UDC 519.254:330.47
DOI: 10.20535/SRIT.2308-8893.2023.4.02

METHODOLOGY OF THE COUNTRIES' ECONOMIC DEVELOPMENT DATA ANALYSIS

V.V. DONETS, V.Y. STRILETS, M.L. UGRYUMOV, D.O. SHEVCHENKO, S.V. PROKOPOVYCH, L.O. CHAGOVETS

Abstract. The paper examines the issue of improving the methods of identification of economic objects and their analysis using algorithms of intelligent data processing. The use of the developed methodology in the economic analysis allows for improvement in the quality of management. It can be the basis for creating decision support systems to prevent potentially dangerous changes in the economic status of the research object. In this work, an improved method of c-means data clustering with agent-oriented modification is proposed, and a radial-basis neural network and its extension are proposed to determine whether the obtained clusters are relevant and to analyze the informativeness of state variables and obtain a subset of informative variables. The effect of applying data compression using an autoencoder on the accuracy of the methods is also considered. According to the results of testing of the developed methodology, it was proved that the probability of incorrect determination of the state was reduced when identifying the states of economic systems, and a reduced value of the error of the third kind was obtained when classifying the states of objects.

Keywords: machine learning, digital development, fuzzy clustering, radial basis neural networks, logistic regression, analysis of variables informativeness.

INTRODUCTION

Analysis of the state of economic systems requires taking into account a large number of factors that have a stochastic nature of development and high dynamism. Continuous monitoring allows taking into account the influence of these factors and maintaining the stable functioning of economic systems in conditions of constant global fluctuations [1]. Machine learning methods make it possible to evaluate these factors, their possible and real impact on macroeconomic processes. The use of machine learning algorithms provides early consideration of the effects of factors that may threaten the stability of economic systems [1].

The use of intelligent methods for the analysis of collected economic data allows to automate the solution of many problems in the management of economic processes [1], which significantly increases its quality and efficiency. Automated systems of economic analysis are used as decision support systems to prevent potentially dangerous changes in the state of economic systems [1; 2]. Existing information systems of economic analysis have modules for solving problems of clustering, classification or forecasting of received data, based on machine learning methods, which allow to improve the accuracy of received decisions.

The aim of the study is to improve the quality of data stratification in the information analysis of economic systems by developing a methodology that includes methods of clustering, classification and analysis of the informativeness

of economic data. The scientific objective of the study is to improve the existing methods of economic data analysis through the introduction of an agent-oriented modification of the clustering method and radial basis neural networks for analyzing the informativeness of state variables. The proposed methods are expected to reduce the probability of erroneous determination of the state in the analysis of the economic system, thus the value of the third-order error in the classification of its state will be reduced.

STATEMENT OF THE RESEARCH PROBLEM

The data obtained as a result of the study of the economic system can be presented in the form:

$$X = \{x_{im}\},$$

where $i = \overline{1,N}$, $m = \overline{1,M}$, X — matrix representing the data sample for analysis; N — number of objects; M — dimension of space.

The problem of data analysis that characterizes the state of the economic system consists of solving a sequence of problems:

– division of a set of data into sets that are similar according to certain characteristics — the task of clustering;

– determination of the current state of the economic system based on a set of characteristics — the task of classification;

– determination of a set of features that best describe the state of the economic system — the task of selecting informative features (reduction of the space of features).

Let's consider the methods of solving each of the problems.

FUZZY DATA CLUSTERING METHOD

For some known set of valid clusters Y it becomes necessary to split the input data X to $|Y|$ subsets (clusters, classes), so that each cluster consists of objects that are close by some metric, or distant by another. Thus, each object will be assigned to the *y*-th cluster.

The result of the clustering algorithm [3] will be the application of the function $cluster : X \to Y$, which matches each object in the input set $x \in X$ matching an object from a set of clusters. Usually, plural $y \in Y$ known in advance for a non-hierarchical approach, or determined in the process for a hierarchical approach. Therefore, the question of determining the optimal number of clusters, as one of the parameters determining the final quality of clustering, often arises.

Let's define the distance between cluster objects as a metric for cluster analysis. Then we define the degree of similarity of objects as the reciprocal of the inter-element distance. Among the works devoted to cluster analysis, can be found a large number of possible metrics for determining the inter-element distance or degree of similarity. The most widespread metric is based on the Euclidean distance, which is a special case of the Minkowski distance [4] with the value of the parameter $\varepsilon = 2$. Generalized Minkowski metric:

$$d_{\varepsilon}(x_i, x_j) = \sqrt[\varepsilon]{\sum_{m=1}^{M} \left| x_{im} - x_{jm} \right|^{\varepsilon}} .$$

The c-means fuzzy clustering method allows fuzzy distribution of objects into clusters or classes. In the c-means method, the object belongs to all clusters, but with a certain value of cluster membership [5].

In the method of fuzzy clustering [6], the membership matrix of elements to a cluster is calculated according to the assumption of a normal distribution of data according to the formula:

$$w_{ij} = \frac{N(d(x_i, c_j) \mid \mu = 0, \sigma_j)}{\sum_{i=1}^{P_j} N(d(x_i, c_j) \mid \mu = 0, \sigma_j)},$$

where x_i — i-th element of the set, $i = (1; P_j)$; c_j — j-th cluster center; $d(x_i, c_j)$ — distance between points x_i and c_j; $N(d(x_i, c_j) \mid \mu = 0, \sigma_j)$ — probability density of a normal distribution at a point $d(x_i, c_j)$.

The cluster centers are adjusted according to the formula

$$c_j \frac{\sum_{i=1}^{P_j} w_{ij} x_i}{\sum_{i=1}^{P_j} w_{ij}} . \qquad (1)$$

The center adjustment process continues until the loss function is minimized:

$$loss = \sum_{j=1}^{K} \sum_{i=1}^{P_j} d(x_i, c_j)^2 w_{ij} \to \min , \qquad (2)$$

or on the condition of reaching some limitation on the number of iterations, or the required classification quality.

Among the important disadvantages of the c-means method are the inability to divide the space with a complex shape of target clusters that go beyond simple *M*-dimensional spheres, and an insufficient level of robustness to noise [5; 7].

For data from real problems, both a complex distribution of object parameters and a high dimensionality of the input data are inherent, which in turn determines the complex form of *M*-dimensional target clusters. Therefore, for the usual method of fuzzy clustering and many of its modifications, clustering with high accuracy is not possible. A modification of the distance metric (together with the membership metric) is proposed in [8]. An interesting approach is the assumption of the Cauchy distribution and the use of the Mahalanobis distance, which were proposed in [9; 10]. Mahalanobis distance was used to improve the calculation algorithm that prevents degeneracy of the inverse matrix [11]:

$$MD(x_i, c_j) = \sqrt{(x_i - c_j)^T \hat{\Sigma}_j^{-1} (x_i - c_j)} ,$$

where $\hat{\Sigma} = \Sigma + \lambda\Sigma$ — is the regularized covariance matrix; λ — is a constant greater than zero.

Taking into account the assumption of the Cauchy distribution in the data, the expression for calculating the value of belonging to a certain cluster [5] has the form:

$$w_{ij} = \frac{\rho(x_i, c_j)}{\sum_{i=1}^{P_j} \rho(x_i, c_j)}, \quad \rho(x_i, c_j) = \left(\pi\eta \left[1 + \frac{MD^2(x_i, c_j)}{\eta^2} \right] \right)^{-1}. \tag{3}$$

Solving the clustering problem for clusters of complex *M*-dimensional form using Gaussian mixture models was considered in works [12; 13], using the derivative in [14] and using the Mahalanobis distance in [5; 9]. According to the obtained results, an improvement in clustering accuracy is noted, but the problem of spatial separation and overuse of input data dependence occurs.

In works [5; 15], the possibility of taking into account the relative entropy of the data distribution was considered when using the c-means method, but the Euclidean distance was chosen as the metric of the distance between the objects of the sample, which reduced the computational load, but did not take into account the entropy of the data.

To overcome the difficulties of using the basic method of fuzzy clustering and its modifications based on Mixture and Gaussian mixture models on data with a complex shape of *M*-dimensional target clusters [12], which is based on an attempt to take into account the entropy of clusters [15] and the Kullback–Leibler distance [16], it was proposed to improve the clustering method.

The Kullback–Leibler distance is an asymmetric measure of the informational difference between two probability distributions. This measure has proven itself well in methods of information processing in physical systems and statistics [16].

According to the previous definition $x_{im} \in X$ — is the *m*-th state variable of the *i*-th vector of the input data sample, where $m \in [1, M]$, M — dimension of the state vector. Let's define $f_s \in F$ as the *s*-th object function from the vector of object functions $s \in [1, S]$, where S — the dimension of the object functions vector. Then $M_\alpha(f_s)$ and $M_\alpha(x_{im})$ are mathematical expectations of f_s and x_{im} respectively. According to this definition $D(f_s)$ and $D(x_{im})$ — dispersion of the relevant variables, and $\sigma(f_s)$ and $\sigma(x_{im})$ — standard deviation. Variance and standard deviation of conditional dependence of f_s from x_{im} an be determined by formulas:

$$D(f_s \mid x_{im}) = var(M_\alpha(f_s(x_{in-m}))), \ \forall n, \ n \neq m, \ x_{in} = const; \tag{4}$$

$$\sigma(f_s \mid x_{im}) = \sqrt{D(f_s \mid x_{im})}. \tag{5}$$

Using expression (4), we obtain estimates of informative state variables:

$$\beta(f_s) = \frac{D(f_s \mid x_{im})}{E(f_s)},$$

where $E(f_s)$ — signal energy.

From (5), we get the influence coefficient (signal to noise ratio):

$$\varphi_{sm} = SNR(f_s \mid x_{im}) = \frac{\sigma(f_s \mid x_{im})}{\sigma(x_{im})}.$$

In [16], the Kullback–Leibler entropy is defined as follows:

$$D_{KL}(f_s, x_i) = \sum_{m=1}^{M} \rho(x_{im} \mid f_s) \log_2 \left(\frac{\rho(x_{im} \mid f_s)}{\rho(x_{im})} \right).$$

Mutual informative dependence is then determined by the formula:

$$H_{sm} = \frac{1}{2} \log_2 SNR^2(f_s \mid x_{im}) = \frac{1}{2} \log_2 \left(\beta(f_s) \frac{E(f_s)}{D(x_{im})} \right).$$

In the proposed method, we replace the loss function (2). Instead, we will get a formula for determining mutual informative dependence, which will be a function of clustering quality assessment, that is, a function of losses in the developed method of fuzzy clustering:

$$H(X,Y) = -\frac{1}{\sum_{j=1}^{k} P_j} \sum_{j=1}^{k} \left[P \left(Y_j^{(t+1)} \times \sum_{i=1}^{P_j} D_{KL}(x_i, Y_j^{(t+1)})` \right) \right] \to \min,$$

where Y_j — state variables belonging to the *j*-th cluster.

AGENT-ORIENTED MODIFICATION OF THE CLUSTERIZATION METHOD

To overcome the non-priority problem, an agent-oriented modification was developed for the classical method of fuzzy clustering considering the *M*-dimensional spatial shape [3; 5], which is considered below.

Let`s introduce special notations for the developed method of fuzzy clustering: X — agents, elements of the input sample, C — centers of clusters, then X_i — agents, cluster elements, Z — agents clusters. According to the agent-oriented approach, the elements-vectors of the input sample and the clusters are agents, these agent-elements choose the cluster agents closest to them, which they join according to a pre-specified metric, thus forming cluster agents. The number of cluster agents is determined by minimizing the loss function. According to the previous definition: the input sample partitioned into clusters is $X = \{P_j\}$, where $j \in (1, K)(1,\ K)$, $N = \sum_{j=1}^{K} \left| P_j \right|$ — the number of elements in the input sample; P_j — set of elements belonging to the *j*-th cluster; K — number of clusters. Than $x_{ij} \in P_j$ — the *i*-th element of the *j*-th cluster.

Four metrics were chosen to compare the possibilities of spatial separation of clusters and computational efficiency:

$$d(x_{ij}, c_j) = \begin{cases} d_1(x_{ij}, c_j), \\ w_{ij}^{-1} d_1(x_{ij}, c_j), \\ -D_{KL}(x_{ij}, c_j), \\ p(x_{ij}, c_j^{t-1}) * \log_2 p(x_{ij}, c_j^{t}), \end{cases} \tag{6}$$

where $d_1(x_{ij}, c_j)$ — Manhattan distance; $w_{ij}^{-1} d_1(x_{ij}, c_j)$ — Mahalanobis distance with the inverse of the membership function; $-D_{KL}(x_{ij}, c_j)$ — Kullback–Leibler divergence; $p(x_{ij}, c_j^{t-1}) * \log_2 p(x_{ij}, c_j^{t})$ — cross entropy.

Having the distance to determine the inter-element distance, we will get an expression for determining the cost function for each cluster, that is, the average measure of the intraclass distance:

$$cl\_loss(P_j) = \frac{1}{|P_j|}\sum_{i=1}^{|P_j|} d(x_{ij}, c_j). \quad (7)$$

Then, using expression (7), we obtain the general cost function for evaluating the current quality of clustering:

$$loss(X^t) = \frac{1}{K^t}\sum_{j=1}^{K^t} cl\_loss(P_j). \quad (8)$$

By combining the classical method of fuzzy clustering with the agent-oriented approach described above, we will obtain a statement of the research problem, according to which it is necessary to determine the number of clusters and such a distribution of elements by clusters that the value of the cost function is minimal:

$$\begin{cases} A = [K^t, X^t], \\ \hat{A} = \arg\min(loss(X^t)). \end{cases}$$

According to the classical clustering method, cluster centers are optimized according to expression (1), and the membership matrix for adjustment is calculated according to expression (3) taking into account the Cauchy distribution assumption. We formulate the clustering algorithm, defined according to the agent-oriented approach, as follows:

1. Determine some initial number of cluster agents $K^t > K$, that is more than the target number of clusters, and set a limit on the number of elements in each cluster $|P_j^t| = N/K^t$ and choose randomly K^t centers of clusters $\{c_j\}$.

2. Select one of the inter-element distances (6) $|P_j^t|$ of the closest elements to each cluster, that is, to form cluster agents P_j^t.

3. For each cluster, calculate the value of the parameters $\rho(x_{ij} | P_j^t)$ distribution and the values of the membership matrix according to expressions (3), and according to expression (1) adjust the cluster centers.

4. To each center of the cluster according to the selected measure $d(x_{ij}, c_j)$ to choose $|P_j^t|$ new agents-elements.

5. For each cluster agent, according to expression (7), determine the value of the cost function (or the average inter-element distance) $cl\_loss\left(P_j^t\right)$.

6. To estimate the current quality of clustering by the loss function according to expression (8). In the case of the operation mode of the algorithm in the automatic search for the optimal number of clusters, and the increase in the value of the cost function, stop the algorithm.

7. To select agent-clusters and discard the agent-cluster with the highest value $cl\_loss(P_j^t)$.

8. To determine the new number of clusters $K^{t+1} = K^t - 1$ and the new number of cluster elements $|P_j^{t+1}| = N / K^{t+1}$.

9. Return to stage 2, if $K^t > K$.

CLASSIFICATION METHOD BASED ON MULTIPLE LOGISTIC REGRESSION

To solve the problem of multiclass classification in the case of spatially separated data, it is proposed to use a radial basis neural network (RBFN) with multiple logistic regression. The application of the RBFN model for multiclass classification will allow checking the assumptions about the correctness of the cluster definition and testing the model's ability to generalize.

RBFN structure: H_0 inputs for each of the parameters, H_1 neurons of the first layer and H_2 output neurons. We define the vector of input data for the *k*-th layer of the neural network (or the vector of output data for the *k*-1 layer) as $\vec{Y}^{(k)} = [Y_1^{(k)}, \ldots, Y_{H_1}^{(k)}]^T$, we define the vector of coordinates of the cents of the activation function for the hidden layer as $\vec{c}_j = [c_{j1}, c_{j2}, \ldots, c_{jH_0}]^T$, where $j = 1..H_1$, and the vector specifying the window width of the activation function of the *j*-th neuron of the hidden layer is defined as $\vec{\sigma}_j = [\sigma_{j1}, \sigma_{j2}, \ldots, \sigma_{jH_0}]^T$. Then the activation function for the neurons of the hidden layer will look like this:

$$\varphi_j = (\vec{Y}_p^{(0)}, \vec{c}_j, \vec{\sigma}_j) = exp\left(-\frac{1}{2} \sum_{h=1}^{H_0} w_{ij} Z_{pjh}^2 \right) \equiv \varphi_{pj},$$

where $Z_{pjh} = \dfrac{Y_{ph}^{(0)} - c_{jh}}{\sigma_{jh}}$; w_{ij} — weighted connection between the *i*-th neuron of the output layer and the *j*-th neuron of the input layer.

Multiple logistic regression [17] is used as the activation function of the output layer, the outputs of which are defined as:

$$\vartheta_j = \frac{\exp(\gamma_j)}{\sum_{k=1}^{H_2} \exp(\gamma_k)}, \text{ де } \gamma_j = \sum_i^{H_1} \varphi_i w_{ij}.$$

A hybrid algorithm was used for training the RBFN, which includes 2 steps, the repetition of which usually leads to fast training of the network, especially if the parameters are successfully generated [18]:

1) selection of linear network parameters (weights) using the pseudo inversion method;

2) optimization of nonlinear parameters of activation functions (window centers and widths).

If there are *P* training pairs $(\vec{Y}_p^{(0)}, \vec{d}_p)$, $p = 1..P$ and fixing the specific values of the centers and window widths of the activation functions, we get a system of equations:

$$\Phi \vec{w}_i = \vec{d}_i,\ i = 1..H_2,$$

where $\Phi = [\varphi_{pj}]$, $p = 1..P$, $j = 0..H_1$, $\varphi_{p0} = 1$, $\vec{w}_i = [w_{i0}, w_{i1}, ..., w_{iH_1}]^T$, $\vec{d}_i = [d_{0i}, d_{1i}, ..., d_{pi}]^T$.

Vector $\vec{w}_i$ can be determined in one step using pseudo matrix inversion Φ: $\vec{w}_i = \Phi^+ \vec{d}_i$, which in practice is calculated using the decomposition of eigenvalues.

At the second stage of the algorithm, when fixing the weights, the excitation signal passes through the network to the initial level, which allows to calculate the error value for the sequence of vectors $\{\vec{Y}_p^{(0)}\}$. After that, there is a return to the hidden layer. The gradient vector of the selection function according to the specific variable cents and window widths is determined by the error value: $\| \vec{Y}^{(2)} - \vec{d} \|_{L_2}$.

Algorithm for forming the "coverage zone" by radial basis functions of *k*-neighbors $\sigma_{jh}^2 = \Sigma_j = \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H_0} (c_{jh} - c_{kh})^2$, $k = 1..K$, $K \in [3,5]$ was used to determine the values of the window widths, which helped reduce the training time of the RBFN.

CHARACTERISTICS INFORMATIVENESS ANALYSIS METHOD

Since it is proposed to use the RBFN network to solve the classification problem, this model can also be used to find the minimum possible subset of informative variables. The input data set can be represented as a Taylor series, keeping only the terms of the first infinitesimal order. For the variance of an arbitrarily obtained linear function of several random variables, the estimate is valid:

$$D_{Y_i} = (\operatorname{grad} Y_i)^T \Sigma_S \operatorname{grad} Y_i = \sum_{j=1}^{J} \left(\frac{\partial Y_i}{\partial s_j} \right)^2 \sigma_{S_j}^2 + \sum_{j=1}^{J} \sum_{l=1, l \neq j}^{J} r_{jl} \frac{\partial Y_i}{\partial s_j} \frac{\partial Y_i}{\partial s_l} \sigma_{S_j} \sigma_{S_l},$$

where Σ_S — covariance matrix of variables $S_1; S_2$, σ_{S_1} — standard deviation; r_{j1} — correlation coefficient between variables S_1 and S_2.

Then the standard deviation and variance of the RBFN output can be estimated according to the architecture chosen for it, and from them determine the energy of the signals by the expression [18]:

$$E_i = \sum_{h=1}^{H_0} \left| D_{Y_i^{(2)} | Y_h^{(0)}} \right|,$$

where $D_{Y_i^{(2)} | Y_h^{(0)}} = \left(\frac{\partial Y_i^{(2)}}{\partial Y_h^{(0)}} \right)^2 \sigma_{Y_h^{(0)}}^2 + \left(\sum_{n=1, n \neq h}^{H_0} r_{hn} \frac{\partial Y_i^{(2)}}{\partial Y_n^{(0)}} \sigma_{Y_n^{(0)}} \right) \frac{\partial Y_i^{(2)}}{\partial Y_h^{(0)}} \sigma_{Y_h^{(0)}}$.

Then the coefficient of informativeness of the variables (the weight of the contribution of $Y_h^{(0)}$ in to $Y_i^{(2)}$) is defined by the expression:

$$\beta_{ih} = \frac{\left| D_{Y_i^{(2)}|Y_h^{(0)}} \right|}{E_i}.$$

DATA PRE-PROCESSING METHODS

In machine learning problems, it has become common practice to use data pre-processing methods (normalization, cleaning from anomalies, and dimensionality reduction) to improve the quality of problem solving [19]. Three methods of the scikit-learn, Python library were used for data *normalization*:

– *RobustScaler* scales parameters with robustness to statistical outliers.

– *StandardScaler* (*Z*-score normalization). Reduces the mean and scales to unit variance.

– *MinMaxScaler* (min-max normalization). Each parameter is scaled and translated individually by the estimator so that it falls within a given range, for example [0,1].

The detection of unusual elements, events, or observations that are significantly different from the main body of data and do not correspond to a well-defined definition of normal behavior is called the process of anomaly detection [20]. Data cleaning techniques remove values that have been identified as outliers and based on anomaly detection.

Two outlier detection methods from the scikit-learn library were used:

– *Interquartile Range* (IQR). By dividing the data set into quartiles, it is used to measure variability;

– *Isolation forest*. The method uses isolation to find anomalies (how far a data point is from the rest of the data) [21; 22].

The *dimensionality reduction* process aims to provide a lower-dimensional representation of the original data set while preserving its important characteristics. Separate scikit-learn and PyTorch libraries were used for dimensionality reduction. Three methods were used:

– *T-distributed Stochastic Neighbor Embedding* (t-SNE) [23];

– *Principal Component Analysis* (PCA) the method is based on SVD, it reduces the dimensionality of the data well [24].

– *Autoencoder*. Is a certain type of feed-forward neural network where the input matches the output. It compresses the input data into a bottleneck (lower dimensional data) and then reconstructs the output data from that representation. The bottleneck is the target compact summation or dimensionality reduction of the input data, also called the latent space representation.

APPLICATION OF METHODOLOGY FOR COUNTRIES DIGITAL DEVELOPMENT DATA ANALYSIS

The developed methodology was tested to identify the state of digital development of the countries of the world. For the classification (positioning of countries) regarding the level of their digital development, the hypothesis of the existence of

homogeneous groups of countries (objects) according to specialized indices was tested. Indices that fully reflect the state of digital development were selected:

– EGIit — Global E-Government Development Index;

– NRIit — network readiness index;

– ICTit — information and communication technologies development index.

By forecasting independent factors — indicators of digital development based on the model, it is possible to estimate the forecast level of social progress of a specific country. The Social Progress Index (SPI) is a combined indicator of the International Research Project The Social Progress Imperative [25; 26] which measures the achievements of the countries of the world in terms of social well-being and social progress. The authors of the study [25; 26] believe that indicators of social development are often considered as an alternative to indicators of economic development. The global e-government development index [26] is an integral indicator that assesses the readiness and capabilities of national government structures in using information and communication technologies (ICT) to provide public services to citizens. The index of network readiness [26] characterizes the level of development of information and communication technologies and the network economy in the countries of the world. Currently, the index is considered one of the most important indicators of the innovative and technological potential of the countries of the world and their development opportunities in the field of high technology and digital economy. The ICT Development Index is a composite index that combines 11 indicators and is used to monitor and compare the development of information and communication technologies (ICT) between countries.

To implement the model, a sample of 115 precedents (observations by country) was collected for 32 variables of the state of social development for each precedent and the 33rd field for the predictive value of the state. The ratio of values of the social progress index SPIt (Social Progress Index) and the average level of income was used to mark the educational sample. All precedents of the sample were distributed according to the respective states:

– "High income" — 45 precedents (I);

– "Upper middle income" — 11 precedents (II);

– "Lower middle income" — 25 precedents (III);

– "Lower income" — 34 precedents (IV).

For this sample, pre-processing of the data was first carried out: normalization and detection of anomalous values. Clustering was performed for the considered economic data, and classification was performed to verify its results. It was decided to use the Kullback–Leibler distance classification method. As a result of its application, an accuracy of 84.3% was achieved, and the value of the flow function was obtained as 0.0117. A matrix of inconsistencies (Table 1) was also constructed to assess the accuracy of the method, as well as graphs of cost function values (Fig. 1) and ROC curves for each of the classes (Fig. 2).

Table 1. The matrix of inconsistencies in the classification of data indicators of the digital development of the countries of the world

Actual class	Predicted class			
	I	II	III	IV
I	37	1	0	7
II	1	8	1	1
III	2	0	21	2
IV	0	1	2	31

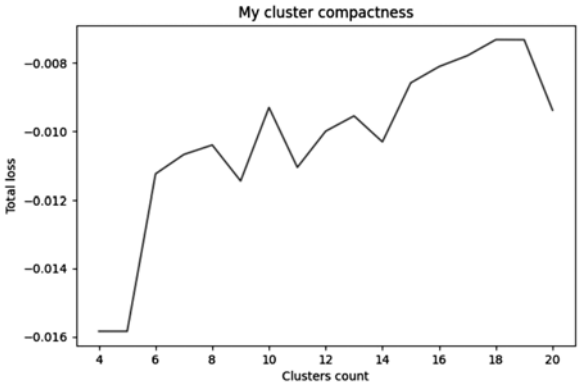

Fig. 1. The ratio of the number of clusters to the value of the cost function for economic indicators of the countries of the world data

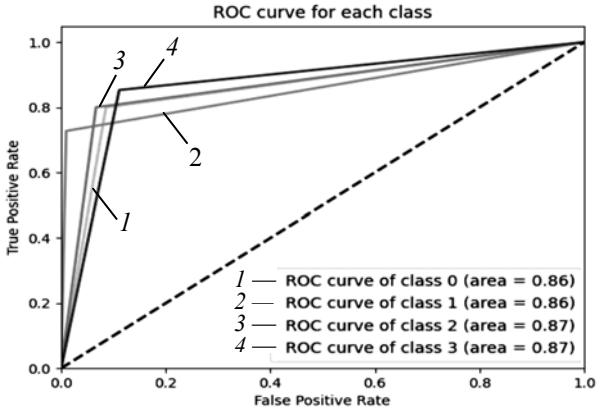

Fig. 2. ROC curves for each of the classes for these economic indicators of the countries of the world

After a series of experiments, it was decided to apply the autoencoder method to reduce the dimensionality of the data with 98% information retention, which made it possible to reduce the dimensionality of 32 to 11 state variables for each case. After this application, an accuracy of 86.9% was achieved, and the value of the cost function became -0.04827. A matrix of inconsistencies (Table 2) was also constructed to assess the accuracy of the method and a graph of the values of the cost function (Fig. 3) and ROC curves for each of the classes (Fig. 4).

Table 2. The matrix of inconsistencies in the classification of compressed data indicators of the digital development of the countries of the world

Actual class	**Predicted class**			
	I	II	III	IV
I	38	0	0	7
II	0	10	0	1
III	1	0	21	3
IV	0	1	2	31

To carry out *multi-class classification* with the help of RBFN, the data of the digital development of countries with a reduced dimension, processed by the autoencoder method, were used. To test the ability of the model to generalize, the data were divided into test and training samples in the ratio of 20% (22 precedents) and 80% (93 precedents), respectively. Previously, the data sample was normalized.

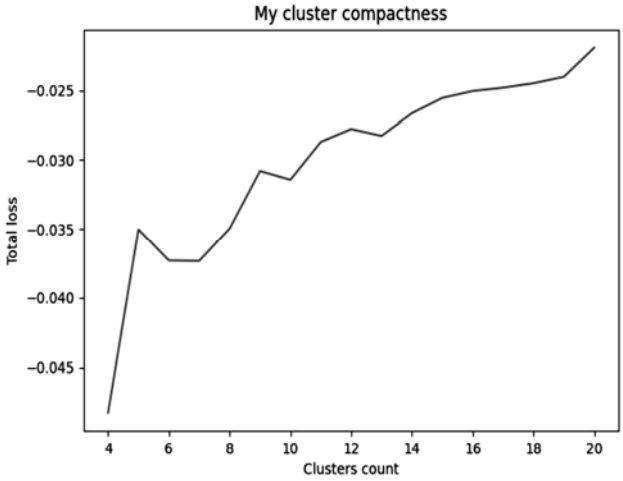

Fig. 3. The ratio of the number of clusters to the value of the cost function for the compressed data of the economic indicators of the countries of the world

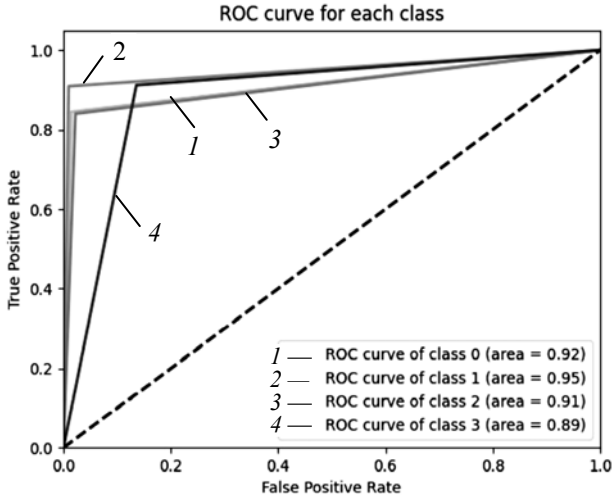

Fig. 4. ROC-curves for each of the classes for compressed data of economic indicators of the countries of the world

RBFN will receive 7 state variables that do not have a defined value at the input, and at the output there will be estimates of state variable values — 4 states. The structure of the proposed RBFN has $H_0 = 7$ inputs for each of the parameters, $H_1 = 90$ neurons of the first layer and $H_2 = 4$ output neurons.

Table 3. Misclassification matrix of the compressed data of the country's digital development indicators of the world

Actual class	Predicted class			
	I	II	III	IV
I	8	0	0	1
II	0	2	0	1
III	0	4	1	0
IV	2	0	0	4

As a result of training on the training sample, an accuracy of 83.87%, while on the test sample — 68.18%. To display the test results, a matrix of inconsistencies was constructed for the training sample (Table 3) and a ROC curve was shown (Fig. 5), which has a smaller coverage area (i.e., worse classification ability), because part of the data was used for training, which reduced the ability of RBFN to generalization.

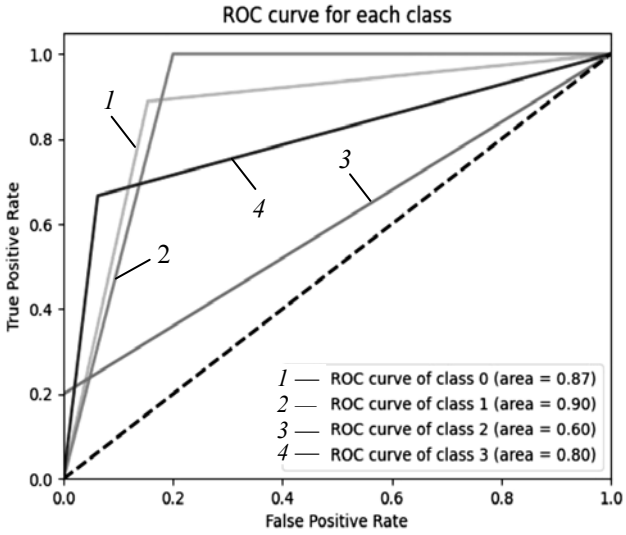

Fig. 5. ROC curves for each of the classes for the PCA test sample of compressed data of indicators of digital development of the countries of the world

An analysis of the sensitivity of the target function was also carried out, i.e. the most informative indicators were determined. The results are shown in Table 4. Based on the results, it can be concluded that a different set of variables is informative for each cluster.

T a b l e 4 . Sensitivity analysis of the variable clusters objective functions

Cluster	Number of precedents	Sensitive cluster variables	Mathematical expectation of the objective function
0	45	TII, ICT, HCI	85.33
1	11	TII, ICT, EGI	52.87
2	25	EPI, HCI, OSI	63.47
3	34	HCI, EPI, EGI	73.60

All numerical studies were carried out using the computer program "Nonlinear estimation methods in the multicriterion problems of system's robust optimal designing and diagnosing under parametric apriority uncertainty (methodology, methods and computer decision support and making system" (ROD&IDS), developed by the authors [27].

CONCLUSIONS

The methods of intelligent data flow processing are widely used during the identification of the states of economic objects. The use of new methods will make it possible to supplement the package of available tools for solving current problems with data processing and will make it possible to increase the stability of the methods to the nature of the data and improve the situation with the use of computing resources.

Presented study examines the problem of improving the methods of classification and clustering of countries according to the state of social and digital development. A multiclass classification method based on radial basis neural networks and a data clustering method based on an agent-oriented modification of the c-means method are proposed.

The proposed RBFN uses multiple logistic regression as the last layer for multiclass classification and the training results of an agent-oriented clustering model as input parameters. The peculiarity of the modification of the c-means method is the introduction of elite selection of clusters.

According to the results of the research, the proposed methodology is proposed to be used for the analysis of economic systems to improve the quality of decision-making, but it should be noted that the method requires a qualitatively prepared sample that covers the largest possible space of input parameters for the target classes.

REFERENCES

1. Mei Yang, Ming K. Lim, Yingchi Qu, Du Ni, and Zhi Xiao, "Supply chain risk management with machine learning technology: A literature review and future research directions," *Computers & Industrial Engineering*, vol. 175, January 2023, 108859. Available: https://doi.org/10.1016/j.cie.2022.108859
2. Benjamin Decardi-Nelson and Jinfeng Liu, "Robust Economic Model Predictive Control with Zone Control," *IFAC-PapersOnLine*, vol. 54, issue 3, pp. 237–242, 2021. Available: https://doi.org/10.1016/j.ifacol.2021.08.248
3. M. Schlesinger and V. Hlavac, *Ten lectures on statistical and structural pattern recognition*. Springer, Dordrecht, 2002. doi: 10.1007/978-94-017-3217-8.
4. *Data clustering: algorithms and applications*, Charu C. Aggarwal and Chandan, K. Reddy (ed.). CRC Press, Taylor & Francis Group, 2014.
5. N. Bakumenko, V. Strilets, and M. Ugryumov, "Application of the C-Means Fuzzy Clustering Method for the Patient's State Recognition Problems in the Medicine Monitoring Systems," *CEUR Workshop Proceedings of 3rd International Conference on Computational Linguistics and Intelligent Systems, COLINS 2019*, vol. I, pp. 218–227, 2019, Available: https://www.researchgate.net/publication/338819685
6. R. Winkler, F. Klawonn, and R. Kruse, "Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets," *Challenges at the Interface of Data Analysis, Computer Science and Optimization*, pp. 79–87, 2012. doi: 10.1007/978-3-642-24466-7_9.
7. Christopher D. Prabhakar Raghavan and Hinrich Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.
8. S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Systems with Applications*, vol. 165, article no. 113856, 2020. doi: 10.1016/j.eswa.2020.113856.
9. Xuemei Zhao, Yu Li, and Quanhua Zhao, "Mahalanobis distance based on fuzzy clustering algorithm for image segmentation," *Digital Signal Processing*, vol. 43, pp. 8–16, Aug 2015. Available: https://doi.org/10.1016/j.dsp.2015.04.009
10. Zarinbala M. Zarandia, M.H. Fazel, and I.B. Turksen, "Relative entropy fuzzy *c*-means clustering," *Information Sciences*, vol. 260, pp. 74–97, 2014. doi: 10.1016/j.ins.2013.11.004.
11. V. Strilets, V. Donets, M. Ugryumov, R. Zelenskyi, and T. Goncharova, "Agent-Oriented data clustering for medical monitoring," *Radioelectronic and Computer Systems*, no. 1, pp. 103–114, 2022. Available: https://doi.org/10.32620/reks.2022.1.08
12. Meng Xing, Yanbo Zhang, Hongmei Yu, Zhenhuan Yang, and Xueling Li, "Predict DLBCL patients' recurrence within two years with Gaussian mixture model cluster oversampling and multi-kernel learning," *Computer Methods Programs in Biomedicine*, vol. 226, 107103, 2022. Available: https://doi.org/10.1016/j.cmpb.2022.107103
13. Lynne A. Kvapil, Mark W. Kimpel, Rasitha R. Jayasekare, and Kim Shelton, "Using Gaussian mixture model clustering to explore morphology and standardized production of ceramic vessels: A case study of pottery from Late Bronze Age Greece,"

Journal of Archaeological Science: Reports, vol. 45, 103543, 2022. Available: https://doi.org/10.1016/j.jasrep.2022.103543

14. Meng Yinfeng, Jiye Liang, Fuyuan Cao and Yijun He, "A new distance with derivative information for functional k-means clustering algorithm," *Information Sciences*, vol. 463–464, pp. 166–185, 2018. Available: https://doi.org/10.1016/ j.ins.2018.06.035
15. Xinmin Tao, Ruotong Wang, Rui Chang, and Chenxi Li, "Density-sensitive fuzzy kernel maximum entropy clustering algorithm," *Knowledge-Based Systems*, vol. 166, pp. 42–57, 2019. Available: https://doi.org/10.1016/j.knosys.2018.12.007.
16. K. Møllersen, S. Dhar and F. Godtliebsen, "On Data-Independent Properties for Density-Based Dissimilarity Measures in Hybrid Clustering," *Applied Mathematics*, vol. 7, no. 15, pp. 1674–1706, 2016. doi: 10.4236/am.2016.715143.
17. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Softmax Units for Multinoulli Output Distributions. Deep Learning*. MIT Press, 2016.
18. V.E. Strilets et al., *Methods of machine learning in the problems of system analysis and decision making: monograph*. Karazin Kharkiv National University, 2020, 195 p.
19. Farbod Farhangi, "Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling," *Intelligent Systems with Applications*, vol. 15, 200100, September 2022. Available: https://doi.org/10.1016/j.iswa.2022.200100
20. Arthur Zimek and Peter Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6), 2018. doi: 10.1002/widm.1280.
21. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation-Based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data*, 6(1), pp. 1–39, 2012. doi:10.1145/2133360.2133363.
22. O.Yu. Lykhach, M.L. Ugryumov, D.O. Shevchenko, and S.I. Shmatkov, "Methods of detecting emissions in test samples during process control in state-based systems," *Bulletin of Karazin Kharkiv National University, ser. "Mathematical modeling. Information Technology. Automated control systems"*, no. 53. pp. 21–40, 2022.
23. L.J.P van der Maaten and G.E. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, 9, pp. 2579–2605, 2008.
24. Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A," *Mathematical, Physical and Engineering Sciences,* 374(2065), 20150202, 2016. doi: 10.1098/rsta.2015.0202.
25. L. Chagovets, N. Chernova, T. Klebanova, O. Dorokhov, and A. Didenko, "Selective Adaptive Model for Forecasting of Regional Development Unevenness Indexes," *Proceedings of the Workshop on the XII International Scientific Practical Conference Modern problems of social and economic systems modelling (MPSESM-W 2020) Kharkiv, Ukraine, June 25, 2020,* pp. 58–76.
26. L.O. Chagovets, S.V. Prokopovych, S.M. Vozniuk, and V.V. Chahovets, "Conceptual basis of modeling telecommunication development of regions by methods of system analysis," *Municipal economy of cities*, vol. 1, no. 161, pp. 230–240, 2021.
27. Computer program "Nonlinear estimation methods in the multicriterion problems of system's robust optimal designing and diagnosing under parametric apriority uncertainty (methodology, methods and computer decision support and making system)" ("ROD&IDS"): Copyright registration certificate no. 82875 / M.L. Ugryumov, Y.S. Meniaylov, S.V. Chernysh, K.M. Ugryumova (Ukraine). Copyright and related rights. Official bulletin. Ministry of Economic Development and Trade of Ukraine. 2018, no. 51, p. 403.

Received 30.06.2023

INFORMATION ON THE ARTICLE

Volodymyr V. Donets, ORCID: 0000-0002-5963-9998, V.N. Karazin Kharkiv National University, Ukraine, e-mail: v.donets@karazin.ua

Viktoriia Y. Strilets, ORCID: 0000-0002-2475-1496, V.N. Karazin Kharkiv National University, Ukraine, e-mail: viktoria.strilets@karazin.ua

Mykhaylo L. Ugryumov, ORCID: 0000-0003-0902-2735, V.N. Karazin Kharkiv National University, Ukraine, e-mail: m.ugryumov@karazin.ua

Dmytro O. Shevchenko, ORCID: 0000-0002-7897-250X, V.N. Karazin Kharkiv National University, Ukraine, e-mail: dimyich24@gmail.com

Svitlana V. Prokopovych, ORCID: 0000-0002-6333-2139, Simon Kuznets Kharkiv National University of Economics, Ukraine, e-mail: prokopovichsv@gmail.com

Liubov O. Chagovets, ORCID: 0000-0003-4064-9712, Simon Kuznets Kharkiv National University of Economics, Ukraine, e-mail: liubov.chahovets@hneu.net

МЕТОДОЛОГІЯ АНАЛІЗУ ДАНИХ ЕКОНОМІЧНОГО РОЗВИТКУ КРАЇН / В.В. Донець, В.Є. Стрілець, М.Л. Угрюмов, Д.О. Шевченко, С.В. Прокопович, Л.О. Чаговець

Анотація. Досліджено питання удосконалення методів ідентифікації економічних об’єктів та їх аналізу з використанням алгоритмів інтелектуального оброблення даних. Використання розробленої методології в економічному аналізі дозволяє підвищити якість управління та може бути основою для створення систем підтримання прийняття рішень для попередження потенційно небезпечних змін економічного стану об’єкта дослідження. Запропоновано удосконалений метод кластеризації даних с-середніх з агентно-орієнтованою модифікацією, для визначення відповідності отриманих кластерів актуальним пропонується радіально-базисна нейромережа та її розширення – для аналізу інформативності змінних стану й отримання підмножини інформативних змінних. Розглянуто вплив застосування стиснення даних за допомогою автокодувальника на точність застосування методів. За результатами тестування розробленої методології було доведено зменшення ймовірності неправильного визначення стану під час ідентифікації станів економічних систем та отримано зменшене значення помилки третього роду під час класифікації станів об’єктів.

Ключові слова: машинне навчання, цифровий розвиток, нечітка кластеризація, радіально базисні нейромережі, логістична регресія, аналіз інформативності змінних.