and proprietary local data, providing probabilistic GMM models of new (unobservable by other LMS) classes.

At the GMO level, a cycle of ensembling the parameters of the local classifier set with the obtained values of new (unobservable) classes of LCD data is implemented, and obtaining a GMO classifier.

5. CONCLUSION

In general, the presented algorithm scheme corresponds to the traditional algorithm scheme for the functioning of a federated machine learning system. At the same time, the inclusion of iterations of an algorithm in the stage of forming classifiers of LMO models, which provides estimates of the probabilistic characteristics of unobserved classes of local data, provides the necessary level of completeness of the classifier of the GMO model.

It is obvious that the estimates obtained during the implementation of the EM algorithm are predictive values, which may affect the classification accuracy of the GMO model. However, in comparison with a multi-element classifier operating under conditions of incompleteness of classes, the obtained accuracy value of the modified algorithm will exceed the capabilities of the traditional algorithm due to the formation of a complete matrix of elements.

REFERENCES

[1] Romansky R. Empirical evaluation of the transfer of information resources in active learning. *International Journal on Information Technologies and Security*, vol. 15, no. 1, 2023, pp. 39-48. DOI: https://doi.org/10.59035/WZFU1905.

[2] Scheja B., Rott B. Mathematical tasks: a review of classification systems. *Mathematics Education Research Journal*, vol. 37, no. 3, 2024. DOI: 10.1007/s13394-024-00506-z.

[3] Balta D. et al. Accountable Federated Machine Learning in Government: Engineering and Management Insights. *Lecture Notes in Computer Science*, vol 12849. https://doi.org/10.1007/978-3-030-82824-0_10.

[4] Mansoori T. Multiclass Classification Methods: A Review. *Communications on Applied Nonlinear Analysis,* vol. 33, no. 1s, 2026, pp. 259-269. DOI: 10.52783/cana.v33.6838.

[5] Pribil J., Pribilova A., Matausek J. Experiment with Evaluation of Quality of the Synthetic Speech by the GMM Classifier. *IEICE Transactions on Information and Systems*, vol. E93D, no. 12, 2010, pp. 3368–3376.

[6] Niu, X., Yuan, J., Fan, L. et al. A federated reinforcement learning framework for balancing rapidity and availability in deep space networks. *Journal of King Saud University Computer and Information Sciences,* vol. 38, no. 6, 2026. https://doi.org/10.1007/s44443-025-00392-w.

[7] Shcherbakov O.V., Zhdanov I.N., Lushin Ya.A. A convolutional autoencoder as a generative model of images for problems of distinguishing attributes and restoring images in missing regions. *Journal of Optical Technology*, vol. 82, no. 8, 2015, pp. 528-532. https://doi.org/10.1364/JOT.82.000528.

[8] Shrinivasa S.R., Prabhakar C.J. Secne Image Classification Based on Combining Extended-Local Directional Ternary Pattern and Dense-SIFT Descriptor. *SN Computer Science*, vol. 6, no. 6, 2025. DOI: 10.1007/s42979-025-04210-y.

[9] Introduction to SIFT. https://docs.opencv.org/4.x/da/df5/tutorial_py_sift_intro.html.

[10] Panic B., Klemenc J., Nagode M. Gaussian Mixture Model Based Classification Revisited: Application to the Bearing Fault Classification. *Strojniški vestnik - Journal of Mechanical Engineering,* vol. 66, no. 4, 2020, pp. 215-226.

Information about the authors:

Oleg Jakovlevich Kravets – Dr. Sci (IT), professor of Voronezh State Technical University, areas of scientific research - system analysis, optimization, simulation of complex objects

Olga Yurievna Zaslavskaya - Dr. Sci (Pedagogy), professor of Moscow City University, areas of scientific research – information technology, system analysis, modelling.

Pavel Viktorovich Bespalov– postgraduate student of of Voronezh State Technical University, areas of scientific research - system analysis, optimization, simulation of complex objects

Yuriy Viktorovich Redkin – PhD, associate professor of Admiral Ushakov Maritime State University, Novorossiysk, areas of scientific research - wireless networks, digital signal processing, control and data processing systems

Oleg Olegovich Varlamov – Dr. Sci (IT), professor of Bauman Moscow State Technical University, areas of scientific research - artificial intelligence, decision-making systems, optimization, simulation of complex objects

Manuscript received on 14 June 2026