

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОІНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

UDC 004.93

Subbotin S. A.

*Dr.Sc., Professor, Professor of Department of software tools,
Zaporizhzhya National Technical University, Ukraine*

THE INSTANCE INDIVIDUAL INFORMATIVITY EVALUATION FOR THE SAMPLING IN NEURAL NETWORK MODEL SYNTHESIS

The problem of mathematical support development is solved to automate the sampling at diagnostic and recognizing model building by precedents. The object of study is the process of diagnostic and recognizing neural network model building by precedents. The subject of study is the sampling methods for neural network model building by precedents. The purpose of the work is to increase the speed and quality of the formation process of selected training samples for neural network model building by precedents. The method of training sample selection is proposed which for a given initial sample of precedents and given feature space partition determines the weights characterizing the term and feature usefulness. It characterizes the individual absolute and relative informativity of instances relative to the centers and the boundaries of feature intervals based on the weight values. This allows to automate the sample analysis and its division into subsamples, and, as a consequence, to reduce the training data dimensionality. This in turn reduces the time and provides an acceptable accuracy of neural model training. The software implementing proposed indicators is developed. The experiments to study their properties are conducted. The experimental results allow to recommend the proposed indicators for use in practice, as well as to determine effective conditions for the application of the proposed indicators.

Keywords: sample, instance selection, data reduction, neural network, data dimensionality reduction.

NOMENCLATURE

δ is a proportion of the original sample in the training subsample;

E_{all} is a neural network model error for the whole original sample;

ep_{tr} is a number of executed epochs of neural network training;

E_{tr} is a neural network model error for the training sample;

$F()$ is a neural network model structure;

$f()$ is a user criterion characterizing the argument quality relatively to the problem being solved;

K_{jk} is a number of classes, which instances hit the k -th interval of j -th feature values;

N is a number of features characterizing original sample;

N' is a number of features in a subsample;

opt is an optimal (desired or acceptable) value of the functional $f()$ for the problem being solved;

S is a number of instances in the original sample;

S' is a number of instances in a subsample;

S_{jk} is a number of instances in the k -th term of the j -th feature;

t_{tr} is a time of neural network model training;

w is a set of controlled (adjusted) parameters of the neural network model;

x_j^s is a value of j -th input feature x_j , characterizing the instance x^s ;

y^s is an output feature value associated with the instance x^s ;

y^{s*} is a calculated output feature value for the s -th instance on the neural model output;

x^s is s -th instance of a sample.

INTRODUCTION

To automate the decision making in problems of technical and medical diagnosis, as well as in pattern recognition problems it is necessary to have a model of a decision dependence from descriptive features, characterizing an instance to be recognized (an observation of the object or process condition at a certain time). As a rule, due to the lack or inadequacy of expert knowledge in practice such model constructed on the basis of observations or precedents (instances).

The one of the most popular and powerful tools for model building by precedents are artificial neural and neuro-fuzzy networks [1] that can learn by precedents, providing their generalization and extracting knowledge from the data.

The object of study is the process of diagnostic and recognizing neural network model building by precedents.

The process of neural model building is typically time-consuming and highly iterative. This is caused by that training time and accuracy of the neural network model are essentially dependent on the dimensionality and quality of the used training sample. Therefore, to improve the construction speed and quality of neural model it is necessary to reduce the dimension of the sample, providing the preservation of its basic properties.

The subject of study is the sampling methods for neural network model building by precedents.

The known sampling methods [2–23] are highly iterative and low speed, as well as characterized by the uncertainty of quality criteria of formed subsample.

The purpose of the work is to increase the speed and quality of the formation process of selected training samples for neural network model building by precedents.

1 PROBLEM STATEMENT

Suppose given the original sample as a set of precedents (instances) $\langle x, y \rangle$, where $x = \{x^s\}$, $x = \{x_j\}$, $x^s = \{x_j^s\}$, $x_j = \{x_j^s\}$, $y = \{y^s\}$, $s = 1, 2, \dots, S$, $j = 1, 2, \dots, N$.

For a given sample of precedents $\langle x, y \rangle$ the problem of neural model synthesis can be presented as the problem of finding $\langle F(), w \rangle$: $y^{s*} = F(w, x^s)$, $f(F(), w, \langle x, y \rangle) \rightarrow \text{opt}$, where the model structure $F()$ usually specified by the user in practice, and the set of controlled parameters w is adjusted based on the training sample.

In turn, the problem of subsample formation from a given sample $\langle x, y \rangle$ is to find such a set of $\langle x', y' \rangle$: $x' \subset \{x^s\}$, $y' = \{y^s | x^s \in x'\}$, $S' \subset S$, $N' = N$, wherein $f(\langle x', y' \rangle, \langle x, y \rangle) \rightarrow \text{opt}$.

2 REVIEW OF THE LITERATURE

The sampling methods for decision-making model building by precedents in [2, 3] are divided into prototype selection methods and prototype construction methods. Here, the prototype means selected subsample relative to the original sample.

The prototype selection methods [4–15] does not modify, but only select the most important instances from the original sample. Depending on the strategy of solution forming these methods are divided into incremental methods

[4, 5] (they successively add instances from the original sample to the subsample) and decremental methods [4–8] (they successively remove instances from the original sample, and obtain a subsample as a result). There are also separated such methods as noise filtering methods [6, 8, 9–11] (they remove instances, which class labels do not equal with most of the neighbor labels), condensation methods [4–7, 12, 13] (this methods add instances from the original sample to the formed subsample, if they bring a new information, but do not add if they have the same class labels as their neighbors), and methods based on stochastic search [12, 14, 15] (they randomly form a subsample from the original sample, considering a set of variants of decisions and selecting the best of them). The common disadvantages of these methods are the high iterativity and big search time, as well as uncertainty in quality criteria selection of formed subsample.

The methods of prototype construction [12, 15–23] based on the original sample build artificial instances, allowing to describe the original sample. Among these methods it is possible to separate the cluster analysis based methods [18, 19, 23] (they replace the original sample by the centers of its clusters), the data squashing methods [17] (they replace the original sample instances by the artificial prototypes having weights obtained on their basis) and the neural network methods (neural network based methods) [16, 20–22] (they train a neural network on the original sample, which is then used for cluster centers extraction as instances of formed subsample). The common disadvantages of these methods is their high iterativity and a big operating time, and the uncertainty in the initial parameter setting. The methods based on a cluster analysis are characterized by disadvantages such as the uncertainty of cluster number, initial parameters, and metric selection for the clustering and training methods. The data squashing methods form prototypes, which are difficult to interpret. The neural network methods have such disadvantages as the difficulty of prototype extraction from the neural network model, the no guarantee of receiving of acceptable neural network model a result of training, the neural network model variability, entailing nonstationarity of constructed prototypes, the orientation on a specific model, the uncertainty in setting the initial parameters of the model and training methods.

Additionally the combined methods are distinguished [3]. They combine the selection and formation of prototypes. The combined methods have the same disadvantages as methods of prototype selection and methods of prototype construction.

Since the prototype construction methods and the hybrid methods related with them are slower than the prototype selection methods, it is advisable to choose the latter as the basis for sampling problem solving.

In order to eliminate the disadvantages of these methods, it is advisable to form a sample without iterative busting of instance combinations by a certain percentage of instance selection from the original sample. This will significantly

reduce the time. Herewith we also need to define the indicators to evaluate the individual instance informativity with regard to their position relatively to the interclass boundaries and to the centers of pseudo-clusters, which. This makes possible to generate a non-random sample, to estimate and guarantee the high quality of selected subsamples.

3 MATERIALS AND METHODS

Let's break feature space into rectangular regions limiting the range of values of each feature by its minimum and maximum values. Then the partition projections into feature axis allow to allocate feature intervals of for each of the rectangular block. The intervals can be formed as cluster projections or as a regular grid, or on the basis of class boundaries in sample one-dimensional projections on the feature axes [24].

Then each such interval can be considered as a term and it is possible to evaluate its importance for decision-making on instance belonging to the cluster with the weight of the k -th term of j -th feature of s -th instance x^s based on a description of the corresponding interval center by the formula (1):

$$w_{C^s_{jk}} = \exp(-(0,5(r_{jk} - l_{jk}) - x^s_j)^2), \quad (1)$$

as well as the weight of the k -th term of j -th feature of the s -th instance x^s relatively to the description of the intercluster boundaries determined by the formula (2):

$$w_{B^s_{jk}} = \exp(-(\min((r_{jk} - x^s_j), (x^s_j - l_{jk}))^2). \quad (2)$$

Then the overall significance of the k -th term of j -th feature of s -th instance x^s relatively to the description of the intercluster boundaries can be estimated using the weight determined by the formula:

$$w^s_{jk} = \max\{w_{C^s_{jk}}, w_{B^s_{jk}}\}.$$

Defining for each s -th instance the term significances, we can also determine the term weights for the whole sample:

$$w_{jk} = \frac{S_{jk}}{SK_{jk}}.$$

Knowing the term significance we can define the feature informativity evaluations by formula (3):

$$w_j = \max_k \{w_{jk}\} \quad (3)$$

or by the formula (4):

$$w_j = \frac{1}{k_j} \sum_{k=1}^{k_j} w_{jk}. \quad (4)$$

It is also possible to use the individual evaluation of the feature informativity in the range [0, 1] defined by the indicators [24].

Based on evaluations of term and feature significance we can determine informativity evaluations for each s -th sample instance by the formula (5):

$$I_1(x^s) = \frac{\sum_{j=1}^N \left(w_j \sum_{k=1}^{k_j} w_{jk} w^s_{jk} \right)}{\sum_{j=1}^N \left(w_j \sum_{k=1}^{k_j} \max_p \{w^p_{jk}\} \right)} \quad (5)$$

or by the formula (6):

$$I_2(x^s) = \frac{1}{N} \sum_{j=1}^N \left(\frac{w_j}{\max_{i=1,2,\dots,N} \{w_i\} k_j} \sum_{k=1}^{k_j} \frac{w_{jk} w^s_{jk}}{\left(\max_{q=1,2,\dots,k_j} \{w_{jk}\} \right) \left(\max_p \{w^p_{jk}\} \right)} \right). \quad (6)$$

Suggested indicators (5) and (6) provide evaluation of individual informativity of instance x^s relatively to the initial sample in the range [0, 1]. The greater the value of corresponding indicator, the more valuable is an instance, and vices versa.

If necessary, the estimates (5) and (6) can be further normalized so that they will give not an absolute but relative value of instance significance in the sample (7):

$$I(x^s) = \frac{I(x^s) - \min\{I(x^p)\}}{\max_p \{I(x^p)\} - \min\{I(x^p)\}}. \quad (7)$$

In this case, the instance with the maximum individual informativity will receive evaluation equal to one, and the instance with minimal informativity will receive evaluation equal to zero. The application of (7) can be useful when it need to simplify the choice of the threshold for separating the sample by the corresponding informativity indicator.

The proposed indicators of evaluation of individual instance significance can be used in the subsample formation from the given original sample by one of the following methods:

1) to form a training subsample of those instances of the original sample, the normalized values of which individual informativity evaluations (7) are greater than some specified threshold;

2) to form a training subsample from the not more than $S' = \delta S$ instances of the original sample with the greatest individual informativity evaluation values;

3) to form a training subsample from the not more than S'/K instances of each class of the original sample with the greatest values of individual informativity evaluations;

4) to use a stochastic search based on evolutionary or multi-agent methods, selecting the best in a some sense combination of instances, using information about individual informativity of instances in the search operators to accelerate the search and focusing it on the most promising solutions.

The first method does not obviously determine the number of instances that will fall into the formed sample. The fourth method is iterative and requires the specification and use of quality indicators, the calculation of which can also be time consuming. Therefore, the second and third methods are the most simple applicable in practice and relatively simple from a computational point of view. They are appropriate to examine together with the proposed measures.

4 EXPERIMENTS

The computer program implementing the proposed method, which complements the «Automated system neural network and neuro-fuzzy model synthesis for non-destructive diagnosis and pattern classification on features» (certificate of copyright registration № 35431 from 21.10.2010) was developed to conduct experiments.

The developed software was studied in solving the Fisher Iris classification problem [25]. The initial data sample contains 150 samples characterized by four input features. The output feature determines instance belonging to one of three classes.

On the basis of the original sample the instance informativity evaluations were obtained and subsets of instances as a training samples were selected by the second and third methods.

To study the second method the 25 %, 50 %, 75 % and 100 % (for the control) instances with the greatest values of individual significance was selected from the whole original sample and included to the training set, respectively. To study the third method the 25 %, 50 %, 75 % and 100 % (for the control) instances with the greatest values of individual significance in each class was selected from the original sample and included to the training set, respectively.

Further, for each sample a model based on a two-layer feed-forward neural network was built. It was trained using the Levenberg-Marquardt method [1]. The number of network inputs was determined by N is the number of features in the corresponding problem. The number of neurons in the second layer of the network corresponds to the number of classes K . The number neurons of the hidden (first) layer was defined as $2K$. All neurons of a network were used the weighted sum as weight (postsynaptic) function and logistic sigmoid as transfer function. The training method parameters were set as follows: the learning rate is 0,01, the allowable number of iterations (epochs) of the method is 1000, the target value of the error function is 10^{-6} .

After neural model training process completion its final characteristics were fixed: the training time t_{tr} and the number of spent training iterations ep_{tr} . After training each model was tested separately on the training and the whole original

samples, for each of which the error was determined, respectively, E_{tr} and E_{all} . Here each error is the number of instances of corresponding sample for which the estimated value did not match the actual value of the output feature.

5 RESULTS

The fragment of the results of conducted experiments is presented in the table 1. Here we use the following notation for the coding method of sampling: G is a regular grid partition, N is an irregular partition based on class boundaries in one-dimensional sample projections on the feature axis, K is instance selection in each class separately, A is instance selection in the whole sample. Calculated instance informativity indicators are encoded as follows: the first digit codes the I calculation method: 1 – by the formula (5), 2 – by the formula (6)); second digit codes the method of w_j calculation: 1 – by the formula (3), 2 – by the formula (4)). For each of the experimentally obtained indicators it is also listed a percentage of formed training sample volume relative to the original sample volume. Markers «min», «average» and «max» are designated, respectively, minimum, average and maximum values.

The table 1 shows that the use of the proposed method of instance significance determining allows in practice to select a subsample of smaller volume from of the original sample, enough to construct neural network models with the required accuracy, reducing the time to build models.

Fig. 1 graphically illustrates the instance placement of the original and formed samples in the space of the first two features (the sepal length in cm on the abscissa axis and the sepal width in cm on the ordinate axis are plotted). Here markers «.», «x» and «+» denote the instances of different classes of the initial sample, a marker «o» indicates instances selected to the training set.

It can be seen from the fig. 1 that the proposed method allows to select the most significant instances of the original sample. In this case the obtained results essentially depends on the subsample formation method, the feature space partitioning method and the method of individual instance informativity evaluation.

6 DISCUSSION

As it evident from the table 1, with the increasing of examples number in the formed sample the accuracy is increased (errors for formed training and for the original samples reducing), the training time and the number of training iterations are increased, and vice versa. At the same time a significant reduction of a sample volume to the 25 % of original leads to deterioration the training process characteristics (the time and number of iterations increase) and also to a decrease in accuracy. This can be explained by the fact that instances critical to describe the class separation can not be included to the sample of small volume.

Even a small reduction of the original sample volume in 25 % (up to 75 % of the original sample volume) yielded acceptable accuracy and reduces training time by more than 1.7 times. Reducing the volume of the original sample by a

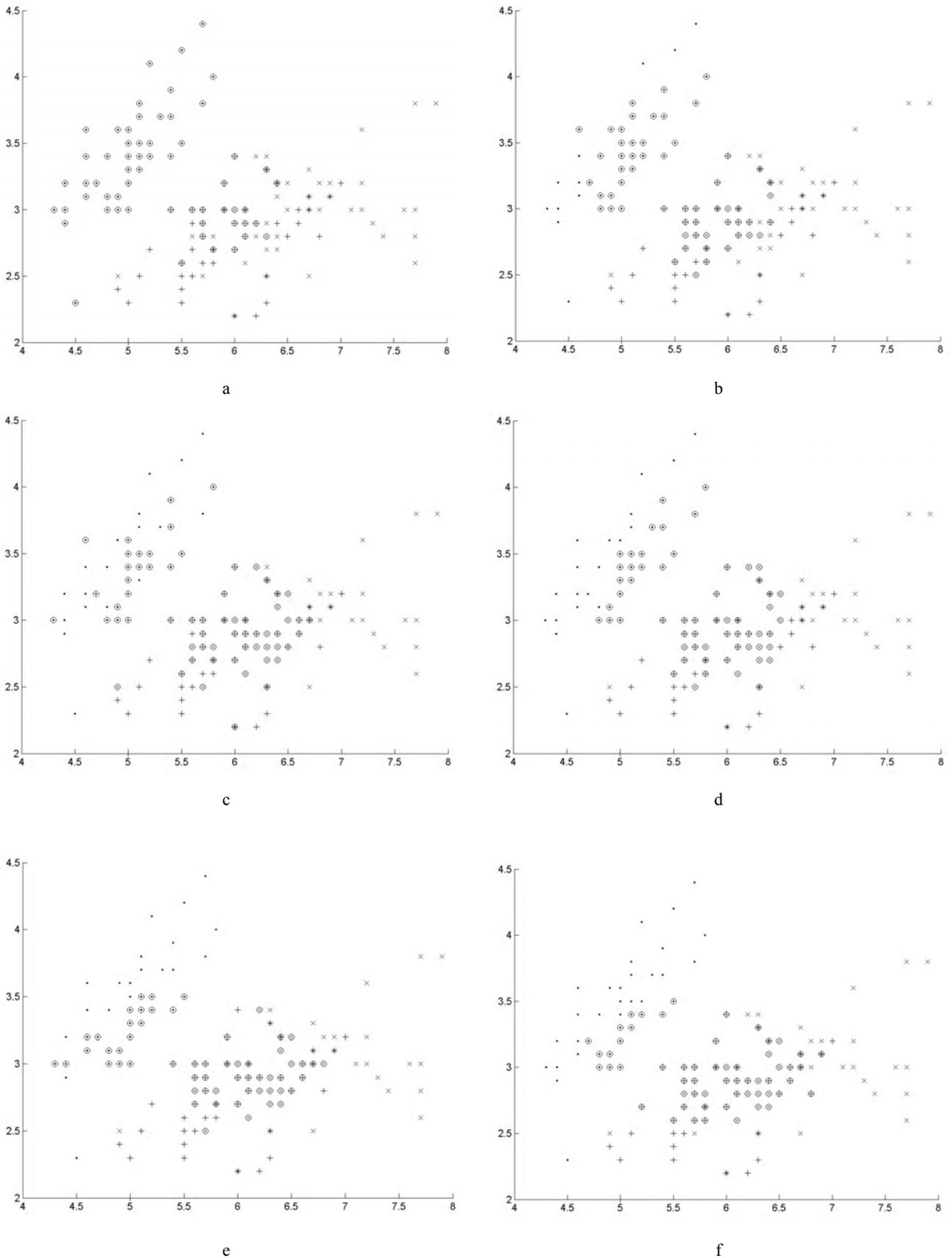


Figure 1 – The example of training samples formation from the original sample for the Fischer iris classification problem in the space of the first two features, where selected 50 % of the instances:

a – I11, G, A; b – I21, G, A; c – I11, G, K; d – I21, G, K; e – I11, N, K; f – I11, N, A

Table 1 – The fragment of experimental results on model building by the formed samples

Method code	I code	$t_{tr. S}$				$ep_{tr.}$				$E_{tr.}$				E_{all}				
		25 %	50 %	75 %	100 %	25 %	50 %	75 %	100 %	25 %	50 %	75 %	100 %	25 %	50 %	75 %	100 %	
G	K	I11	0.094	0.218	0.172	0.655	17	51	28	123	0	0	0.006	0	0.064	0.024	0.004	0
		I12	1.888	0.234	0.343	0.250	630	49	67	33	0	0	0	0	0.091	0.044	0.002	0
		I21	0.094	0.218	2.371	0.125	21	31	583	19	0	0	0	0	0.058	0.031	0	0
		I22	0.125	0.172	0.203	0.203	24	35	38	28	0	0	0	0	0.044	0.049	0.022	0
N	K	I11	0.094	0.546	2.418	1.076	16	137	594	239	0	0	0	0	0.073	0.036	0.004	0
		I12	0.094	0.250	0.203	1.591	15	55	39	325	0	0.009	0	0	0.044	0.040	0.013	0
		I21	0.203	0.374	0.203	4.680	50	92	26	1001	0	0	0	0	0.060	0.018	0.013	0
		I22	0.125	0.140	0.640	2.855	27	25	137	634	0	0	0	0	0.053	0.024	0.024	0
G	A	I11	0.296	0.421	0.218	0.203	7	101	40	172	0	0.027	0	0	0.444	0.169	0.004	0
		I12	0.109	0.562	0.312	0.624	9	137	75	114	0	0	0	0	0.229	0.076	0.004	0
		I21	0.218	2.730	0.343	0.406	68	809	75	90	0.018	0	0	0	0.149	0.044	0.007	0
		I22	0.109	0.109	0.515	0.577	18	21	114	105	0	0	0	0	0.036	0.064	0.029	0
N	A	I11	2.964	0.577	3.931	2.964	1001	151	1001	637	0	0	0	0	0.033	0.013	0	0
		I12	2.980	0.281	0.265	2.714	976	70	62	595	0	0	0	0	0.029	0.033	0.009	0
		I21	0.125	0.234	0.187	1.326	42	52	45	274	0	0	0	0	0.249	0.009	0.004	0
		I22	0.640	1.888	0.218	0.468	199	545	32	101	0	0	0	0	0.240	0.009	0.009	0
min		0.094	0.109	0.172	0.125	7	21	26	19	0	0	0	0	0.029	0.009	0	0	
average		0.635	0.560	0.784	1.295	195	148	185	281	0.001	0.002	0	0	0.119	0.043	0.009	0	
Max		2.980	2.730	3.931	4.680	1001	809	1001	1001	0.018	0.027	0.006	0	0.444	0.169	0.029	0	

half afforded the gain in speed by 2.3 times. This confirms expediency of application of the proposed mathematical support in the neural network model building by precedents.

A method of instance selections in which the subsample is extracted considering the instance significance in the whole original sample (fig. 1a, fig. 1b, fig. 1f), leads to the selection of less informative instances in comparison with the instance selection considering the significance of instances in each class separately (fig. 1c, fig. 1d, fig. 1e). This is because the frequencies of each class instances may be different and in the selection of instances excluding class numbers it is possible to pass a locally important instances. Another cause may be that the instances describing external borders of classes, but do not important for the separation of adjacent classes can be recognized individually as significant, if we ignore their belonging to classes.

We should also note that the method of calculation of individual instance informativity indicators not only quantitatively but also qualitatively effect on the formed sample. It has been established that the indicators I11, defined by formulas (5) and (1), and I12, defined by formulas (5) and (2), respectively, mostly lead to the similar results, which differ significantly from the results for indicators I21, defined by the formulas (6) and (1), and I22, defined by the formulas (6) and (2), respectively. At the same time the indicators I21 and I22 are more resistant to the instance selection method, and the indicators I11 and I12 are the most effective in the instance selection using the instance importance in each class separately.

The significant influence of a feature space partitioning method on the results of the significance evaluation and

selection of instances by the results of conducted experiments can be explained by that the method of irregular partitioning with allocation of class intervals on the axis of each feature [24] allows usually get the best partition in comparison to the regular grid partition method. However, reducing the width of the interval, and correspondingly increasing the number of intervals of each feature axis can improve the latter method results. The selection of the optimum width of the interval is a separate problem that should be carried out taking into account the complexity characteristics of the particular application.

The closest analogue to the proposed method for determining the instance informativity is a set of indicators proposed in [26]. In contrast to the proposed in this paper, the indicators [26] characterize separately instance properties to be informative relatively to the external and internal borders, as well as to the class centers, which is their advantage in the problems of the data visualization and analysis. However, their disadvantages are low speed due to the need to calculate distances between instances, as well as the need and ambiguity of indicator integration to the comprehensive measures of instance informativity.

The advantage of the indicators proposed in this paper is that there is no need to calculate the distances between instances, but disadvantage is the necessary to divide the feature space. However, this disadvantage can be seen as an advantage in the case of large samples: if we use a partition that is simple from a computational point of view (for example, a regular grid) and know the minimum and maximum values of each feature than the computational cost of the proposed indicators will be less than the using of a set [26].

CONCLUSIONS

The urgent problem of mathematical support development is solved to automate the sampling at diagnostic and recognizing model building by precedents.

The method of training sample selection is firstly proposed. It determines the weights characterizing the term and feature usefulness for a given initial sample of precedents and given feature space partition. It characterizes the individual absolute and relative informativity of instances relative to the centers and the boundaries of feature intervals based on the weight values. This allows to automate the sample analysis and its division into subsamples, and, as a consequence, to reduce the training data dimensionality. This in turn reduces the time and provides an acceptable accuracy of neural model training.

The practical significance of obtained results is that the software realizing the proposed indicators is developed, as well as experiments to study their properties are conducted. The experimental results allow to recommend the proposed indicators for use in practice, as well as to determine effective conditions for the application of the proposed indicators.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Zaporizhzhya National Technical University «Intelligent information technologies of automation of designing, simulation, control and diagnosis of manufacturing processes and systems» (state registration number 0112U005350) and by the international project «Centers of Excellence for young REsearchers» of European Commission (№544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

REFERENCES

- Engelbrecht A. Computational intelligence: an introduction / A. Engelbrecht. – Sidney : John Wiley & Sons, 2007. – 597 p. DOI: 10.1002/9780470512517
- Jankowski N. Comparison of instance selection algorithms I. Algorithms survey / N. Jankowski, M. Grochowski // Artificial Intelligence and Soft Computing : 7th International Conference ICAISC-2004, Zakopane, 7–11 June, 2004 : proceedings. – Berlin : Springer, 2004. – P. 598–603. – (Lecture Notes in Computer Science, Vol. 3070). DOI: 10.1007/978-3-540-24844-6_90
- Reinartz T. A unifying view on instance selection / T. Reinartz // Data Mining and Knowledge Discovery. – 2002. – № 6. – P. 191–210. DOI: 10.1023/A:1014047731786
- Hart P. E. The condensed nearest neighbor rule / P. E. Hart // IEEE Transactions on Information Theory. – 1968. – Vol. 14. – P. 515–516. DOI: 10.1109/TIT.1968.1054155
- Aha D. W. Instance-based learning algorithms / D. W. Aha, D. Kibler, M. K. Albert // Machine Learning. – 1991. – № 6. – P. 37–66. DOI: 10.1023/A:1022689900470
- Gates G. The reduced nearest neighbor rule / G. Gates // IEEE Transactions on Information Theory. – 1972, Vol. 18, № 3. – P. 431–433. DOI: 10.1109/TIT.1972.1054809
- Kibbler D. Learning representative exemplars of concepts: an initial case of study / D. Kibbler, D. W. Aha // Machine Learning : 4th International Workshop, Irvine, 22–25 June 1987 : proceedings. – Burlington : Morgan Kaufmann, 1987. – P. 24–30. DOI: 10.1016/b978-0-934613-41-5.50006-4
- Wilson D. L. Asymptotic properties of nearest neighbor rules using edited data / D. L. Wilson // IEEE Transactions on Systems, Man, Cybernetics. – 1972. – Vol. 2, № 3. – P. 408–421. DOI: 10.1109/TSMC.1972.4309137
- Tomek I. An experiment with the edited nearest-neighbor rule / I. Tomek // IEEE Transactions on Systems, Man, and Cybernetics. – 1976. – Vol. 6. – P. 448–452. DOI: 10.1109/TSMC.1976.4309523
- Jankowski N. Data regularization / N. Jankowski // Neural Networks and Soft Computing : Fifth Conference, Zakopane, 6–10 June 2000 : proceedings. – Czkstochowa : Polish Neural Networks Society, 2000. – P. 209–214.
- Broadley C. E. Addressing the selective superiority problem: automatic algorithm/model class selection / C. E. Broadley // Machine Learning : Tenth International Conference, Amherst, 27–29 June, 1993 : proceedings. – Burlington : Morgan Kaufmann, 1993. – P. 17–24. DOI: 10.1016/b978-1-55860-307-3.50009-5
- Wilson D. R. Reduction techniques for instancebased learning algorithms / D. R. Wilson, T. R. Martinez // Machine Learning. – 2000. – Vol. 38, № 3. – P. 257–286. DOI: 10.1023/A:1007626913721
- An algorithm for a selective nearest neighbor decision rule / [G. L. Ritter, H. B. Woodruff, S. R. Lowry, T. L. Isenhour] // IEEE Transactions on Information Theory, 1975. – Vol. 21, № 6. – P. 665–669. DOI: 10.1109/TIT.1975.1055464
- Domingo C. Adaptive sampling methods for scaling up knowledge discovery algorithms / C. Domingo, R. Gavaldá, O. Watanabe // Discovery Science : Second International Conference, DS'99 Tokyo, 6–8 December 1999 : proceedings. – Berlin : Springer, 1999. – P. 172–183. DOI: 10.1007/3-540-46846-3_16
- Skalak D. B. Prototype and feature selection by sampling and random mutation hill climbing algorithms / D. B. Skalak // Machine Learning : Eleventh International Conference, New Brunswick, 10–13 July 1994 : proceedings. – Burlington : Morgan Kaufmann, 1994. – P. 293–301. DOI: 10.1016/b978-1-55860-335-6.50043-x
- Kohonen T. Learning vector quantization / T. Kohonen // Neural Networks. – 1988. – Vol. 1, P. 303 DOI: 10.1016/0893-6080(88)90334-6
- Likelihood-based data squashing: a modeling approach to instance construction / [D. Madigan, N. Raghavan, W. DuMouchel, M. Nason, C. Posse, G. Ridgeway] // Data Mining and Knowledge Discovery. – 2002. – Vol. 6, № 2. – P. 173–190. DOI: 10.1023/A:1014095614948
- Support cluster machine / [B. Li, M. Chi, J. Fan, X. Xue] // Machine Learning : 24th International Conference, Corvallis, 20–24 June 2007 : proceedings. – New York, 2007. – P. 505–512. DOI: 10.1145/1273496.1273560
- Evans R. Clustering for classification: using standard clustering methods to summarise datasets with minimal loss of classification accuracy / R. Evans. – Saarbrücken: VDM Verlag, 2008. – 108 p.
- Sane S. S. A Novel supervised instance selection algorithm / S. S. Sane, A. A. Ghatol // International Journal of Business Intelligence and Data Mining. – 2007. – Vol. 2, № 4. – P. 471–495. DOI: 10.1504/IJBIDM.2007.016384

21. Reeves C. R. Using genetic algorithms for training data selection in RBF networks / C. R. Reeves, D. R. Bush // Instance Selection and Construction for Data Mining / Eds.: H. Liu, H. Motoda. – Norwell : Kluwer, 2001. – Part VI. – P. 339–356. DOI: 10.1007/978-1-4757-3359-4_19
22. Suykens J. A. Least squares support vector machine classifiers / J. A. Suykens, J. Vandewalle // Neural Processing Letters. – 1999. – Vol. 9, № 3. – P. 293–300. DOI: 10.1023/A:1018628609742
23. Koskimaki H. Two-level clustering approach to training data instance selection: a case study for the steel industry / H. Koskimaki, I. Juutilainen, P. Laurinen, J. Roning // Neural Networks : International Joint Conference (IJCNN-2008), Hong Kong, 1–8 June 2008 : proceedings. – Los Alamitos: IEEE, 2008. – P. 3044–3049. DOI: 10.1109/ijcnn.2008.4634228
24. UCI machine learning repository [Electronic resource]. – Access mode: <http://archive.ics.uci.edu/ml/datasets/>
25. Subbotin S. The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition / S. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2013. – Vol. 22, № 2. – P. 97–103. DOI: 10.3103/s1060992x13020082
26. Subbotin S. A. Methods of sampling based on exhaustive and evolutionary search / S. A. Subbotin // Automatic Control and Computer Sciences. – 2013. – Vol. 47, № 3. – P. 113–121. DOI: 10.3103/s0146411613030073

Article was submitted 03.08.2014.

Субботин С. А.

Д-р техн. наук, профессор, Запорожский национальный технический университет, Украина

ОЦЕНИВАНИЕ ИНДИВИДУАЛЬНОЙ ИНФОРМАТИВНОСТИ ЭКЗЕМПЛЯРОВ ДЛЯ ФОРМИРОВАНИЯ ВЫБОРОК ПРИ ПОСТРОЕНИИ НЕЙРОМОДЕЛЕЙ

Решена задача разработки математического обеспечения для автоматизации формирования выборок при построении диагностических и распознающих моделей по прецедентам. Объектом исследования являлся процесс построения диагностических и распознающих нейромоделей по прецедентам. Предмет исследования составляют методы формирования выборок для построения нейромоделей по прецедентам. Цель работы – повысить скорость процесса формирования и качество выделяемых обучающих выборок для построения нейромоделей по прецедентам. Предложен метод формирования обучающих выборок, который для заданной исходной выборки прецедентов и заданного разбиения пространства признаков определяет веса, характеризующие полезность термов и признаков, с учетом значений которых позволяет охарактеризовать индивидуальную абсолютную и относительную информативность экземпляров относительно центров и границ интервалов признаков, что позволяет автоматизировать анализ выборки и ее разделение на подвыборки, и, как следствие, сократить размерность обучающих данных, что, в свою очередь, позволит сократить время и обеспечить приемлемую точность обучения нейромоделей. Разработано программное обеспечение, реализующее предложенные показатели. Проведены эксперименты по исследованию их свойств. Результаты экспериментов позволяют рекомендовать предложенные показатели для использования на практике, а также определять эффективные условия применения предложенных показателей.

Ключевые слова: выборка, отбор экземпляров, редукция данных, нейронная сеть, сокращение размерности данных.

Субботін С. О.

Д-р техн. наук, професор, Запорізький національний технічний університет, Україна

ОЦІНЮВАННЯ ІНДИВІДУАЛЬНОЇ ІНФОРМАТИВНОСТІ ЕКЗЕМПЛЯРІВ ДЛЯ ФОРМУВАННЯ ВИБІРОК ПРИ ПОБУДОВІ НЕЙРОМОДЕЛЕЙ

Вирішено завдання розробки математичного забезпечення для автоматизації формування вибірок при побудові діагностичних і розпізнавальних моделей за прецедентами. Об'єктом дослідження був процес побудови діагностичних і розпізнавальних нейромоделей за прецедентами. Предмет дослідження становлять методи формування вибірок для побудови нейромоделей за прецедентами. Мета роботи – підвищити швидкість процесу формування та якість виділюваних навчальних вибірок для побудови нейромоделей за прецедентами. Запропоновано метод формування навчальних вибірок, який для заданої вихідної вибірки прецедентів і заданого розбиття простору ознак визначає ваги, що характеризують корисність термів і ознак, з урахуванням значень яких дозволяє охарактеризувати індивідуальну абсолютну і відносну інформативність примірників щодо центрів і меж інтервалів ознак, що дозволяє автоматизувати аналіз вибірки і її поділ на підвибірки, і, як наслідок, скоротити розмірність навчальних даних, що, у свою чергу, дозволяє скоротити час і забезпечити прийнятну точність навчання нейромоделей. Розроблено програмне забезпечення, що реалізує запропоновані показники. Проведені експерименти з дослідження їхніх властивостей. Результати експериментів дозволяють рекомендувати запропоновані показники для використання на практиці, а також визначити ефективні умови застосування запропонованих показників.

Ключові слова: вибірка, відбір екземплярів, редукція даних, нейронна мережа, скорочення розмірності даних.

REFERENCES

1. Engelbrecht A. Computational intelligence: an introduction. Sidney, John Wiley & Sons, 2007, 597 p. DOI: 10.1002/9780470512517
2. Jankowski N., Grochowski M. Comparison of instance selection algorithms I. Algorithms survey, *Artificial Intelligence and Soft Computing : 7th International Conference ICAISC-2004, Zakopane, 7–11 June, 2004 : proceedings*. Berlin, Springer, 2004, pp. 598–603. – (Lecture Notes in Computer Science, Vol. 3070). DOI: 10.1007/978-3-540-24844-6_90
3. Reinartz T. A unifying view on instance selection, *Data Mining and Knowledge Discovery*, 2002, No. 6, pp. 191–210. DOI: 10.1023/A:1014047731786
4. Hart P. E. The condensed nearest neighbor rule, *IEEE Transactions on Information Theory*, 1968. Vol. 14, pp. 515–516. DOI: 10.1109/TIT.1968.1054155
5. Aha D. W., Kibler D., Albert M. K. Instance-based learning algorithms, *Machine Learning*, 1991, No. 6, pp. 37–66. DOI: 10.1023/A:1022689900470

6. Gates G. The reduced nearest neighbor rule, *IEEE Transactions on Information Theory*, 1972, Vol. 18, No. 3, pp. 431–433. DOI: 10.1109/TIT.1972.1054809
7. Kibbler D., Aha D. W. Learning representative exemplars of concepts: an initial case of study, *Machine Learning : 4th International Workshop*, Irvine, 22–25 June 1987, proceedings. – Burlington, Morgan Kaufmann, 1987, pp. 24–30. DOI: 10.1016/b978-0-934613-41-5.50006-4
8. Wilson D. L. Asymptotic properties of nearest neighbor rules using edited data, *IEEE Transactions on Systems, Man, Cybernetics*, 1972, Vol. 2, No. 3, pp. 408–421. DOI: 10.1109/TSMC.1972.4309137
9. Tomek I. An experiment with the edited nearest-neighbor rule, *IEEE Transactions on Systems, Man, and Cybernetics*, 1976, Vol. 6, pp. 448–452. DOI: 10.1109/TSMC.1976.4309523
10. Jankowski N. Data regularization, *Neural Networks and Soft Computing : Fifth Conference*, Zakopane, 6–10 June 2000 : proceedings, Czestochowa, Polish Neural Networks Society, 2000, pp. 209–214.
11. Broadley C. E. Addressing the selective superiority problem: automatic algorithm/model class selection, *Machine Learning*, Tenth International Conference, Amherst, 27–29 June, 1993, proceedings, Burlington, Morgan Kaufmann, 1993, pp. 17–24. DOI: 10.1016/b978-1-55860-307-3.50009-5
12. Wilson D. R., Martinez T. R. Reduction techniques for instancebased learning algorithms, *Machine Learning*, 2000, Vol. 38, No. 3, pp. 257–286. DOI: 10.1023/A:1007626913721
13. Ritter G. L., Woodruff H. B., Lowry S. R., Isenhour T. L. An algorithm for a selective nearest neighbor decision rule, *IEEE Transactions on Information Theory*, 1975, Vol. 21, No. 6, pp. 665–669. DOI: 10.1109/TIT.1975.1055464
14. Domingo C., Gavalda R., Watanabe O. Adaptive sampling methods for scaling up knowledge discovery algorithms, *Discovery Science : Second International Conference, DS'99* Tokyo, 6–8 December 1999 : proceedings. Berlin. Springer, 1999, pp. 172–183. DOI: 10.1007/3-540-46846-3_16
15. Skalak D. B. Prototype and feature selection by sampling and random mutation hill climbing algorithms, *Machine Learning, Eleventh International Conference*, New Brunswick, 10–13 July 1994 : proceedings, Burlington, Morgan Kaufmann, 1994, pp. 293–301. DOI: 10.1016/b978-1-55860-335-6.50043-x
16. Kohonen T. Learning vector quantization, *Neural Networks*, 1988, Vol. 1, pp. 303 DOI: 10.1016/0893-6080(88)90334-6
17. Madigan D., Raghavan N., DuMouchel W., Nason M., Posse C., Ridgeway G. Likelihood-based data squashing: a modeling approach to instance construction, *Data Mining and Knowledge Discovery*, 2002, Vol. 6, No. 2, pp. 173–190. DOI: 10.1023/A:1014095614948
18. Li B., Chi M., Fan J., Xue X. Support cluster machine, *Machine Learning*, 24th International Conference, Corvallis, 20–24 June 2007, proceedings. New York, 2007, pp. 505–512. DOI: 10.1145/1273496.1273560
19. Evans R. Clustering for classification: using standard clustering methods to summarise datasets with minimal loss of classification accuracy. Saarbrücken, VDM Verlag, 2008, 108 p.
20. Sane S. S., Ghatol A. A. A Novel supervised instance selection algorithm, *International Journal of Business Intelligence and Data Mining*, 2007, Vol. 2, No. 4, pp. 471–495. DOI: 10.1504/IJBIDM.2007.016384
21. Reeves C. R., Bush D. R. Using genetic algorithms for training data selection in RBF networks, *Instance Selection and Construction for Data Mining*, Eds.: H. Liu, H. Motoda. Norwell, Kluwer, 2001, Part VI, pp. 339–356. DOI: 10.1007/978-1-4757-3359-4_19
22. Suykens J. A., Vandewalle J. Least squares support vector machine classifiers, *Neural Processing Letters*, 1999, Vol. 9, No. 3, pp. 293–300. DOI: 10.1023/A:1018628609742
23. Koskimaki H., Juutilainen I., Laurinen P., Roning J. Two-level clustering approach to training data instance selection: a case study for the steel industry, *Neural Networks : International Joint Conference (IJCNN-2008)*, Hong Kong, 1–8 June 2008 : proceedings. Los Alamitos, IEEE, 2008, pp. 3044–3049. DOI: 10.1109/ijcnn.2008.4634228
24. UCI machine learning repository [Electronic resource]. – Access mode: <http://archive.ics.uci.edu/ml/datasets/>
25. Subbotin S. The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition, *Optical Memory and Neural Networks (Information Optics)*, 2013, Vol. 22, No. 2, pp. 97–103. DOI: 10.3103/s1060992x13020082
26. Subbotin S. A. Methods of sampling based on exhaustive and evolutionary search, *Automatic Control and Computer Sciences*, 2013, Vol. 47, No. 3, pp. 113–121. DOI: 10.3103/s0146411613030073