UDC 004.8

Olena O. Arsirii[1], Doctor of Technical Sciences, Professor, Head of Information Systems Department, E-mail: e.arsiriy@gmail.com, ORCID: https://orcid.org/0000-0001-8130-9613
Olga S. Manikaeva[1], postgraduate student of Information Systems Department, E-mail: manikaeva@gmail.com, ORCID: http://orcid.org/0000-0002-0631-8883
[1]Odessa National Polytechnic University, Avenue Shevchenko, 1, Odessa, Ukraine, 65044

MODELS AND METHODS OF INTELLECTUAL ANALYSIS FOR MEDICAL-SOCIOLOGICAL MONITORING'S DATA BASED ON THE NEURAL NETWORK WITH A COMPETITIVE LAYER

Abstract*: In this scientific publication, we suggest using the system of intellectual analysis of medical and sociological monitoring's data using a neural network with a competitive Kohonen layer to automate the process of obtaining knowledge (metadata) about the state of public health of the target audience. The following specialized tools have been developed to implement the system: models and a method for presenting detailed and aggregated medical and sociological data in area of primary and secondary features; the method of neural network classification of respondents based on machine learning of a neural network with a competitive layer; the procedure for labeling neurons of the Kohonen layer, taking into account the classification decisions received from the sociologist-analyst (initial markers). At at the first step, a two-dimensional histogram of pairwise coincidences of neuron numbers and existing initial class markers was constructed, and then it was corrected by lines and by columns in accordance with the developed rule. The result of the correction is the correspondence matrix of the numbers of neurons of the Kohonen layer and existing markers of classification decisions. The testing of the developed models and methods is based on a system of intellectual analysis using real medical-sociological monitoring's data. The research results show that it is possible to increase the relative share of correct classification decisions by an average of 20 % and reduce the share of false decisions by 50% compared with the sociologist-analyst for tasks of intellectual analysis of medical and sociological monitoring's data. These tasks were related to determining the working conditions of respondents.*

Keywords: *data mining; medical and sociological monitoring; neural networks with a competitive layer*

The introduction and the research problem's formulation. The computational capabilities of modern computer systems make possible collecting, accumulating and analyzing data on the state of public health, based on the population's self-esteem of their health, quality of life, and satisfaction with medical and social services. Data, collected continuously in the mode of medical and sociological monitoring, are of particular value [1]. Creating a permanent system for collecting and evaluating information – medical and sociological monitoring (MSM) involves organizing surveys, questionnaires or interviews with the subsequent analysis of answers containing heterogeneous information. It allows the sociologist-analyst to "extract" valuable knowledge about the target (analyzed) audience in the form of quantitative metrics or qualitative assessments of behavioral, medical, socio-demographic, psychophysical, geographical, or any other characteristics. So, for example, the intellectual activity of the sociologist in making classification decisions about the impact of working conditions on the health of the target audience is associated with the organization of monitoring not only by the complex characteristics of the levels of aerosol, electromagnetic, acoustic, chemical and biological effects, ionizing radiation, microclimate, lighting, vibration in industrial premises, but also with the constant collection of information on gender and age composition, bad habits, sports, participation in medical examinations and other characteristics of the quality of life of the analyzed workers [2].

However, the computer form for the representation of heterogeneous empirical data of MSM contains the necessary information for an expert assessment of the state of the target audience only in an implicit form. In order to make a classification decision, it is necessary to use special methods of data analysis to extract this information.

As a result, a class of tasks related to data mining was singled out. Data mining is an approach that combines methods that can detect previously unknown, non-trivial, practically useful and accessible interpretations of knowledge in raw empirical data of MSM, necessary for making decisions regarding the studied target audience, using a specific technique [3]. The current level of development of information technology allows you to automate the process of conducting an intellectual analysis of empirical quantitative and categorical data of MSM, both for constructing various data models in the source and feature spaces, and for their visual representation in the decision space [4-5].

The analysis of existing scientific achievements and publications

In the general case, new knowledge about the target audience is extracted on the basis of the analysis of empirical data of the sociological survey, and according to [6-7] this process is presented in the form of a cyclic sequence consisting of the following steps:

1) Awareness of the theoretical or practical insufficiency of the existing knowledge of the target audience;

2) Formulation of the problem and the hypothesis (in qualitative research hypotheses are usually presented at the last stages of research);

3) Collecting of empirical material on the basis of which hypothetical assumptions can be confirmed or refuted;

4) Analysis of empirical data using various methods, strategies, research programs and models;

5) Interpretation of the processed data and decision-making, explanation of the social phenomenon with the use of them;

6) Redefinition and clarification of a problem or hypothesis leading to a new research cycle (return to stage 3).

On the other hand, recently, due to the advent of modern software, Data Mining methods are gradually becoming the most popular tools for the sociologist- analyst. From the point of view of the existing standards describing the organization of the Data Mining process and the development of Data Mining systems, the most popular and common methodology is CRISP-DM (The Cross Industry Standard Process for Data Mining) [8,9]. In accordance with the CRISP-DM standard, Data Mining is a continuous process with many cycles and feedbacks and includes the following six steps: Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The seventh step is sometimes added to this sequence of stages - Control; it ends the circle. Using the CRISP-DM methodology, Data Mining becomes a business process during which Data Mining technology focuses on solving specific business problems.

Let’s consider some of the problems of organizing the Data Mining process of the sociological survey data for a sociologist-analyst using the CRISP-DM standard.

Thus, according to [6-7] at the stage of understanding business, the sociologist-analyst on the basis of his knowledge or lack of knowledge about the target audience, solves the problem of hypothesizing, which must be confirmed or refuted as a result of a sociological research. According to [7] such a hypothesis is a partially substantiated pattern of knowledge, serving either for the connection between various empirical facts or for explaining a fact or a group of facts. For example, as a formalized goal of the sociological survey there exists a hypothesis that the dependent variable (lifestyle) varies depending on some reasons (quality of food, alcohol consumption, playing sports, etc.) that are independent variables. However, this variable is not initially dependent or independent. It becomes such at the stage of understanding the business.

At the stage of data understanding, the sociological expert solves the problems connected with collecting sociological (or empirical) data. Such data can be defined as primary information of any kind, obtained as a result of one of the many types of sociological information collecting. [10]. As a rule, to conduct deep analytical studies, data are collected through questionnaires and interviews with a “complex” structure. Any empirical data is always structured.

Depending on the degree of structuredness, the data can be subdivided into the following types:

– *unstructured data* – text-type data, obtained in the process of conducting different types of interviews (narrative, keynote, etc.), the texts of answers to open questions with an unlimited search field for answers and any other texts or documents to which the sociologist could address [9-10];

– *strongly-structured data* – data, existing in the form of matrices of any type (for example, tables), obtained through the part of the questionnaire (interview) according to the scheme of closed questions;

– *weakly-structured data* – data of an intermediate type, not only quantitative but also categorical (nominal and ordinal) types, as well as existing in textual form, however, while being specially organized. Examples of such kind of data – data with a limited number of unique values or categories. As a rule, they contain answers to open-ended questions of an interview (interview) with a limited field of search for answers, either obtained by the method of incomplete sentences (test for sentence completion), or by using the method of repertory grids (G. Kelly's theory of personality constructs) [11-12].

At the stage of *data preparation*, the sociologist-analyst faces the tasks associated with the organization of a *multidimensional attribute*

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

space – the formalization and structuring of sociological survey data. When transforming unstructured data into feature space, special text recognition and analysis methods (*OCR* and *Text Mining*) are used for the further modeling. Rigidly structured data is a quantitative data type that is susceptible to the formalization and automatically transformed feature space. For the structuring and formalization of poorly structured data, various methods of preliminary data processing are used, for example, cleaning, filtering, rating and coding, etc. As a result of this preliminary processing, the weakly structured sociological survey data is transformed into a multidimensional feature space.

To conduct the sociological survey data modeling stage in the multidimensional attribute space, the sociologist-analyst solves the problems associated with the choice of methods and models for conducting Data Mining. Moreover, he has at his disposal the entire arsenal of regression, discriminant, dispersive, cluster, correlation, factor analysis methods [12-13].

To improve the responsiveness of MSM Data Mining, you can use machine learning [14] or precedent training. In that situation, there is a multidimensional set of independent variables (empirical facts) and many possible values of the dependent variable (answers) in accordance with the research hypothesis. There is some relationship between the variables, but it is unknown. There is only a finite set of precedents – labeled pairs of "facts, answer", called the training sample. Learning with a teacher (supervised learning) is learning when each precedent is a pair of "facts, answer" and it is required to build an algorithm that accepts a lot of facts at the input and gives an answer at the output. Learning without a teacher (unsupervised learning) haven't answers set, and you need to look for dependencies between empirical facts. In this case, the training sample consists of unlabeled use cases (data). It is often used teaching with the involvement of a teacher (partial training - semi-supervised learning) with MSM Data Mining. Teaching with the involvement of a teacher occupies an intermediate position. Each use case is a pair of "facts, answer", but the answers are known only on a part of the use cases. In this case, training sample consists of weakly labeled use cases [15].

At the stages of results evaluating and implementing, the tasks of interpreting the so-called sociological survey metadata or extracted knowledge obtained after the simulation, solving the characteristics of the target audience and explaining the studied social phenomenon with their help are solved [6-7].

Classification errors of the 1st and 2nd kind and their relative fractions of truly positive cases and truly negative cases are calculated to compare the results of the examination of the object's state and the classification decisions obtained in the IMS system of MSM for all examples of the training sample [11]:

$$TPR = \frac{TP}{TP+FN} * 100\ \%, \qquad (1)$$

where: *TPR* – relative share of true positive cases (True Positives Rate – TPR);

TP – true positive cases (correctly classified positive examples);

FN – positive examples classified as negative (I type error);

$$FPR = \frac{FP}{TN+FP} * 100\ \%, \qquad (2)$$

where: *FPR* – relative share of true negative cases (False Positives Rate – FPR);

FP – negative examples classified as positive (II type error);

TN – true negative cases (correctly classified negative examples).

Thus, the complexity of developing data mining tools for solving practical problems of extracting knowledge about the target audience based on MSM data is associated with the need to solve two types of tasks:

1. MSM data were defined as weakly structured at the stage of understanding data because they were collected through surveys, questionnaires and interviews with a "complex" structure, were interpreted using various and not always connected scales, and could contradict each other. Therefore, at the stage of data preparation, when feature spaces is constructing, it is required to solve the problems of structuring and formalizing poorly structured data.

2. Only weakly labeled precedents can be used to build the training sample to conduct the MSM Data Mining, using machine learning, the explanation of this situation is that expert decision on the labeling of available MSM data (facts) requires a long manual analysis of multidimensional features, depends on the qualification of the sociologist-analyst and is often ambiguous. Therefore, at the modeling stage, it is required to solve problems associated with the organization of training with the partial involvement of a teacher.

To eliminate the above-mentioned deficiencies of the sociological survey of data processing systems and to facilitate the work of the sociologist-analyst, a problem-oriented Data Mining system has been developed. The system consists of subsystems of initial data preparing, modeling and decision-making support, interconnected with the help of interface tools for a decision maker – sociologist-analyst [16; 17; 18]. The creation of such problem-oriented intellectual systems is always based on previously developed models, methods and information technologies (IT).

The purpose and tasks of the research. The purpose of this study is to develop models and methods for the intellectual analysis of medical and sociological monitoring's data to automate the extraction of knowledge about the target audience. This extraction will improve the efficiency and reliability of the classification of the respondents.

The following tasks must be solved to achieve purpose of this scientific publication:

1. The formalization of the MSM hypothesis in terms of classification / clustering problems; developing of detailed and aggregated MSM data.

2. The developing the method of neural network classification of poorly labeled data with the partial involvement of a teacher.

3. The testing of developed models and methods in the framework of the MSM Data Mining system, using the "working conditions" data.

The main research material includes a consistent presentation of the results of solving the formulated problems within the framework of a single technological approach to the organization of Data Mining.

The formal statement of the MSM Data Mining problem, models of detailed and aggregated MSM data

The task of extracting knowledge about the target audience (obtaining metadata), as a result of the MSM Data Mining, can be reduced to the task of classification in terms of machine learning [14]. Then, the target audience (TA) consists of many respondents (R):

$$TA = \{R_1, R_2, \dots, R_i, \dots, R_n\}, \quad (3)$$

where: R_i – respondent, we are studying;

n – number of respondents.

Each respondent can be characterized by a multidimensional set of variables denoting its properties:

$$R_i = \{x_1, x_2, \dots, x_j, \dots, x_m\}, \quad (4)$$

where: x_j – independent variables indicating the respondent's properties;

m – the number of respondent properties, which are less than the analyzed respondents – $m < n$.

A multidimensional set of independent variables (4) during MSM is a set of poorly structured data, the type and format of which depends on the method of obtaining the initial information. For example, respondent data can be obtained not only by using completed survey forms, questionnaires and interviews, but also extracted from electronic mail messages, business cards, pdf- and txt-files, instant messages in various messengers (WhatsApp, Viber, Facebook Messenger, Skype, ICQ, Google Hangouts, etc.), documents, web pages, bills, audio/ video, checks and contracts, pictures, etc. Such extracted from various sources data, called detailed data.

Therefore, taking into account representation (4), we propose a model of detailed data D_{R_i} of the respondent D_{R_i}in the m-dimensional property space, which is formally defined via the tuple

$$D_{R_i} = \langle V_j, T_j, F_j, S_j, Q_j, Mt_j \rangle, = 1, m\ , \quad (5)$$

where: $V_j, T_j, F_j, S_j, Q_j, Mt_j$ – value, type, format, source, quality grade, data transformation method of j-th property.

To increase the reliability of making the classification decision t_i, the detailed data D_{R_i} must be aggregated (combined) and, using the Mt_j method, transformed into the space of attributes: $X_j = \vee_\varphi Mt_j(V_j)$, for example, min, max, Σ, C (), etc. are used as an aggregation function for detailed data D_{R_i}. The procedures of the transformation stage of MSM's data are described in detail in [19].

Then, taking into account the representation (4), the model of aggregated data A_{R_i} of the respondent R_i in the q-dimensional space of attributes must be specified via the tuple:

$$A_{R_i} = \langle x_1, \dots, x_j, \dots, x_q \rangle, \quad (6)$$

where: q – dimension of the space of attributes, $q < m$.

On the other hand, the sociologist-expert hypothesizes that each respondent R_i of the target audience TA belongs to any class t_i from the set of

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

values of classes $T = \{t_1, \dots, t_i, \dots, t_p\}$, p is the number of classes into which a subset of respondents is divided (3).

The value of t_i can be determined by analyzing the set of aggregated data A_{R_i} from the q - dimensional attribute space. This analysis is performed by the sociologist-expert, using his own experience and knowledge. The accuracy of rules $a: X \to T$ constructing depends directly on the previously put forward hypothesis about the power of the classes set – $|T|$. The adoption of this hypothetic is ambiguous and determined by the qualification of the sociologist-expert. As a result of the analysis, we obtained the next equation:

$$A_{R_i} = \langle x_1, \dots, x_j, \dots, x_q, t_i \rangle, \qquad (7)$$

where: t_i – label of the class, to which the respondent R_i belongs to, according to the decision of the expert.

In case, when we use MSM's data as initial data, it is not possible to obtain a sufficient amount of tagged data, because it's requires a large investment of time and other resources of the sociologist. Then, in accordance with (3) and (6), we have:

1. Data set $A_{TA} = \{X_1, X_2, \dots, X_n\}$; label set $T = \{t_1, t_2, \dots, t_p\}$;
2. Labeled data set: $(X_l, T_l) = \{(X_{1:l} T_{1:l})\}$;
3. Unlabeled data set $X_u = \{X_{l+1:n}\}$, used in the learning, generally $l \ll n$;
4. Unlabeled data set $X_{test} = \{X_{n+1:g}\}$, not used in the learning (test set).

Taking into account (6) and (3), the labeled training data set in a matrix form is a set of aggregated data of respondents R_i of the target audience TA:

$$A_{TA_l} = \langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{lq} \end{bmatrix}, \begin{bmatrix} t_1 \\ \dots \\ t_p \end{bmatrix} \rangle, \qquad (8)$$

where: X matrix – the set of descriptions of respondents R_i in a q-dimensional space of attributes, T vector – finite set of class numbers (names, labels).

The training selection of unlabeled data (unlabeled training set) and the test selection (test set) have form:

$$A_{TA_u} = \langle \begin{bmatrix} x_{l+11} & \dots & x_{l+1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix} \rangle, \qquad (9)$$

$$A_{TA_{test}} = \langle \begin{bmatrix} x_{n+11} & \dots & x_{n+1q} \\ \dots & \dots & \dots \\ x_{g1} & \dots & x_{gq} \end{bmatrix} \rangle. \qquad (10)$$

Then, we can formulate the task of classifying MSM's data in terms of machine learning with partial involvement of a teacher as follows: there is an unknown target relationship – $a: X \to T$, the values of which are known only for a finite number of respondents R_i from the labeled training selection A_{TA_l}. Considering the data from the unlabeled training selection A_{TA_u}, it is necessary to construct an algorithm $a^*: X \to T$, which capable to classify with specified accuracy the arbitrary responder $R_i \in A_{TA_{test}}$.

We propose to use Data clustering to solve the problem of classifying MSM's data with partial involvement of a teacher. Data clustering solves the problem of dividing labeled A_{TA_l} and unlabeled A_{TA_u} training data into disjoint subsets, called clusters, so that each cluster consists of similar respondents, and respondents from different clusters are significantly different.

In this case, the task of clustering respondents from the combined set A_{TA}, consisting the sets of aggregated data A_{TA_l} and A_{TA_u}, includes constructing the set:

$$Z = \{z_1, \dots, z_i, \dots, z_p\}, \qquad (11)$$

where: z_i – the cluster, containing similar A_{R_i} (6) from the (8) set:

$$z_i = \left\{ A_{R_i}, A_{R_j} \middle| \begin{matrix} A_{R_i} \in A_{TA}, A_{R_j} \in A_{TA}, \\ d(A_{R_i}, A_{R_j}) < \sigma \end{matrix} \right\}, \qquad (12)$$

where: σ – value that determines the proximity measure for inclusion in one cluster, and $d(A_{R_i}, A_{R_j})$ – the measure of proximity between objects called distance.

The resulting data set of aggregated data A_{TA} has form:

$$A_{TA} = \langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix}, \begin{bmatrix} z_1 \\ \dots \\ z_p \end{bmatrix} \rangle. \qquad 13)$$

The obtained MSM's aggregated data sets (8) and (13) are loading into the data warehouse and then used to construct the classification algorithm $a: X \to T$.

Thus, you can use a specialized subsystem, built on the principle of ETL systems (Extraction, Transformation Loading) to collect, transform and store the MSM's data [17].

The development of the method for neural network classification with partial involvement of a teacher of poorly labeled data. A neural network classification method with partial involvement of a teacher for poorly labeled MSM's data, based on preliminary cluster analysis using a competitive layer of Kohonen W, was developed during creating a problem-oriented MSM Data mining system for the implementation of the modeling stage (Fig. 1).

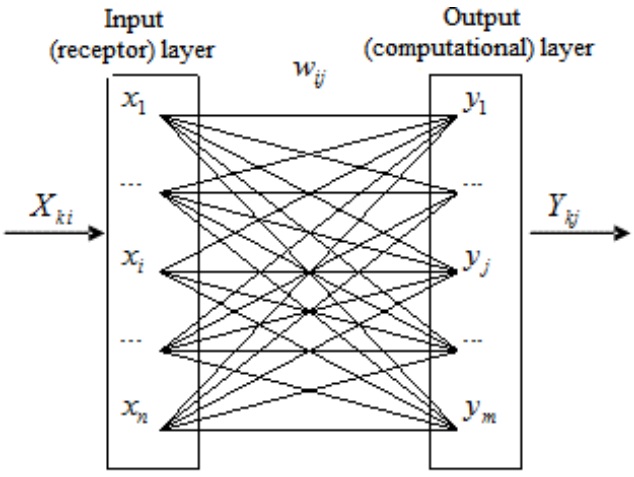


Fig. 1. The neural network model of the Kohonen W-layer

To obtain aggregated A_{TA} data (13), we performe machine learning of the Kohonen layer, using aggregated labeled A_{TA_l} and unlabeled A_{TA_u} data that are extracted from the storage and combined to form learning selection as:

$$A_{TA_{train}} = \left\langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix} \right\rangle. \qquad (14)$$

The implementation of machine learning of the Kohonen layer reckons for the sequential implementation of the self-organization steps of the competitive Kohonen layer's neurons, graduation of the elements of the output vector of the training selection and final labeling of Kohonen layer's neurons (Fig. 2).

First stage. The classic self-organization procedure of a Kohonen competitive layer is implemented via the iterative WTA (Winner Takes All) algorithm [20]. According to the WTA algorithm, the values of the vectors $\{x_{ki}\}$ from the training set $A_{TA_{train}}$, $i = \overline{1,q}$, $k = \overline{1,n}$, where q – number of attributes, and n – number of vectors in the selection, are sequentially sent to the input of the Kohonen layer (Fig. 1) (14).

The goal of the self-organization is to minimize the difference between distances:

$$d(x_{ki}, w_{ij}) \longrightarrow \min d(x_{ki}, w_{ij}), \qquad (15)$$

of the input vectors's x_{ki} elements and the weight coefficients w_{ij} of the Kohonen layer's neuron-winner, in accordance with correction formula:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)[x_{ki} - w_{ij}(t)], \quad (16)$$

where: $\eta(t)$ –time-varying correction step factor.

Usually, the monotonically decreasing function $(0 < \eta(t) < 1)$ is chosen as $\eta(t)$. The Euclidean distance is used as a measure of distance:

$$d(x_{ki}, w_{ij}) = \sqrt{\sum_{i=1}^{q} (x_{ki} - w_{ij})^2}. \qquad (17)$$

Second stage. The values of the training selection vectors $\{x_{ki}\}$ are sequentially sent to the input of the Kohonen self-organizing layer during performing the calibration procedure for the elements of the output vector of the training selection. On the input of the Kohonen layer for each of these values, on the basis of (14), values of the output vector $\{y_{kj}\}$, $j = \overline{1,p}$ are formed (p is the number of neurons (classes)). The serial number of the winning j neuron of the k-th vector $\{y_{kj}\}$ is assigned as the value to the k-th element of the calibration vector $\{z_k\}$. Thus, after performing the calibration procedure, we have a preliminary solution in the following form:

$$R^* = \langle X, Y, Z \rangle \left\langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix}, \begin{bmatrix} y_{11} & \dots & y_{1p} \\ \dots & \dots & \dots \\ y_{n1} & \dots & y_{np} \end{bmatrix}, \begin{bmatrix} z_1 \\ \dots \\ z_p \end{bmatrix} \right\rangle, \qquad (18)$$

where: the elements of the X matrix satisfy formula (14), the elements of the Y binary matrix of the desired outputs are formed as a result of the competitive Kohonen layer training. Vector y_{kj}= 1, if j is the winner's neuron number and y_{kj}= 0 in all other cases. The elements of the Z calibration vector are the index number of the winning neuron j.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

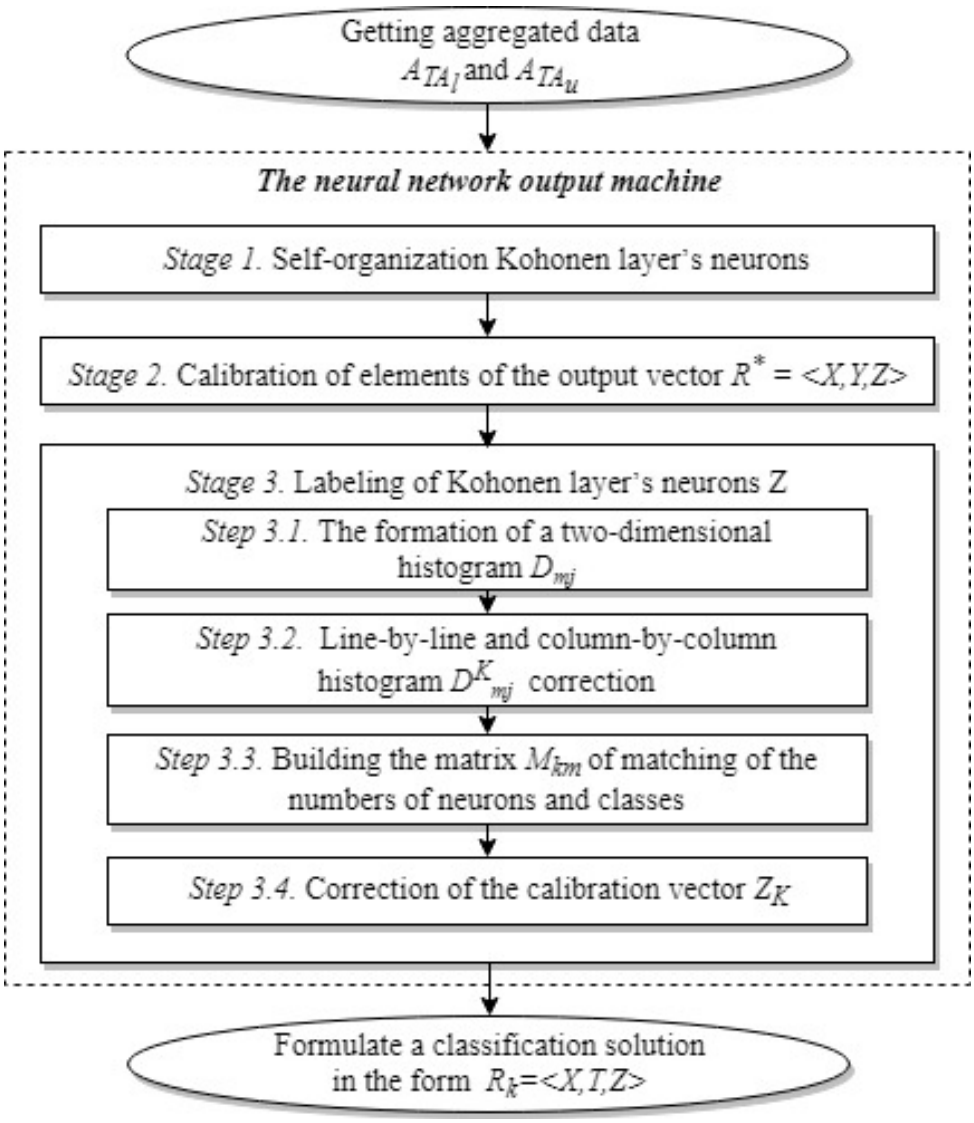


Fig. 2. Method's of the neural network classification with partial involvement of a teacher of poorly labeled data stages

Third stage. The procedure for labeling Kohonen layer's neurons by class numbers of the training selection's aggregated data has been proposed to determine the matching between the value of the final grade $\{t_k\}$ (class number) from the labeled training set A_{TA_l} (8) and the value of the neuron-winner number of the calibration vector $\{z_k\}$ (18), (8). The labeling procedure requires three steps (Fig. 2).

Step 3.1., We must form a two-dimensional histogram D_{mj} (square matrix) of pairwise coincidences of the neurons and classes numbers $z_{km} = t_{kj}$, where $m, j = \overline{1,p}$, for all examples of the training selection. Histogram D_{mj} building on a mnemonic code:

```
D=0;    /*the matrix elements nullifying*/
for m = 1:p
for j = 1:size(Y,2)/*the number of classes*/
for k = 1:size(Y,1) /* the number of training selection examples */
if ((P(k) == m) && (Z(k,l)== m))
D(m,j) = D(m,j)+1;
end
end
end
end
```

Step 3.2. We must perform line-by-line and column-by-column histogram D_{mj} correction, leaving unchanged only those values of elements that satisfy the condition:

$$\begin{cases} \max(D_m) = \max(D_j) \\ m = j \\ \max(D_{mj}) <> 0 \end{cases}. \quad (19)$$

The remaining values of the intersecting row and column are setting to zero. Line-by-column histogram D_{mj} correction on the mnemonic code:

```
for j = 1:size(D,1)
 for m = 1:size(D,1)
 if (D(m,j) == max(D(:,j)))
 && (D(p,j) == max(D(p,:))) &&
 (D(p,j)> 0 )
 D(:,j) = 0;
 D(m,:) = 0;
 end
 end
end
```

As you can see, the two-dimensional histogram D_{mj} correction is performed iteratively, while only one non-zero value remains in each line and column. The result of the correction is saving in D_{mj}^{K}.

Step 3.3. The corrected pairwise matches two-dimensional histogram D_{mj}^{K} can be transformed into the matrix of numbers of neurons and classes matching for all samples of the training selection ($M_{km}, k = \overline{1,2}$):

```
M=0
for m = 1:size(Dk,1)
 for j = 1:size(Dk,1)
       if Dk(m,j)>0
        M(1,m)= p;
       M(2,m)= j;
 end
 end
end
```

Step 3.4. At the last step, we perform the final correction of the calibration vector $\{z_k\}$, using the matrix of numbers of neurons and classes matching. Based on the (8) and the resulting vector$\{z_k\}$, we can formulate a classification solution (20) and transmit it to compare the results and evaluation of the expert marks' quality T on the state of the object by comparing it with solution Z:

$$R_k = \langle X, T, Z \rangle \langle \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix}, \begin{bmatrix} t_1 \\ \cdots \\ t_p \end{bmatrix}, \begin{bmatrix} z_1 \\ \cdots \\ z_p \end{bmatrix} \rangle. \quad (20)$$

The developed models and methods testing within the MSM Data mining system on the "working conditions" data

The research shows that the intellectual activity of the sociologist-expert in making classification decisions about the impact of working conditions on the health of the target audience is related to the organization of MSM's data for the "working conditions" group [17; 19; 21]. This group includes quantitative and qualitative attributes that set the levels of aerosol, electromagnetic, acoustic, chemical and biological effects, ionizing radiation, microclimate, lighting and vibration. As a result of the MSM's data analysis, in accordance with the proposed scale (which depends on the level of maximum permissible concentration of the studied factor), the sociologist-expert evaluates the state of working conditions as: "optimal", "acceptable", "harmful", "dangerous", "extreme".

The results of the testing of developed models and methods as part of the MSM Data mining system during a comprehensive examination of the level of aerosol exposure, as one of the factors of working conditions affecting the health of the target audience, are shown in Fig. 3. The training selection consisted of 1200 examples. The sociologist-expert, based on the analysis of the converted data on the dispersion composition, concentration, exposure time and type of aerosol particles (which are displayed in a special electronic questionnaire (Fig. 3–1), determines the general level of aerosol exposure. The aggregated data A_{TA_l} (Fig. 3–2) are transmitted through the storage to the neural network output machine to check the quality of the expert's decision. As a result of the self-organization and calibration steps, a preliminary decision R_p, which is transmitted to Step 3, is formed (Fig. 3–3). To label the elements of the calibration vector *Z*, a two-dimensional pairwise coincidences of the numbers of neurons and classes histogram *D* is formed from the preliminary solution R_p (Fig. 3–4). Then, in order to get the matrix of the numbers of neurons and classes matching *M* (Fig. 3–5), line-by-line and column-by-column correction is performed (Fig. 3–6).

To form the classification solution R_k via the matrix of matching, the final correction of the calibration vector *Z* is performed (Fig. 3–7 – *DSS decision*).

Thus, the sociologist-expert can compare his decision and the solution, proposed by the MSM Data mining system, and correct his decision. In addition to the determining of aerosol exposure level, the developed MSM Data mining system was used to determine the general levels of electromagnetic, acoustic, chemical and biological effects, ionizing radiation, microclimate, illumination and vibration. To compare the decision of the sociologist-expert (*Final level*) and the

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

classification decision (*DSS decision*) via the (1) and (2), the values of *TPR* and *FPR* were calculated. The results of the *TPR* and *FPR* calculations for all samples of the training selection for all of the determined exposure levels of the working conditions group before and after correction are shown in Fig. 4 and Fig. 5, a, b, respectively.

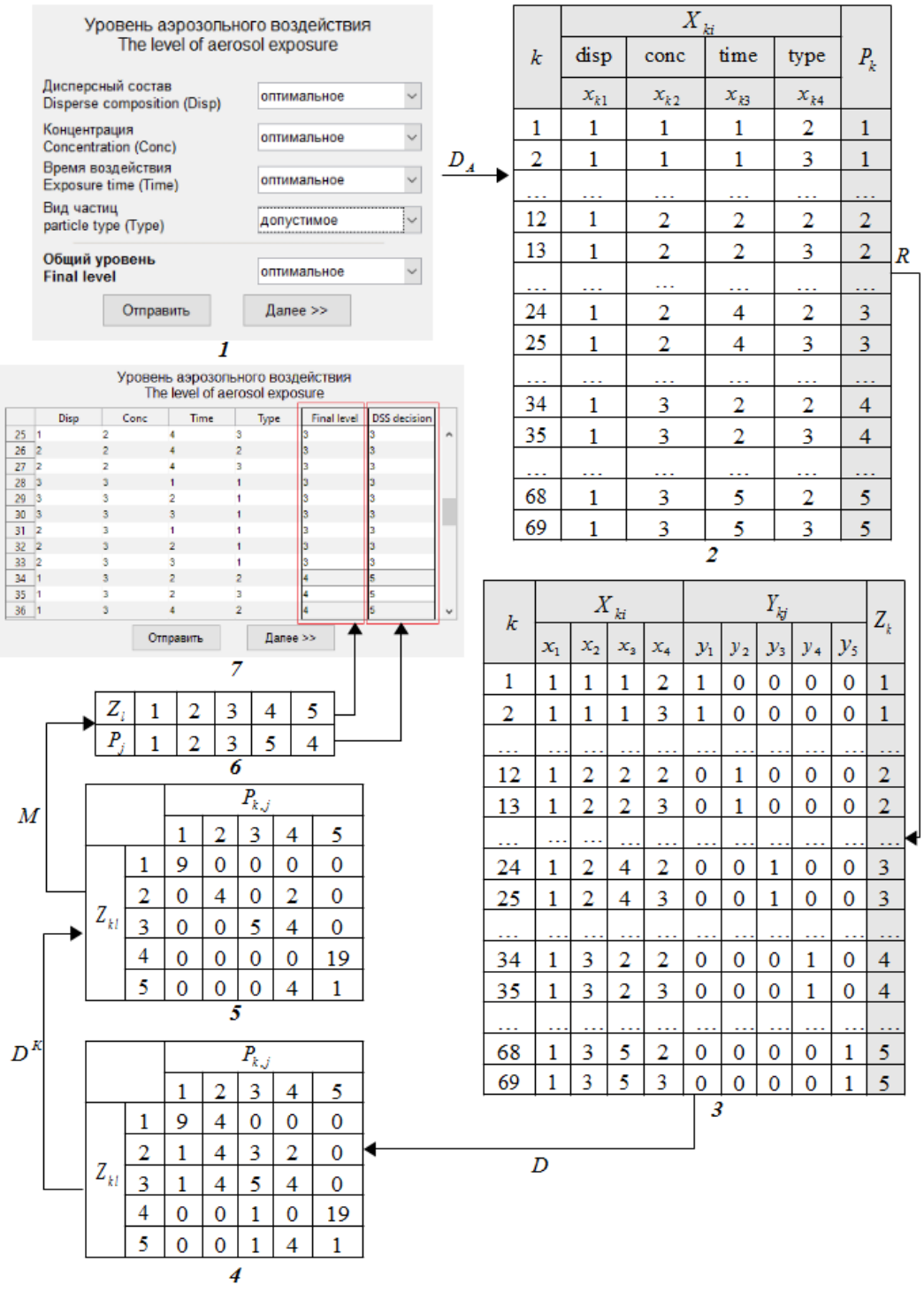
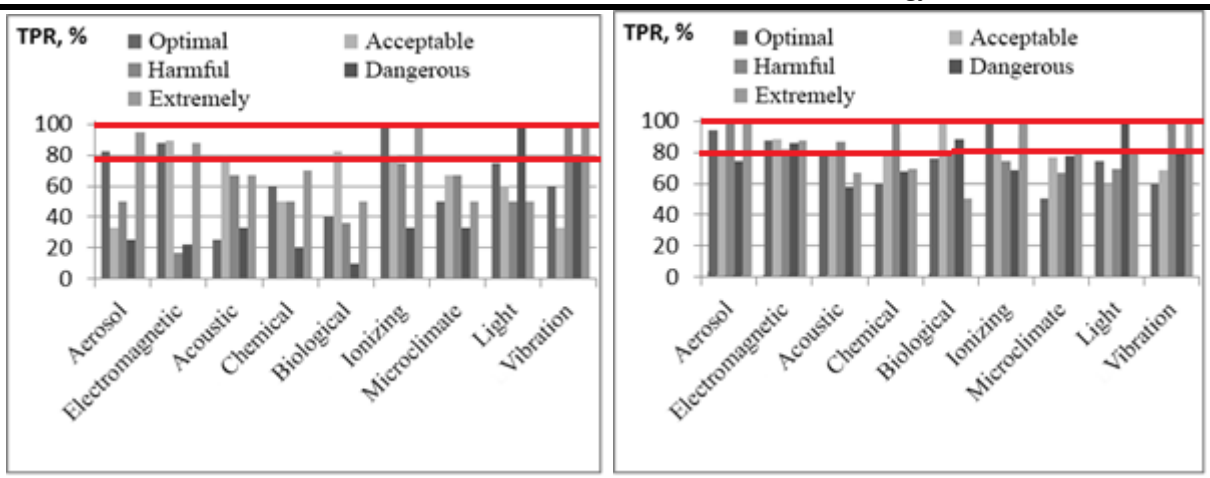
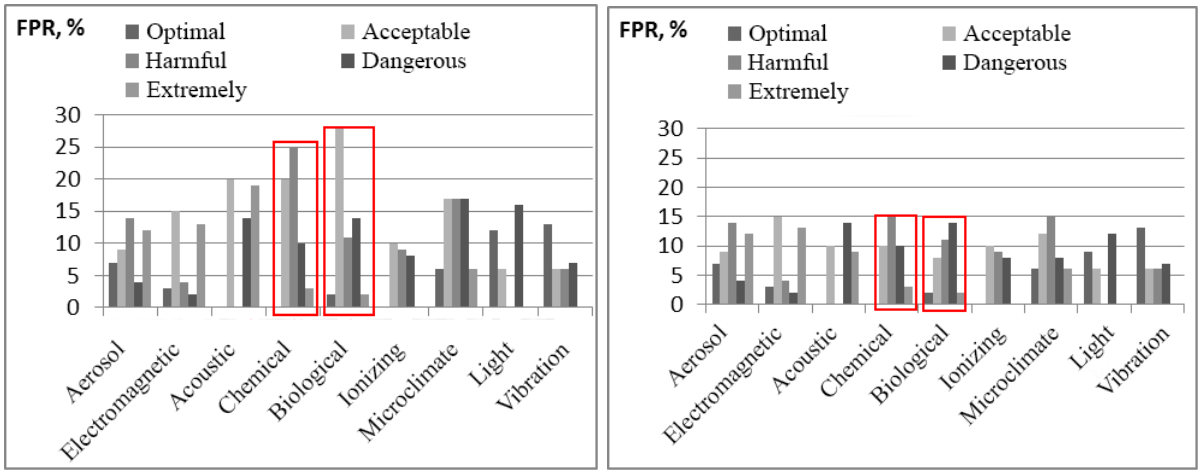


Fig. 3. The results of testing the operation of the MSM Data mining system:
1 – subsystem for the preparation of initial data; 2 – aggregated data from the storage receiving; 3 – the result of the calibration procedure of the output vector's elements; 4 – two-dimensional histogram formatting; 5 – line-by-line and column-by-column histogram correction; 6 – matrix of the matching of numbers of neurons and classes; 7 – the decisions of the sociologist-expert and classification decisions comparison



a b

Fig. 4. Comparative *TPR results*:
before – a and after – b of the correction



a b

Fig. 5. Comparative *FPR results*:
before – a and after – b of the correction

The calculation analysis of the values of first (Fig. 4) and second type errors (Fig. 5) before and after the correction shows an increase in the relative share of TPR by an average of 20 % and a decrease of 50 % in FPR for all groups of working conditions. Moreover, after the sociologist-expert decisions correction, the TPR marks are more evenly distributed for all of the studied exposure levels from the group of working conditions (Fig. 4). The decrease in the relative share of truly negative cases (*FPR*) is especially noticeable for the levels of chemical and biological effects (Fig. 5).

The conclusion and the prospects for further research

The testing results analysis showed that the use of the developed models and methods within the MSM Data mining system to determine the impact of working conditions on the health of workers allowed to increase the reliability of expert decisions. The following tasks were solved.

1. Based on the analysis of literary sources, it is shown that new knowledge about the target audience is extracted from the MSM's empirical data as a result of a cyclic sequence, consisting of six mandatory steps.

2. The problems of organizing the MSM Data mining system for a sociologist-analyst, based on the *CRISP-DM* standard, are considered.

3. The functional structure of the MSM Data mining system was proposed. The problem-oriented MSM Data mining system, according to the proposed structure, consists of subsystems of initial MSM's data preparation and aggregated data storage, neural network modeling approach and decision making on the basis of MSM's data supporting.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

4. The formalization of the MSM hypothesis was proposed. The detailed and aggregated MSM's data models were developed in terms of classify-cation/clustering problems.

5. The method of neural network classification with a partial involvement of a teacher of poorly labeled data has been developed.

6. The MSM Data mining system was developed and tested on the "working conditions" data.

The use of the developed within the MSM Data mining system models and methods allows you to correct the sociologist-expert decisions. This can be performed in accordance with the decisions, which was made by the neural network classification with partial involvement of a teacher of poorly labeled data. It's proven that it is possible to increase the relative share of correct expert's evaluation by an average of 20 % and reduce by 50 % for a number of MSM Data mining tasks the false evaluations. Among the latter: determination of enterprise's working conditions, according to the levels of aerosol, electromagnetic, acoustic, chemical and biological effects levels, ionizing radiation, microclimate, lighting and vibration.

References

1. Efimenko, S. A. (2019). "Mediko - sotsiologicheskiy monitoring kak instrument sovre-mennyih tehnologiy v upravlenii zdorovem patsientov". **[**Medical and sociological monitoring as an instrument of modern technologies in managing patient health], *Internet conference Health: problems of organization, management and levels of responsibility* [Electronic resource]. – Available at: http://ecsocman.hse.ru/text/16207043/. – Active link 01.07.2019 (in Russian).

2. Rudenko, A. I., & Arsirii, E. A. (2018). "Metodika intellektualnogo analiza slabostrukturirovannyih mnogomernyih dannyih sotsiologicheskih oprosov". [The methodology of intellectual analysis of poorly structured multidimensional data from sociological surveys], *Materials of the Eighth International Conference of Students and Young Students of Modern Information Technology 2018 (Modern Information Technology* systems *2018)* (May 23-25, 2018) / *MES of Ukraine;* Odessa Nat. polytech. un-t; Int Compute, Odessa, Ukraine, *Ecology*, pp. 168-169 (in Russian)**.**

3. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Data bases". [Текст], *AI Magazine*, Vol. 17, No. 3, pp. 37-54. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

4. Semenov, V. E. (2009). "Analiz i interpretatsiya dannyih v sotsiologii: uchebnoe posobie". [Analysis and interpretation of data in sociology: a training manual], Vladimir state. un-t. Vladimir, Russian Federation, *Publishing House of VlSU*, 131 p. ISBN 978-5-89368-916-7 (in Russian).

5. Kislova, O. N. (2005). "Intellektualnyiy analiz dannyih: vozmozhnosti i perspektivyi primeneniya v sotsiologicheskih issledovaniyah. MetodologIya, teorIya ta praktika sotsIologIchnogo analIzu suchasnogo suspIlstva". [Data Mining: Possibilities and Perspectives of Application in Sociological. Methodology, theory and practice of sociological analysis of modern society: Collection of scientific works Research], *ZbIrnik naukovih prats*, Kharkiv, Ukraine, pp. 237-243 (in Russian).

6. Babilunha, O., Arsirii, E. A., Manikaeva, O., & Rudenko, O. (2018). "Automation of the preparation process weakly-structured multi-dimensional data of sociological surveys in the data mining system". *Herald of Advanced Information Technology*, Vol. 1, No. 1, pp. 11-20. DOI://10.15276/hait.01.2018.1.

7. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). "CRISP-DM 1.0: Step- by-Step Data Mining Guide". *SPSS*, Copenhagen.

8. Wirth, R., & Hipp, J. (2000). "CRISP-DM: Towards a Standard Process Model for Data Mining", *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29-30.

9. Praveen, S., & Chandra, U. (2017). "Influence of Structured, SemiStructured, Unstructured data on various data models", *International Journal of Scientific & Engineering Research*, Vol. 8, Issue 12, pp. 67-69.

10. Corbetta, P. (2011). "Social Research: Theory, Methods and Techniques", London: *Sage*, 328 p.

11. Hunter, M. G. (2002). "The Repertory Grid Technique: A Method for the Study of Cognition in Information Systems", MIS Quarterly, 26(1), pp. 39-57.

12. Kim, J. O., & Muller, C. W. (1989). "Faktornyiy, diskriminantnyiy i klasternyiy analiz". [Factor, discriminate and cluster analysis], Moscow, Russian Federation, *Finance and Statistics* (in Russian).

13. Härdle, W., & Simar, L. (2012). "Applied Multivariate Statistical Analysis Free preview", Berlin; New York : *Springe*r, 486 p.

14. Witten, I. H., & Frank, E. (2005). "Data Mining: Practical Machine Learning Tools and Techniques" (Second Edition), Morgan Kaufmann,

Germany. ISBN 0-12-088407-0 https://www.cs.waikato.ac.nz/~ml/weka/ book.html.

15. Chapelle, O, Schölkopf, B, & Zien, A. (2006). "Semi-Supervised Learning (Adaptive Computation and Machine Learning series)", *The MIT Press*, (September 22, 2006), 528 p.

16. Merkert, J., Mueller, M., & Hubl, M. A (2015). "Survey of the Application of Machine Learning in Decision Support Systems", *Twenty-Third European Conference on Information Systems (ECIS 2015)*, Münster, Germany, pp. 1-15.

17. Arsiriy, E. A., Manikaeva, O. S., Vasilevskaya, A. P. (2015). "Razrabotka podsistemyi podderzhki prinyatiya resheniy v sistemah neyrosetevogo raspoznavaniya obrazov po statisticheckoy informatsii". [Development of a decision support subsystem in neural network pattern recognition systems based on statistical information], *East European Journal of Advanced Technologies*, Vol. 6, No. 4 (78), pp. 4-12.

18. Barsegyan, A. A., Kupriyanov, M. S., Stepanenko, V. V., & Holod, I. I. (2004). "Metodyi i modeli analiza dannyih: OLAP i Data Mining". [Methods and models of data analysis: OLAP and Data Mining], SPb., Russian Federation, *BHV-Petersburg*, 336 p.: ill. (in Russian)

19. Arsiri, O., Antoshchuk, S., Babilunha, O., Manikaeva, O., & Nikolenko, A. (2019). "Intellectual Information Technology of Analysis of Weakly-Structured Multi-Dimensional Data of Sociological Research", *International Scientific Conference "Intellectual Systems of Decision Making and Problem of Computational Intelligence" ISDMCI 2019*, Lecture Notes in Computational Intelligence and Decision Making, pp. 242-258. [Electronic resource]. – Available at: https://link.springer.com /chapter/10.1007/978-3-030-26474-1_18.

20. Kohonen, T. (1990). "The self-organizing map", *Proceedings of the IEEE*, Vol. 78, Issue 9, pp. 1464-1480. Doi: 10.1109/5.58325.

21. (2019). "Dannyie sotsiologicheskogo issledovaniya "Ukraina–stil zhizni" [Data from the sociological study "Ukraine– Lifestyle"]. [Electronic resource]. – Access mode: http://edukacjainauka.pl/limesurvey/ index. php/lang-pl 23. – Active link 01.07.2019 (in Russian).

Received 15.05.2019

УДК 004.8

[1]**Арсірій, Олена Олександрівна,** д-р техн. наук, проф., зав. кафедри інформаційних систем,
E-mail: e.arsiriy@gmail.com, , ORCID: https://orcid.org/0000-0001-8130-9613
[1]**Манікаєва, Ольга Сергіївна,** аспірант каф. інформаційних систем,
E-mail: manikaeva@gmail.com, ORCID: http://orcid.org/0000-0002-0631-8883
[1]Одеський національний політехнічний університет, пр. Шевченка, 1, м. Одеса, Україна, 65044

МОДЕЛІ І МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ МЕДИКО-СОЦІОЛОГІЧЕСКОГО МОНІТОРИНГУ НА ОСНОВІ НЕЙРОННОЇ МЕРЕЖІ З КОНКУРУЮЧИМ ШАРОМ

Анотація. *Для автоматизації процесу отримання знань (метаданих) про стан громадського здоров'я цільової аудиторії запропоновано використовувати систему інтелектуального аналізу даних медико-соціологічного моніторингу з використанням нейронної мережі з конкуруючим шаром Кохонена. Для реалізації системи розроблені наступні спеціалізовані засоби: моделі і метод представлення деталізованих і агрегованих медико-соціологічних даних в просторах первинних і вторинних ознак; метод нейромережевої класифікації респондентів на основі машинного навчання нейронної мережі з конкуруючим шаром; процедура маркування нейронів шару Кохонена з урахуванням класифікаційних рішень отриманих від соціолога-аналітика (первинних маркерів). При виконанні процедури маркування на першому кроці будуватися двовимірна гістограма попарних збігів номерів нейронів і існуючих первинних маркерів класів, далі виконується її коригування по рядках та по стовбцях відповідно до розробленого правилом. Результатом виконання коригування є матриця відповідностей номерів нейронів шару Кохонена і існуючих маркерів класифікаційних рішень. Апробація розроблених моделей і методів проводилася на основі системи інтелектуального аналізу з використанням реальних даних медико-соціологічного моніторингу. Показано, що вдалося підвищити відносну частку правильних класифікаційних рішень в середньому на 20 % і знизити на 50 % частку помилкових рішень в порівнянні з соціологом-аналітиком для ряду задач інтелектуального аналізу даних медико-соціологічного моніторингу пов'язаних з визначенням умов праці респондентів.*

Ключові слова*: інтелектуальний аналіз даних; медико-соціологічний моніторинг; нейронні мережі з конкуруючим шаром*

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

УДК 004.8

[1]**Арсирий Елена Александровна,** д-р техн. наук, проф., зав. кафедрой информационных систем,
E-mail: e.arsiriy@gmail.com, , ORCID: https://orcid.org/0000-0001-8130-9613
[1]**Маникаева Ольга Сергеевна,** аспирант каф. информационных систем,
E-mail: manikaeva@gmail.com, ORCID: http://orcid.org/0000-0002-0631-8883
[1]Одесский национальный политехнический университет, пр. Шевченко, 1, г. Одесса, Украина, 65044

МОДЕЛИ И МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ МЕДИКО-СОЦИОЛОГИЧЕСКОГО МОНИТОРИНГА НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ С КОНКУРЕНТНЫМ СЛОЕМ

Аннотация. *Для автоматизации процесса получения знаний (метаданных) о состоянии общественного здоровья целевой аудитории предложено использовать систему интеллектуального анализа данных медико-социологического мониторинга с использованием нейронной сети с конкурирующим слоем Кохонена. Для реализации системы разработаны следующие специализированные средства: модели и метод представления детализированных и агрегированных медико-социологических данных в пространствах первичных и вторичных признаков; метод нейросетевой классификации респондентов на основе машинного обучения нейронной сети с конкурирующим слоем; процедура маркировки нейронов слоя Кохонена с учетом классификационных решений полученных от социолога-аналитика (первоначальных маркеров). При выполнении процедуры маркировки на первом шаге строиться двумерная гистограмма попарных совпадений номеров нейронов и существующих первоначальных маркеров классов, далее выполняется ее построчно и постолбцовая корректировка в соответствии с разработанным правилом. Результатом выполнения корректировки является матрица соответствий номеров нейронов слоя Кохонена и существующих маркеров классификационных решений. Апробация разработанных моделей и методов проводилась на основе системы интеллектуального анализа с использованием реальных данных медико-социологического мониторинга. Показано, что удалось повысить относительную долю правильныхклассификационных решений в среднем на 20 % и снизить на 50 % долю ложных решений по сравнению с социологом-аналитиком для ряда задач интеллектуального анализа данных медико-социологического мониторинга,связанных с определением условий труда респондентов.*

Ключевые слова: *интеллектуальный анализ данных; медико-социологический мониторинг; нейронные сети с конкурирующим слоем*

Olena Oleksandrivna Arsirii
Doctor of Technical Sciences, Professor,
Research field: Information technology, decision support systems, machine learning, neural networks

Olga Sergiivna Manikaeva
postgraduate student Department of Information Systems
Research field: artificial intelligence methods and systems, neural networks