e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

UDC 004.93

THE FRACTAL ANALYSIS OF SAMPLE AND DECISION TREE MODEL

Subbotin S. A. – Dr. Sc., Professor, Head of the Department of Software Tools, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Gofman Ye. A. – PhD, Senior Researcher of the Research Unit, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

ABSTRACT

Context. The problem of decision tree model synthesis using the fractal analysis is considered in the paper. The object of study is a decision trees. The subject of study is a methods of decision tree model synthesis and analysis.

Objective. The objective of the paper is a creation of methods and fractal indicators allowing jointly solving the problem of decision tree model synthesis and the task of reducing the dimension of training data from a unified approach based on the principles of fractal analysis.

Method. The fractal dimension for a decision tree based model is defined as for whole training sample as for specific classes. The method of the fractal dimension of a model based on a decision tree estimation taking into account model error is proposed. It allows to built model with an acceptable error value, but with optimized level of fractal dimensionality. This makes possibility to reduce decision tree model complexity and to make it mo interpretable. The set of indicators characterizing complexity of decision tree model is proposed. The set of indicators characterizing complexity of decision tree model is proposed. It contains complexity of node checking, complexity of node achieving, an average model complexity and worst tree model complexity of computations. On the basis of proposed set of indicators a complex criterion for model building is proposed. The indicators of the fractal dimension of the decision tree model error can be used to find and remove the non-informative features in the model.

Results. The developed indicators and methods are implemented in software and studied at practical problem solving. As results of experimental study of proposed indicators the graphs of their dependences were obtained. They include graphs of dependencies of number of hyperblocks covering the sample in the features space from size of block side: for whole sample, for each class, for different set error values and obtained error values, for varied values of resulted number of features and instances, also as graphs of dependencies between average and worst tree complexities, decision tree fractal dimensionality and tree average complexity, joint criterion and indicator of feature set reduction, and between joint criterion and tree fractal dimensionality/

Conclusions. The conducted experiments confirmed the operability of the proposed mathematical support and allow recommending it for use in practice for solving the problems of model building by the precedents.

KEYWORDS: decision tree, sample, fractal dimension, indicator, tree complexity.

NOMENCLATURE

ε is a maximum acceptable error value;

ω is a set of model parameters;

c_i^a is a complexity of achieving of *i*-th leaf node;

c_i is a complexity of checking for *i*-th node it's can be obtained as number of *i*-th node's successors;

c_{tree}^w is a worse complexity of computations for the tree model;

c_{tree}^a is an average complexity of computations for the tree model;

D is a fractal dimension;

$<Dc>$ is a correlation dimension;

D_{tree} is a data fractal dimension relatively the accuracy (error) of the synthesized model;

$D^{(k)}$ is a fractal dimension of *k*-th class;

D is a fractal dimension of the sample;

E is a model error;

f is a model quality criterion;

$F()$ is a model structure;

F_{tree} is a joint multiplicative criterion for decision tree model;

I_N is a coefficient of features reduction;

j is a number of feature;

K is a number of classes;

DOI 10.15588/1607-3274-2020-1-11

K is a number of classes;

L a number of intervals on which the ranges of feature values will be separated;

l is a hypercube side length;

l is a length of the interval;

L is a number of intervals;

N is an number of input features;

N' is a feature subset size;

$n(l)$ is a number of hyperblocks of side with the *l* size covering the sample;

$n_{i,q}$ is a number of instances belonging to a rectangular hyperblock formed by feature intervals;

$n_{i,q,k}$ is a number of *k*-th classes exemplars, which are given in each rectangular hyperblock formed by features intervals;

$n(l)$ is a number of hyperblocks with the side of *l* size covering the sample;

$n_k(l)$ is a number of hyperblocks with the side of *l* size covering the sample for *k*-th class;

$n_{(k)}$ is a number of hyperblocks with the side of *l*-size covering the *k*-th class of the sample;

opt is a symbol of optimum;

Q is a number of clusters;

r is a heuristically defined cut-off radius;

r_k is a Euclidean distance between pair of points;

S' is a subsample size;

S is a number of precedents;

tree is a tree recognizing model;

U is a total number of tree nodes;

u_i is a type of the *i*-th node of the tree;

X is a data sample;

x_j is a *j*-th input feature;

x^s is a *s*-th instance of a sample;

x^s_j is a value of *j*-th input for *s*-th instance;

$x_j^{\max}$ is a maximal value of x_j;

$x_j^{\max}$ is a minimal value of x_j;

Y^i is a class of majority of instances hit to the *i*-th node, which is a leaf;

y is an output feature vector;

y^s is a value of output feature values for *s*-th instance.

INTRODUCTION

Decision trees are a popular tool for solving problems of building models on precedents in diagnostics, pattern recognition, and forecasting in various practical areas [1–4]. One of the most significant advantages of models based on decision trees is their interpretability (convenience for human perception and analysis).

The **object of study** is a decision trees.

It is now known a large number of methods to synthesize a model based on decision trees [5–11]. However, as a rule, the known methods in their goal functions (the criteria for the training quality) do not take into account the characteristics of the training sample. This in practice can lead to the construction of non-optimal models.

On the other hand, the model synthesis for big data sets unusually requires the preliminary reduction of the data dimensionality size, which is explained by the high iterativity of the known training methods, as well as the need to obtain a model that provides a good generalization of the data. At the same time, the traditionally used methods of informative feature selection [12–15] and of sample formation [16–21] have such common disadvantage as they are not directly related to each other and come from different points of view on the informativeness of features or instances.

The **subject of study** is methods of decision tree model synthesis and analysis.

One of the promising areas of data analysis is a fractal analysis [22–31]. There are various approaches to the definition of fractal parameters for data [25, 27]. However, they are also not interconnected with each other and with the decision tree model training process.

The **objective** of the paper is a creation of methods and fractal indicators allowing jointly solving the problem of decision tree model synthesis and the task of reducing the dimension of training data from a unified approach based on the principles of fractal analysis.

1 PROBLEM STATEMENT

Let we have an original data sample $X = \langle x, y \rangle$ a set of *S* precedents (instances. exemplars, observations) characterizing dependence $y(x)$, where $x=\{x^s\}$, $y=\{y^s\}$, $s=1, 2, \ldots S$, characterized by the set of *N* input features $\{x_j\}$, $j = 1, 2, \ldots, N$, and output feature *y*. Each *s*-th precedent can be noted as $\langle x^s, y^s \rangle$, $x^s=\{x^s_j\}$, $y^s \in \{1, 2, \ldots, K\}$, $K>1$.

Then the problem of model synthesis of the dependence $y(x)$ will be considered in a search of such structure $F()$ and adjusting such values of parameters ω of a model $\langle F(), \omega \rangle$, which will satisfy the model quality criterion $f(F(), \omega, \langle x, y \rangle) \rightarrow opt$. Usually, the model quality criterion is defined as a model error [2]:

$$E = \frac{1}{2}\sum_{s=1}^{S}(y^s - F(\omega, x^s))^2 \rightarrow \min.$$

2 REVIEW OF THE LITERATURE

The key concept in a fractal analysis is a fractal dimension, which is defined as coefficient describing the fractal structure or the set on the basis of a quantitative assessment of its complexity as the coefficient of variation in details and with a scale conversion.

The Hausdorff-Besicovich dimension according to [25, 28] is defined as

$$D \approx \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

One of the most affordable ways to determine the Hausdorff-Besicovich dimension is box-counting method [29, 30], which consists in repeating fractal object coating

DOI 10.15588/1607-3274-2020-1-11

by hypercubes of equal size and counting minimum number of hypercubes which contain points of the object.

By consistently reducing the hypercubes size l we will get a set of points with coordinates $(\log(n(l)),\ \log(l^{-1}))$, which define a curve, which slope determined by the linear regression, is a fractal dimension:

$$D = \lim_{l \to 0} \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

The Takens' method [31] is used to determine the correlation dimension:

$$\langle Dc \rangle = -\left\{ \frac{1}{|R|} \sum_{k=1}^{|R|} r_k \right\}^{-1},$$

where $R = \{r_k |\ r_k < r\}$, $|R|$ is a cardinality of the set R, $r>0$.

The common disadvantage of considered methods [22–31] for determining the fractal dimension is that the cardinality of the set must satisfy the inequality $N<2\log_{10}S$, which shows that the number of data points S required for accurate dimension estimation of the N-dimensional set must be at least $10^{\frac{N}{2}}$. It leads to large N values even for small sets.

The common feature of all above described methods for determining the fractal dimension is that sample dimension and dimension of the model trained on its basis are defined with no connection to each other. It limits their practical application.

In [32] the methods for estimating the fractal dimension allowing characterize properties of the sample. The sample instances are represented as points in the feature space. Then clusters will correspond to the compact areas in a feature space, which will be combined into classes. Different geometric shapes can describe clusters. Fractal analysis of the sample in the feature space can be performed by setting the elemental form for clustering and varying the size of the cluster for partitioning the sample into fragments. For the sample fractal dimension analysis the method [32] contains following stages.

Initialization stage. Set a learning sample $<x,y>$ and L the number of intervals on which the ranges of feature values will be separated.

Sample normalization stage. If feature values are non-normalized, they should be normalized by mapping to the interval [0, 1]: $x_j^s = \left(x_j^s - x_j^{\min}\right) / \left(x_j^{\max} - x_j^{\min}\right)$.

Clustering stage. Divide the range of each feature values on L intervals of length l: $l = 1/L$. Form clusters as rectangular blocks at the different features interval intersection.

Data analysis stage. Determine the number of instances belonging to a rectangular hyperblock formed by feature intervals $n_{i,q}$. Determine the number of k-th classes exemplars, which is given in each rectangular hyperblock formed by features intervals $n_{i,q,k}$.

Determine the number of hyperblocks with the side of l-size covering the k-th class of the sample in the N features space:

$$n_{(k)} = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \mid n_{i,q,k} > 0 \right\}.$$

Determine the number of hyperblocks with the side of l size covering the sample in the N features space:

$$n(l) = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \mid n_{i,q} > 0 \right\} = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \left| \sum_{k=1}^{K} n_{i,q,k} > 0 \right. \right\}.$$

Stage of fractal dimension estimation. Determine at a given l the fractal dimension of k-th class, $k = 1, 2, ..., K$: $D^{(k)} = \log(n_{(k)}) / \log(l^{-1})$.

Determine the fractal dimension of the sample at a given l: $D = \log(n(l)) / \log(l^{-1})$.

This method operates with rectangular blocks of the same size, covering the feature space by them. The single controlled parameter of the method is defined by the number of intervals L, which are divided in ranges of the feature values.

It is obvious that number of clusters $Q \geq K$, $Q=L^N$, and for each feature $L \geq 2$. To provide generalization properties of clusters we impose restriction $Q \leq NS$.

Thus, we obtain $K \leq L^N \leq NS$, $L \geq 2$. Taking logarithms $\log(K) \leq N\log(L) \leq \log(NS)$ we obtain after transformations: $2 \leq L \leq \sqrt[N]{NS}$. Note that minimum step for varying the L values is 1. If the upper limit value $\sqrt[N]{NS}$ is less than 2, it can be replaced with S. This is due to the fact that on the each feature axis will not be more than the S points and feature axis partition on more than S intervals will obviously lead to occurrence of the empty intervals. For large N values, the given number of partitions S on each feature will lead to forming of a huge number of blocks equal S^N, which make a computation very hard, and in some cases practically non-realizable. Therefore, it is reasonable in this case to set the value of the upper limit of L by the round(log(S)), where round is function of rounding to the nearest integer number. Evaluation of indicator D for small values of L requires high cost of computing resources and computer memory resources than for large L values. However, the analysis accuracy for small L values will be lower while the generalization level will be higher than for large L values.

Consider possible ways to implement this method. If we assume that the data structure will be created containing the counters of instances numbers belonging to each of rectangular hyper-block in the feature space, it will require at least $2L^N$ memory cells where 2 bytes will be given to represent L^N integers. In turn, for each hyper-block we need to evaluate belonging of the sample instances, which would require about $2SL^N$ comparisons. This approach, obviously, is practically applicable only for small N. Since to determine the fractal dimension it is not important to know how many instances hit in each

DOI 10.15588/1607-3274-2020-1-11

block, but it is important to know how many blocks contains instances, then to reduce the computational and memory costs are encouraged to use the following approach.

The advantage of the described method and of the sample quality indicator determined on its basis is the fact that they does not depend of the model synthesis method, and of the results of its work and allow to evaluate the properties of the single sample.

The disadvantages of this method are the uncertainty in the choice of the L parameter value, and absence of relation between the method and the quality of the synthesized model.

3 MATERIALS AND METHODS

The decision tree model consists of nodes connected by the links. The node can be a root (having no parents), a leaf (having no successors), or an internal (having parent and successors nodes). Each node of the tree (excluding leafs) contains check on one of the features. As a result of checking the recognized instance on this node, it will be redirected to one of the successor nodes of this node, depending on what interval of checked feature values it falls into.

For a decision tree based model, we define the fractal dimension as the minimum number of rectangular blocks in the feature space needed to cover the training data set. Since the leaf nodes of the model based on the decision tree correspond to rectangular areas in the feature space, and the instances of the training sample belong by the model only to these areas, the number of leaf nodes in the tree is the fractal dimension of the decision tree.

Let u_i is a type of the i-th node of the tree (u_i = 1 if i-th node is a leaf; u_i = 0, otherwise), U is a total number of tree nodes. Then the number of hyperblocks with the side of l size covering the sample in the normalized feature space can be evaluated as:

$$n(l) = \sum_{i=1}^{U} u_i \ .$$

By analogy for k-th class in the sample we can define:

$$n_k(l) = \sum_{i=1}^{U} \left\{ 1 \mid u_i = 1, Y^i = k \right\}, \ k = 1, 2, ..., K,$$

where Y^i is a class of majority of instances hit to the i-th node, which is a leaf, K is a number of classes.

It is obviously, that

$$n(l) = \sum_{k=1}^{K} n_k(l) \ .$$

To estimate the fractal dimension of a model based on a decision tree, we will use an approach similar to neural networks [33–35].

Initialization stage. Set the training sample <x, y>, the model synthesis method, the model training quality criterion as error function E, and the maximum acceptable error value ε.

Sample normalization stage. If feature values are non-normalized, they should be normalized by mapping to the interval [0, 1].

Formation and analysis of data partition stage. Sequentially changing the value of L = 2, ..., S:

– determine the length of the interval l;

– quantize the sample features, partitioning their ranges of values on L intervals;

– determine the number of hyperblocks of side with the l size covering the sample in the space of the N features $n(l)$;

– prune a recognizing model *tree* by a given method, minimizing the error function E to achieve an acceptable level ε;

– estimate the error E of the constructed recognizing model *tree*.

The fractal dimension determining stage. For every l, for which the model error E is acceptable, determine the data fractal dimension relatively the accuracy (error) of the synthesized model *tree*:

$$D_{tree} = \left\{ \left. \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)} \right| E(tree) \le \varepsilon \right\}.$$

This method operates by rectangular blocks of equal size, covering by them the feature space. The single controlled parameter of the method is given threshold value of the model error ε.

Obviously, the smaller the given ε value, the more detailed model should be, i.e., it will need to form a larger number of clusters Q, and hence the greater should be the L value. Accordingly, with a decrease of the given ε, the cost of computing resources and computer memory resources will increase for the sample analysis.

The advantage of the proposed method and sample quality indicator determined on its basis is the fact that they are related with quality indicator of the synthesized model, and automatically sets the optimum value of the L.

The disadvantages of the proposed method are the uncertainty of ε parameter values choice and its dependence on the training quality and model functioning principles on which it is defined. It should also be noted that the error function used in the method is only one of the synthesized model characteristics, but it does not take into account the model dimension and generalizing properties.

Therefore, the fractal dimension of the trained model is proposed to be determined on the basis of the below method taking into account the model dimension.

In addition to the tree fractal dimensionality we can take into account complexity of calculations.

For i-th node it's complexity of checking c_i can be obtained as number of i-th node's successors.

DOI 10.15588/1607-3274-2020-1-11

For *i*-th leaf node the complexity of achieving c_i^a can be evaluated as a sum of complexities of checking of all nodes in the path from the tree root to the *i*-th leaf node.

For the tree model the worse complexity of computations can be estimated as the maximal complexity of the leaf nodes:

$$c_{tree}^w = \max_{i=1,2,...,U} \{c_i^a \mid u_i = 1\}.$$

For the tree model the average complexity of computations can be estimated as the average complexity of the leaf nodes:

$$c_{tree}^a = \frac{1}{n(l)} \sum_{i=1}^{U} \{c_i^a \mid u_i = 1\}.$$

Generally, when the model error is acceptable, we need to minimize the average computational complexity of leaf nodes reaching, as well as minimize the worst complexity of leaf nodes reaching.

For the decision tree model synthesis the proposed set of fractal indicators allows to define a system of criterions:

$$\begin{cases} D_{tree} \to \min \\ c_{tree}^a \to \min \\ c_{tree}^w \to \min \end{cases}.$$

Since in the general case the number of features in the original set and the number of features used by the model based on the decision tree may differ, it is advisable to consider it as a characteristic of the model quality. Then it is possible to determine on its basis the coefficient of features reduction as:

$$I_N = \frac{N}{N'}, 1 \le N' \le N.$$

Obviously, the greater the coefficient of feature reduction, the simpler the model, provided that acceptable accuracy is achieved. In the best case $I_N = N$, in the worst case $I_N = 1$.

It is possible also to define one joint multiplicative criterion for decision tree model synthesis based on the fractal analysis as:

$$F_{tree} = \frac{D_{tree} c_{tree}^a c_{tree}^w}{I_N} \to \min.$$

Using proposed fractal indicators as individual, and as combined it is possible to solve different tasks in the process of decision tree model synthesis from unified point of view.

4 EXPERIMENTS

To study the complex of proposed sample and model fractal indicators they were implemented in software. The developed software was used in computational experiments to study the applicability of proposed indicators for solving the problems of automatic classification.

Several datasets for different tasks [4, 36–38] characterized in Table 1 were used for experimental study.

Table 1 – Characteristics of the tasks for methods experimental study

Task	Source	*N*	*S*	*K*
Fisher Iris	[36]	4	150	3
Agricultural plant classification on remote sensing data	[37]	55	248	2
Diagnosis of Arrhythmia	[38]	279	452	2
Air-engine blade diagnosis	[4]	10240	32	2

For this datasets several series of experiments were conducted.

The first series of experiments were devoted to study the methods of data dimensionality reduction using fractal indicators for model synthesis. Here it is needed to evaluate fractal dimensionalities of original datasets and their classes. Then is possible to study dependencies of $n(l)$ from l^{-1} for the entire sample and the classes, for different given ε values and obtained values of error E, as well as dimensions of the formed data subsample: subsample size S' and feature subset size N'.

The second series of experiments were concerned to study the methods of decision tree model synthesis using fractal indicators. For each task we need to built a tree model and study dependencies between sample properties and proposed indicators.

5 RESULTS

For each data set as a result of the experiments, the fractal dimensions of the data and the decision tree models constructed on their basis are calculated.

For example, the computed fractal dimension of the sample for the Fisher Iris data set [36] is $D = 0.59034$, and fractal dimension assessments of the classes: $D^{(1)} = 0.68223$, $D^{(2)} = 0.6212$, $D^{(3)} = 0.53407$.

Graphs of dependencies from l^{-1} in a logarithmic system of coordinates for the entire sample and the classes are shown in Fig. 1 and Fig. 2, respectively. On Fig. 2 markers of different sizes encode the different classes.

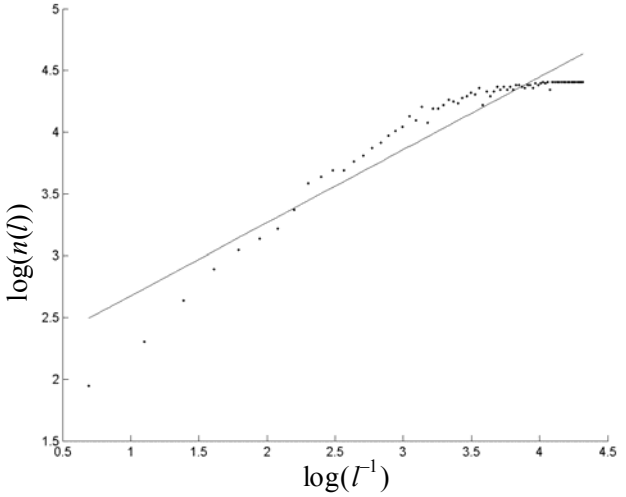

Figure 1 – Graph of dependency of $n(l)$ of a sample from from l^{-1} in a logarithmic coordinate system

DOI 10.15588/1607-3274-2020-1-11

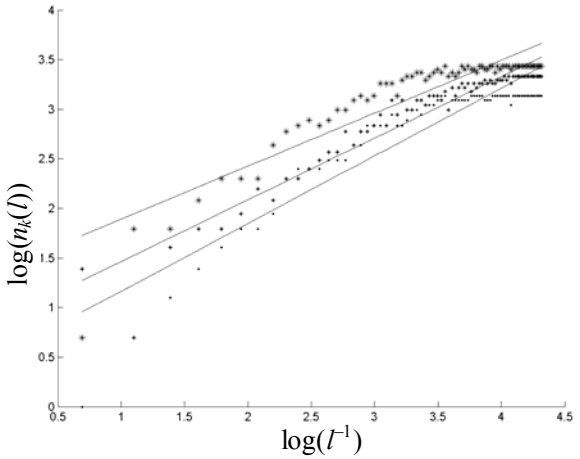

Figure 2 – Graphs of dependencies of $n_k(l)$ of classes from l^{-1} in a logarithmic coordinate system

Fig. 3 shows the schematic graph of generalized dependencies of $n(l)$ of sample from l^{-1} in a logarithmic system of coordinates for different set values of ε and obtained values E.

Fig. 4 presents the schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system for varied values of N' and S'.

Fig. 5 shows the schematic graph of generalized dependencies between c^a_{tree} and c^w_{tree}.

Fig. 6 presents the schematic graph of generalized dependencies between D_{tree} and c^a_{tree}.

Fig. 7 shows the schematic graph of generalized dependency between F_{tree} and I_N.

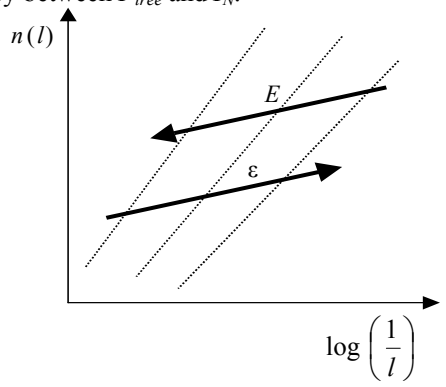

Figure 3 – Schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system

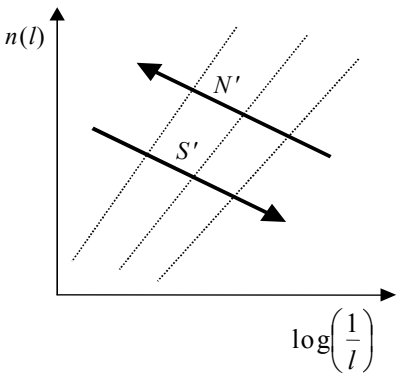

Figure 4 – Schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system

DOI 10.15588/1607-3274-2020-1-11

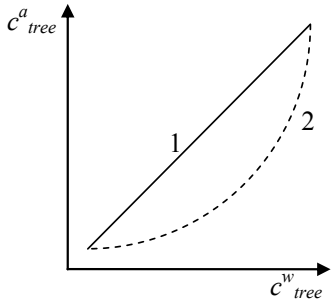

Figure 5 – Schematic graph of generalized dependencies between c^a_{tree} and c^w_{tree}: 1 – $c^a_{tree} = c^w_{tree}$, 2 – $c^a_{tree} < c^w_{tree}$

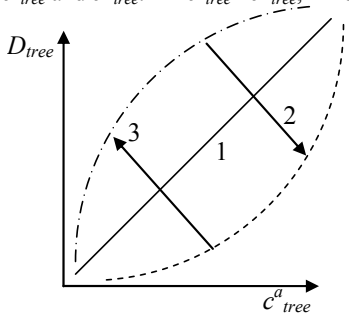

Figure 6 – Schematic graph of generalized dependencies between D_{tree} and c^a_{tree}:
1 – basic relation; 2 – for decreasing of l or $n(l)$ or U or c^w_{tree};
3 – for increasing of l or $n(l)$ or U or c^w_{tree}

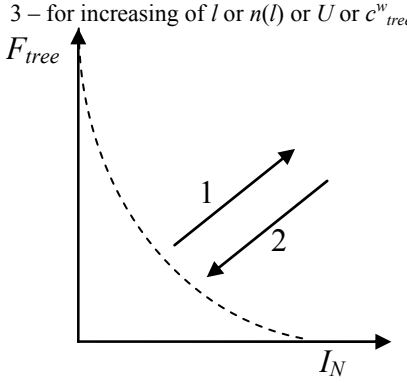

Figure 7 – Schematic graph of generalized dependency between F_{tree} and I_N :
1 – for increasing of model complexity of computations (c^a_{tree} and/or c^w_{tree}), 2 – for decreasing of model complexity of computations (c^a_{tree} and/or c^w_{tree})

Fig. 8 presents the schematic graph of generalized dependencies between F_{tree} and D_{tree}.

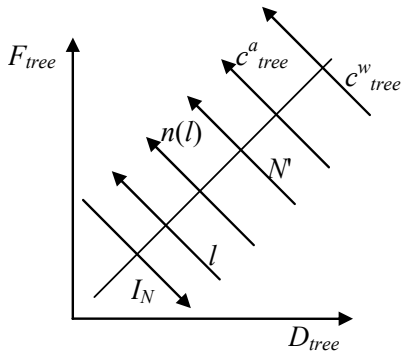

Figure 8 – Schematic graph of generalized dependency between F_{tree} and D_{tree}

6 DISCUSSION

As it can be seen from Fig. 1 and Fig. 2 the proposed indicators of the fractal dimension allow show the differences between classes. These indicators can be used in methods of sample selection, defining quality criteria of formed subsamples on the base of the proposed indicators of the fractal dimension.

If formed subsample or its classes on indicators of the fractal dimension are significantly differ from similar parameters of the original sample, it is possible that, the sample does not have the representativeness relative to the original sample. Also the proposed indicators at the several subsamples-candidates comparing could be used as their quality measures: among subsamples-candidates should be preferred that which have indicators of the fractal dimension with closest values to the original sample indicators values.

As it can be seen from Fig. 3, a change of the specified components of formed subsample dimension (number of features N' and the number of instances S'), also as E and ε affect the position of the straight line connecting points of dependence $n(l)$. The greater the ε value the greater the $n(l)$ and the less the E value the greater the $n(l)$.

From the Fig. 4 we can see that the less S' value and the bigger the N' value the bigger the n(l) value,

As it can bee seen from thr Fig. 5 the bigger the c^{w}_{tree} the bigger the c^{a}_{tree}. The smaller c^{a}_{tree} comparing with c^{w}_{tree} the slower c^{a}_{tree} grow.

Fig. 6 indicates that the bigger c^{a}_{tree} value the bigger the D_{tree} value. If we have decreasing of l or $n(l)$ or U or c^{w}_{tree}; then D_{tree} value will grow slowly comparing with increasing of l or $n(l)$ or U or c^{w}_{tree}

From the Fig. 7 we can bring that the bigger the I_N value the less the F_{tree} value. If model complexity of computations (c^{a}_{tree} and/or c^{w}_{tree}) increased then the bigger F_{tree} and vice versa.

As it can be seen from the Fig. 8 the Ftree indicator will receive the greater value the greater the D_{tree}, l, $n(l)$, N', c^{a}_{tree}, c^{wt}_{ree} and the less the I_N indicator value.

CONCLUSIONS

The urgent problem of decision tree model synthesis using the fractal analysis is considered in the paper.

The scientific novelty of obtained results is that the fractal dimension for a decision tree based model is defined as for whole training sample as for specific classes. The method of the fractal dimension of a model based on a decision tree estimation taking into account model error is proposed. It allows to build model with an acceptable error value, but with optimized level of fractal dimensionality. This makes possibility to reduce decision tree model complexity and to make it mo interpretable.

The set of indicators characterizing complexity of decision tree model is proposed. It contains complexity of node checking, complexity of node achieving, an average model complexity and worst tree model complexity of computations. On the basis of proposed set of indicators a complex criterion for model building is proposed. The indicators of the fractal dimension of the decision tree model error can be used to find and remove the non-informative features in the model.

The practical significance of obtained results is that the developed indicators and methods are implemented in software and studied at practical problem solving. The conducted experiments confirmed the operability of the proposed software and allow recommending it for use in practice for solving the problems of model building by the precedents.

The **prospects for further study** may include the optimization of software implementation of proposed methods and indicators, also as experimental study of proposed indicators on the larger complex of practical problems having different nature and dimension.

ACKNOWLEDGEMENTS

The work was conducted in the framework of the state budget scientific project “Development and research of intelligent methods and software for diagnosing and non-destructive quality control of military and civilian equipment” (State register No. 0119U100360) of National University "Zaporizhzhia Polytechnic" under partial support of international project "Innovative Multidisciplinary Curriculum in Artificial Implants for Bio-Engineering BSc/MSc Degrees” (BIOART, Ref. no. 586114-EPP-1-2017-1-ES-EPPKA2-CBHE-JP)" co-funded by the Erasmus+ Programme of the European Union and "Virtual Master Cooperation on Data Science (ViMaCs) funded by the DAAD.

REFERENCES

1. Geurts P., Irrthum A., Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology, *Molecular Biosystems*, 2009, Vol. 5, No. 12, pp. 1593–1605.
2. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Boca Raton, Chapman and Hall/CRC, 1984, 368 p.
3. Heath D., Kasif S., Salzberg S. Induction of oblique decision trees [Electronic resource]. Baltimor, Johns Hopkins University, 1993, 6 p. Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf
4. Rabcan J., Levashenko V., Zaitseva E., Kvassay M., Subbotin S. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree, *Engineering Structures.* – Vol. 197, 109396.
5. Quinlan J. R. Induction of decision trees, *Machine learning*, 1986, Vol. 1, No. 1, pp. 81– 106.
6. Breiman L. Bagging predictors , *Machine Learning*, 1996, Vol. 24, No. 2, pp. 123–140.
7. Utgoff P. E. Incremental induction of decision trees, *Machine learning*, 1989, Vol. 4, No. 2, pp. 161–186. DOI:10.1023/A:1022699900025
8. Hyafil L., Rivest R. L. Constructing optimal binary decision trees is np-complete, *Information Processing Letters.* – 1976, Vol. 5, No. 1, pp. 15–17.
9. Subbotin S. A. Postroyeniye derev’yev resheniy dlya sluchaya maloinformativnykh priznakov, *Radío Electronics, Computer Science, Control*, 2019, No. 1, pp. 122–131.

DOI 10.15588/1607-3274-2020-1-11

10. Amit Y., Geman D., Wilder K. Joint induction of shape features and tree classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, Vol. 19, No. 11. pp. 1300–1305.
11. Subbotin S. A. Metody sinteza modeley kolichest-vennykh zavisimostey v bazise derev'yev regressii, realizuyushchikh klaster-regressionnuyu approksimatsiyu po pretsedentam, *Radío Electronics, Computer Science, Control,* 2019, No. 3, pp. 76-85.
12. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 300 p.
13. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis, *Applications of Computational Intelligence in Biomedical Technology*. Cham, Springer, 2016, pp. 215–228. DOI: 10.1007/978-3-319-19147-8_13
14. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees, *IEEE Transactions on Computers*, 1989, Vol. 38, № 1, pp. 130–141.
15. Tolosi L., Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions, *Bioinformatics,* 2011, Vol. 27, No. 14, pp. 1986–1994. DOI:10.1093/bioinformatics/btr300
16. Chaudhuri A., Stenger H.. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p. DOI: 10.1201/9781420028638
17. 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
18. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis, *Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings*. Los Alamitos, IEEE, 2014, pp. 332–343. DOI: 10.1109/dt.2014.6868734
19. Lavrakas P.J. Encyclopedia of survey research methods. – Thousand Oaks: Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.4135/9781412963947.n159
20. Łukasik S., Kulczycki P. An algorithm for sample and data dimensionality reduction using fast simulated annealing, *Advanced Data Mining and Applications, Lecture Notes in Computer Science*. Berlin: Springer, 2011, Vol. 7120, pp. 152–161. DOI: 10.1007/978-3-642-25853-4_12
21. Subbotin S. A. The training set quality measures for neural network learning, *Optical Memory and Neural Networks (Information Optics),* 2010, Vol. 19, No. 2, pp. 126–139. DOI: 10.3103/s1060992x10020037
22. Cheng Q. Multifractal Modeling and Lacunarity Analysis, *Mathematical Geology*, 1997, Vol. 29, No. 7, pp. 919–932. DOI:10.1023/A:1022355723781
23. Eftekhari A. Fractal Dimension of Electrochemical Reactions, *Journal of the Electrochemical Society,* 2004, Vol. 151, No. 9, pp. E291–E296. DOI:10.1149/1.1773583.
24. Dubuc B., Quiniou J., Roques-Carmes C., Tricot C., Zucker S. Evaluating the fractal dimension of profiles, *Physical Review,* 1989. – Vol. 39, No. 3. – P. 1500–1512. DOI:10.1103/PhysRevA.39.1500
25. Camastra F. Data Dimensionality Estimation Methods: A survey, *Pattern Recognition*, 2003, Vol. 36, No. 12, pp. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
26. de Sousa P. M., Traina C., Traina A. J. M., Wu L., Faloutsos C. A fast and effective method to find correlations among attributes in databases, *Data Mining and Knowledge Discovery*, 2007, Vol. 14, Issue 3, pp. 367–407. DOI: 10.1007/s10618-006-0056-4
27. Roberts A., Cronin A. Unbiased estimation of multi-fractal dimensions of finite data sets, *Physica A: Statistical Mechanics and its Applications*, 1996, Vol. 233, No. 3–4, pp. 867–878. DOI:10.1016/s0378-4371(96)00165-3
28. Kumaraswamy K. Fractal Dimension for Data Mining [Electronic resource]. Access mode: https://www.ml.cmu.edu/research/dap-papers/skkumar_kdd_project.pdf.
29. Li J. Du Q., Sun C. An improved box-counting method for image fractal dimension estimation, *Pattern Recognition*, 2009, Vol. 42, No. 11, pp. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
30. Popescu D. P., Flueraru C., Mao Y., Chang S., Sowa M.G.,Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue, *Biomedical Optics Express,* 2010, Vol. 1, No. 1, pp. 268–277. DOI:10.1364/boe.1.000268
31. Takens F. On the numerical determination of the dimension of an attractor, *Dynamical Systems and Bifurcations Workshop Groningen, 16–20 April 1984 : proceedings*. Berlin, Springer, 1985, pp. 99–106. DOI: 10.1007/bfb0075637
32. Subbotin S. A. Metriki kachestva vyborok dannykh i modeley zavisimostey, osnovannyye na fraktal'noy razmernosti, *Radío Electronics, Computer Science, Control*, 2017, No. 2, pp. 70–81.
33. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal, *Journal of Theoretical and Applied Information Technology*, 2013, Vol. 49, No. 1, pp. 342–347.
34. Crişan D. A., Dobrescu R. Fractal dimension spectrum as an indicator for training neural networks, *Universitatea Politehnica Bucuresti Sci. Bull. Series C,* 2007, Vol. 69, No. 1, pp. 23–32.
35. Subbotin S. A. The neural network model synthesis based on the fractal analysis, *Optical Memory and Neural Networks,* 2017, Vol. 26, No. 4, pp. 257–273.
36. Fisher Iris dataset [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
37. Dubrovin V., Subbotin S., Morshchavka S., Piza D. The plant recognition on remote sensing results by the feed-forward neural networks, *International Journal of Smart Engineering System Design,* 2001, Vol. 3, No. 4, pp. 251–256.
38. Arrhythmia Data Set [Electronic resource]. Access mode: http://archive.ics.uci.edu/ml/datasets/arrhythmia

Received 20.11.2019.
Accepted 16.02.2020.

УДК 004.93

ФРАКТАЛЬНИЙ АНАЛІЗ ВИБІРОК І МОДЕЛЕЙ НА ОСНОВІ ДЕРЕВ РІШЕНЬ

Субботін С. О. – д-р техн. наук, професор, завідувач кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя, Україна.

Гофман Є. О. – канд. техн. наук, старший науковий співробітник науково-дослідної частини Національного університету «Запорізька політехніка», Запоріжжя, Україна.

АНОТАЦІЯ

Актуальність. У статті розглядається проблема синтезу моделі на основі дерева рішень з використанням фрактального аналізу. Об'єктом дослідження є дерева рішень. Предметом дослідження є методи синтезу та аналізу моделей на основі дерев рішень.

DOI 10.15588/1607-3274-2020-1-11

Мета роботи – створення методів і фрактальних індикаторів, що дозволяють спільно вирішити задачу синтезу моделі на основі дерева рішень і завдання скорочення розмірності навчальних даних за допомогою єдиного підходу, заснованого на принципах фрактального аналізу.

Метод. Фрактальна розмірність для моделі на основі дерева рішень визначена як для всієї навчальної вибірки, так і для кожного класу. Запропоновано метод визначення фрактальної розмірності моделі, заснований на оцінюванні дерева рішень з урахуванням похибки моделі. Це дозволяє побудувати модель з прийнятним значенням помилки, але з оптимізованим рівнем фрактальної розмірності, що дозволяє зменшити складність моделі дерева рішень і зробити її більш зрозумілою. Запропоновано набір показників, що характеризують складність моделі на основі дерева рішень. Він містить складність перевірки вузлів, складність досягнення вузла, середню і найгіршу складність обчислень моделі дерева. На основі запропонованого набору показників запропоновано комплексний критерій побудови моделі. Індикатори фрактальної розмірності помилки моделі дерева рішень можуть бути використані для пошуку і видалення неінформативних ознак в моделі.

Результати. Розроблені показники і методи реалізовані в програмному забезпеченні і вивчені при вирішенні практичних завдань. В результаті експериментального дослідження запропонованих показників отримані графіки залежностей між ними, включаючи графіки залежностей числа гіперблоків, що охоплюють вибірку в просторі ознак, від розміру боку блоку: для всієї вибірки, для кожного класу, для різних встановлених значень помилок і отриманих значень помилок, для різних значень результуючих чисел ознак і екземплярів, також графіків залежностей між середньою і найгіршою складнощами дерева, фрактальної розмірністю дерева рішень і ср днів складністю дерева, об'єднаним критерієм і індикатором скорочення набору ознак, а також між спільним критерієм і фрактальної розмірністю дерева.

Висновки. Проведені експерименти підтвердили працездатність запропонованого математичного забезпечення та дозволяють рекомендувати його для практичного використання для вирішення завдань побудови моделей по прецедентах.

КЛЮЧОВІ СЛОВА: дерево рішень, вибірка, фрактальна розмірність, індикатор, складність дерева.

УДК 004.93

ФРАКТАЛЬНЫЙ АНАЛИЗ ВЫБОРОК И МОДЕЛЕЙ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ

Субботин С. А. – д-р техн. наук, профессор, заведующий кафедрой программных средств Национального университета «Запорожская политехника», Запорожье, Украина.

Гофман Е. А. – канд. техн. наук, старший научный сотрудник научно-исследовательской части Национального университета «Запорожская политехника», Запорожье, Украина.

АННОТАЦИЯ

Актуальность. В статье рассматривается проблема синтеза модели на основе дерева решений с использованием фрактального анализа. Объектом исследования являются деревья решений. Предметом исследования являются методы синтеза и анализа моделей на основе деревьев решений.

Цель работы – создание методов и фрактальных индикаторов, позволяющих совместно решить задачу синтеза модели на основе дерева решений и задачу сокращения размерности обучающих данных с помощью единого подхода, основанного на принципах фрактального анализа.

Метод. Фрактальная размерность для модели на основе дерева решений определена как для всей обучающей выборки, так и для каждого класса. Предложен метод определения фрактальной размерности модели, основанный на оценке дерева решений с учетом погрешности модели. Это позволяет построить модель с приемлемым значением ошибки, но с оптимизированным уровнем фрактальной размерности, что позволяет уменьшить сложность модели дерева решений и сделать ее более понятной. Предложен набор показателей, характеризующих сложность модели на основе дерева решений. Он содержит сложность проверки узлов, сложность достижения узла, среднюю и наихудшую сложность вычислений модели дерева. На основе предложенного набора показателей предложен комплексный критерий построения модели. Индикаторы фрактальной размерности ошибки модели дерева решений могут быть использованы для поиска и удаления неинформативных признаков в модели.

Результаты. Разработанные показатели и методы реализованы в программном обеспечении и изучены при решении практических задач. В результате экспериментального исследования предложенных показателей получены графики зависимостей между ними, включающие графики зависимостей числа гиперблоков, охватывающих выборку в пространстве признаков, от размера стороны блока: для всей выборки, для каждого класса, для различных установленных значений ошибок и полученных значений ошибок, для различных значений результирующих чисел признаков и экземпляров, также графиков зависимостей между средней и наихудшей сложностями дерева, фрактальной размерностью дерева решений и средней сложностью дерева, объединенным критерием и индикатором сокращения набора признаков, а также между совместным критерием и фрактальной размерностью дерева.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для практического использования для решения задач построения моделей по прецедентам.

КЛЮЧЕВЫЕ СЛОВА: дерево решений, выборка, фрактальная размерность, индикатор, сложность дерева.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Geurts P. Supervised learning with decision tree-based methods in computational and systems biology / P. Geurts, A. Irrthum, L. Wehenkel // Molecular Biosystems. – 2009. –Vol. 5, № 12. – P. 1593–1605.
2. Classification and regression trees / [L. Breiman J. H. Friedman, R. A. Olshen, C. J. Stone]. – Boca Raton: Chapman and Hall/CRC, 1984. – 368 p.
3. Heath D. Induction of oblique decision trees [Electronic resource] / D. Heath, S. Kasif, S. Salzberg. – Baltimor : Johns Hopkins University, 1993. – 6 p. – Access mode:

DOI 10.15588/1607-3274-2020-1-11

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf
4. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree / [J. Rabcan, V. Levashenko, E. Zaitseva et al.] // Engineering Structures. – Vol. 197, 109396.
5. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine learning. – 1986. – Vol. 1, № 1. – P. 81–106.
6. Breiman L. Bagging predictors / L. Breiman // Machine Learning. – 1996. – Vol. 24, № 2. – P. 123–140.
7. Utgoff P. E. Incremental induction of decision trees / P. E. Utgoff // Machine learning, 1989. – Vol. 4, № 2. – P. 161–186. DOI:10.1023/A:1022699900025
8. Hyafil L. Constructing optimal binary decision trees is np-complete / L. Hyafil, R. L. Rivest // Information Processing Letters. – 1976. – Vol. 5, № 1. – P.15–17.
9. Субботин С. А. Построение деревьев решений для случая малоинформативных признаков / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2019. – № 1. – С. 122–131.
10. Amit Y. Joint induction of shape features and tree classifiers / Y. Amit, D. Geman, K. Wilder // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19, № 11. – P.1300–1305.
11. Субботин С. А. Методы синтеза моделей количественных зависимостей в базисе деревьев регрессии, реализующих кластер-регрессионную аппроксимацию по прецедентам / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2019. – № 3. – С. 76–85.
12. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. –Hoboken: John Wiley & Sons, 2008. – 300 p.
13. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis / S. Subbotin // Applications of Computational Intelligence in Biomedical Technology. – Cham : Springer, 2016. – P. 215–228. DOI: 10.1007/978-3-319-19147-8_13
14. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees / M. Miyakawa // IEEE Transactions on Computers. – 1989. – Vol. 38, № 1. – P. 130–141.
15. Tolosi L. Classification with correlated features: unreliability of feature ranking and solutions / L. Tolosi, T. Lengauer // Bioinformatics. – 2011. – Vol. 27, № 14. – P. 1986–1994. DOI:10.1093/bioinformatics/btr300
16. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p. DOI: 10.1201/9781420028638
17. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search / S. A. Subbotin // Automatic Control and Computer Sciences. – 2013. – Vol. 47, No. 3. – P. 113–121. DOI: 10.3103/s0146411613030073
18. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis / S. A. Subbotin // Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings. – Los Alamitos: IEEE, 2014. – P. 332–343. DOI: 10.1109/dt.2014.6868734
19. Lavrakas P.J. Encyclopedia of survey research methods. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.4135/9781412963947.n159
20. Łukasik S. An algorithm for sample and data dimensionality reduction using fast simulated annealing / S. Łukasik, P. Kulczycki // Advanced Data Mining and Applications, Lecture Notes in Computer Science. – Berlin : Springer, 2011. – Vol. 7120. – P. 152–161. DOI: 10.1007/978-3-642-25853-4_12
21. Subbotin S.A. The training set quality measures for neural network learning / S. A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19, No. 2. – P. 126–139. DOI: 10.3103/s1060992x10020037
22. Cheng Q. Multifractal Modeling and Lacunarity Analysis / Q. Cheng // Mathematical Geology. – 1997. – Vol. 29, No. 7. – P. 919–932. DOI:10.1023/A:1022355723781
23. Eftekhari A. Fractal Dimension of Electrochemical Reactions / A. Eftekhari // Journal of the Electrochemical Society. – 2004. – Vol. 151, No. 9. – P. E291–E296. DOI:10.1149/1.1773583.
24. Evaluating the fractal dimension of profiles / [B. Dubuc, J. Quiniou, C. Roques-Carmes et al.] // Physical Review. – 1989. – Vol. 39, No. 3. – P. 1500–1512. DOI:10.1103/PhysRevA.39.1500
25. Camastra F. Data Dimensionality Estimation Methods: A survey / F. Camastra // Pattern Recognition. – 2003. – Vol. 36, No. 12. – P. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
26. A fast and effective method to find correlations among attributes in databases / [P. M. de Sousa, C. Traina, A. J. M. Traina et al.] // Data Mining and Knowledge Discovery. – 2007. – Vol. 14, Issue 3. – P. 367–407. DOI: 10.1007/s10618-006-0056-4
27. Roberts A. Unbiased estimation of multi-fractal dimensions of finite data sets / A. Roberts, A. Cronin // Physica A: Statistical Mechanics and its Applications. – 1996. – Vol. 233, No. 3–4. – P. 867–878. DOI:10.1016/s0378-4371(96)00165-3
28. Kumaraswamy K. Fractal Dimension for Data Mining [Electronic resource] / K. Kumaraswamy. – Access mode: https://www.ml.cmu.edu/research/dap-papers/skkumar_kdd_project.pdf.
29. Li J. An improved box-counting method for image fractal dimension estimation / J. Li, Q. Du, C. Sun // Pattern Recognition. – 2009. – Vol. 42, No. 11. – P. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
30. Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue / [D. P. Popescu, C. Flueraru, Y. Mao et al.] // Biomedical Optics Express. – 2010. – Vol. 1, No. 1. – P. 268–277. DOI:10.1364/boe.1.000268
31. Takens F. On the numerical determination of the dimension of an attractor / F. Takens // Dynamical Systems and Bifurcations Workshop Groningen, 16–20 April 1984 : proceedings. – Berlin: Springer, 1985. – P. 99–106. DOI: 10.1007/bfb0075637
32. Субботин С. А. Метрики качества выборок данных и моделей зависимостей, основанные на фрактальной размерности / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2017. – № 2. – С. 70–81.
33. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal / Y. Zong-Chang // Journal of Theoretical and Applied Information Technology. – 2013. – Vol. 49, No. 1. – P. 342–347.
34. Crişan D.A. Fractal dimension spectrum as an indicator for training neural networks / D. A. Crişan, R. Dobrescu, // Universitatea Politehnica Bucuresti Sci. Bull. Series C. – 2007. – Vol. 69, No. 1. – P. 23–32.
35. Subbotin S. A. The neural network model synthesis based on the fractal analysis / S. A. Subbotin // Optical Memory and Neural Networks. – 2017. – Vol. 26, No. 4. – P. 257–273.
36. Fisher Iris dataset [Electronic resource].– Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
37. The plant recognition on remote sensing results by the feed-forward neural networks / [V. Dubrovin, S. Subbotin, S. Morshchavka, D. Piza] // International Journal of Smart Engineering System Design. – 2001. – Vol. 3, No. 4. – P. 251–256.
38. Arrhythmia Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/arrhythmia

DOI 10.15588/1607-3274-2020-1-11