The Regression Tree Model Building Based on a Cluster-Regression Approximation for Data-Driven Medicine

Sergey Subbotin[1][0000-0001-5814-8268], Elena Kirsanova[2][/0000-0001-5337-2439],

[1]Zaporizhzhia National Technical University, Zhukovsky str., 64,Zaporizhzhia, 69063, Ukraine
subbotin@zntu.edu.ua
[2]Zaporizhzhia State Medical University, Maiakovskyi avenue 26, Zaporizhzhia, 69035, Ukraine
kirsanova@zsmu.zp.ua

Abstract. The problem of quantitative dependency model building on precedents for data-driven medicine is considered. A tree-cluster-regression approximation method is proposed. It makes possible to ensure acceptable model accuracy, high levels of interpretability and generalization of data, and to reduce the complexity of the model. The software that implements the proposed methods is developed. The developed software is studied at solving the problem of children health indicator modelling.

Keywords: data-driven diagnosis, regression, cluster analysis, cluster-regression approximation, regression tree, neural network

1 Introduction

The data-driven diagnosis in medicine means decision making based on observations of a set of descriptive diagnostic features characterizing the patient's condition.

In contrast to the traditional expert approach involving human expert, data-driven diagnostics do not require the direct involvement of a person in the decision-making process (this is provided by the model), and also does not require expert knowledge in the form of regularities and rules for building the model (this is ensured by automatic extraction of knowledge from observations in the model learning process).

This allows using the data-driven diagnosis for a quick decision-making under time constraints (screening diagnostics), as well as automating decision-making and control of decisions made in the context of a lack of expert knowledge or experts.

A special class of models for making diagnostic decisions is constituted by quantitative models for estimating or predicting the values of the output real variable.

The known methods for model constructing of quantitative dependencies on observations such as regression analysis [1, 2], neural networks [3, 4] and the Group Method of Data Handling [5, 6] strive to build a single model in the entire feature space by solving an optimization problem that requires a lot of time and leads to a complex model, and also requires a large amount of observations, which is not always possible in practice.

Other well-known methods, such as regression trees [7, 8, 9] and neuro-fuzzy networks [10–13] build a combination of primitive partial models for local areas in the feature space, which allows to simplify the obtained model, but significantly reduces it accuracy, however, at the same time, it allows to synthesize models based on a smaller set of observations.

A compromise between the above groups of methods is the cluster-regression approximation method [14], which allows to obtain sufficiently accurate, as well as simple and interpretable models, minimizing the number of used features.

At the same time, the cluster-regression approximation method [14] is strongly dependent on the cluster analysis procedure, which, as a rule, is time-consuming and requires the setting of cluster separation principles.

The aim of the paper is to simplify the cluster regression approximation models by indirectly implementing the cluster analysis in the process of model building.

2 Formal problem statement

Let we have a training set of S precedents (observations, instances, cases) $<x, y>$, where $x = \{x^s\}$, $y=\{y^s\}$, $s = 1, 2, ..., S$, x^s is an s-th instance of the sample, y^s is an output feature value associated with the s-th instance, $x^s = \{x^s_j\}$, $j = 1, 2, ..., N$, x^s_j is a value of the j-th input feature of the s-th instance, N is a number of input features.

Then the model constructing task for the dependence $y=f(w, x)$ is to find such a model structure f and such values of model parameters w for which the model quality criterion F is satisfied. As a rule, for the problems of approximation the model quality criterion is determined as a function of the model error (1):

$$E = \sum_{s=1}^{S} \left(y^s - f(w, x^s) \right)^2 \to 0 . \quad (1)$$

3 Literature review

The regression analysis methods [1, 2] allow for a given sample of observations to obtain polynomial models, which coefficients are determined in the general case by solving the optimization problem of minimizing the error criterion in the space of model coefficients. The advantage of these methods is their universality. The disadvantages of these methods are the non-interpretativity of the models obtained with a large number of features, as well as the problem of choosing the optimal structure and complexity of the model.

The methods for constructing the predictive models based on neural networks [3, 4], as a rule, also, as in the case of regression analysis, for model building (training) require solving of the optimization problem to minimize the error criterion in the model weights space. However, the model has, as a rule, a specific structure, which can be considered as a hierarchical combination of nonlinear functions and linear polynomials. The method's advantage is its universality. The disadvantages of the

method are non-interpretativity and high complexity of obtained models, the problem of choosing the optimal structure of the model and the parameters of its elements, as well as high interactivity and time-consumity of the method.

The Group Method of Data Handling proposed by A.G. Ivakhnenko [5, 6] assumes enumeration of dependency models based on support functions the polynomials of various degrees, combining different combinations of input variables. For each model, the coefficients are determined by the method of regression analysis [1, 2]. Among all the models the several best are selected. The best selected models are used at the next iteration of the method as arguments for the formation of more complex models. The model quality is determined by the error or the determination coefficient, or by the coefficient of pair correlation between output and input features. If an acceptable model is found, the method terminates. A peculiarity of the method is the convertibility of the obtained models into polynomial neural networks (a kind of deep neural networks). The advantages of the method are its universality, suitability for small-size samples, the ability to solve the problem of informative features selection in the process of model building, the structuredness and hierarchy of the resulting models and, as a result, their interpretability. The disadvantages of the method are a significant increase in the complexity of the obtained models with an increase in the volume of training data and requirements for the accuracy of the model, as well as an increase in the number of used input features, high interactivity and considerable time-consumity of the method.

A common feature of all the methods listed above is that they strive to build a single polynomial model throughout the entire space of the original features, the coefficients of which are selected by solving an optimization problem. This approach turns out to be very costly with a large number of features and instances, and is also fraught with the problem of choosing the starting point of the search, setting the optimal structure and parameters of the network.

The methods of regression tree constructing [7–9] hierarchically divide the initial space into regions (clusters), in which they evaluate the output average value for the instances hit in the cluster. This value is assigned to the output feature of all instances hitting in this cluster. The advantage of this group of methods is the simplicity and interpretability of the resulting models, as well as the possibility of passing the cluster analysis tasks and selecting informative features. The disadvantages of the methods of this group are the low accuracy of the obtained models, as well as the ambiguity in the hierarchical combination of checks for assigning an instance to a cluster.

The predictive model constructing methods based on neuro-fuzzy networks [10–13] represent a model of the dependence as a combination of linear regression models of dependencies of an output feature from input features for individual regions in the feature space (clusters) defined by rules, specified by experts or derived from a cluster analysis. The advantage of the resulting models is their interpretability. The disadvantages of these methods are that the resulting models are not accurate enough, when there are a large number of initial features then model is an extremely complex, and the time to build a model also significantly increases. In addition, the disadvantages include the dependability of this group of methods on the availability of expert knowledge or the need for a preliminary cluster analysis.

A common peculiarity of the decision trees and neuro-fuzzy networks in numerical model construction is that they instead of constructing a single model for entire feature space build a set of hierarchically derived linear models as a result of cluster analysis. The advantage of the obtained models is their interpretability. The disadvantage of the obtained models is their low accuracy.

The cluster-regression approximation method proposed in [14] combines the advantages and eliminates the disadvantages of the above described groups of methods. The clusters are allocated in the original feature space. Then for each cluster the method builds partial model. At the same time, in each cluster a set of models is built that use a smaller in size set of the most informative features to build a partial model, while the partial models are built as linear or neural network based. This method also has a neural network and neuro-fuzzy interpretation [14].

The advantage of the cluster-regression approximation method is that it is allow to obtain more accurate models than neuro-fuzzy networks and decision trees, but, unlike neural networks, it builds separate models for each cluster, trying to minimize the number of features for each partial model, which makes possible to obtain simpler and more interpretable models. The disadvantage of the cluster-regression approximation method [14] is that it requires to use a cluster analysis [15, 16], during which it uses the whole set of initial features.

Therefore, an urgent task is to improve the method of cluster-regression approximation by indirectly implementing cluster analysis in the process of model building.

4 The modified method of a cluster-regression approximation

Consider the basic method of cluster regression approximation [14]. It splits the sample in a feature space into compact groups of instances called clusters. For instances of each cluster, it builds partial regression models, seeking to minimize the complexity of each model and the number of used features. However, the resulting model may still be redundant. Therefore, it is proposed to simplify the model after building by contrasting the model weights. Formally, this method can be written as follows.

1. The cluster allocation stage. On the basis of instances of a given training sample $<x, y>$ using a given cluster analysis method [16, 17], allocate Q clusters in the feature space by determining the coordinates of their centers $\{C^q\}$, $C^q = \{C^q_j\}$, where C^q_j is a coordinate of a q-th cluster center on j-th feature, $q = 1, 2, ..., Q$, $j = 1, 2, ..., N$.

2. The stage of a training set cluster analysis. Determine the belonging of each s-th instance x^s of the sample $<x, y>$ to the clusters, $s = 1, 2, ..., S$:

– find the distances $R(x^s, C^q)$ from the instance x^s to the center of each cluster C^q, $q = 1, 2, ..., Q$, in the metric of the corresponding cluster analysis method;

– assign the instance x^s to the q^s cluster, the distance from the instance to which center is the smallest, where q^s is obtained form (2):

$$q^s = \arg \min_{q=1,2,...,Q} \{R(x^s, C^q)\}. \quad (2)$$

3. The stage of partial model building for clusters. For instances of each q-th cluster, $q = 1, 2, ..., Q$:

– if only one instance hit into q-th cluster, then accept it as a singleton, specifying the partial model-function of the q-th cluster as a constant: $f^q = \{x^s \mid \exists! s : q^s = q\}$;

– if more than one instance hit into q-th cluster, then evaluate the individual informativity of each j-th feature relatively to the output feature y for instances of the q-th cluster $I^q(x_j, y)$ [17, 18], build for the instances of q-th cluster the one-dimensional linear regression model of the output feature y dependence on the individually most informative input feature $f^q = w^q{}_1 x^s{}_j + w^q{}_0$, where $w^q{}_j$ is a q-th cluster model weight coefficient for j-th feature obtained using the least squares method [1, 2], estimate the error of the resulting partial q-th model E^q using (1), if the error E^q is acceptable (for example, if $E^q(S^q/S) \le \varepsilon$, where ε is the user-specified maximal allowable error value for the entire sample), then proceed the next cluster, otherwise: build for the q-th cluster instances a multidimensional linear regression model of the output feature dependence on the entire set of input features in the form (3) [1, 2]:

$$f^q = w_0^q + \sum_{j=1}^{N} w_j^q x_j^s , \tag{3}$$

estimate the error of the obtained q-th partial model E^q, if the error E^q is acceptable (for example, if $E^q(S^q/S) \le \varepsilon$), then proceed the next cluster, otherwise: build a multidimensional non-linear model of the dependence of the output feature from the entire set of input features for instances of q-th cluster based on a single-layer perceptron working according to (4) [19]:

$$f^q = \psi^q \left(w_0^q + \sum_{j=1}^{N} w_j^q x_j^s \right), \tag{4}$$

where ψ^q is a nonlinear activation function, for example, sigmoid (in this case may be we will need to normalize the output parameter, mapping its values to the interval of ψ^q function values), using the Widrow-Hoff method [19] for training, estimate the error of the resulting partial q-th model E^q, if the error E^q is acceptable (for example, if $E^q(S^q/S) \le \varepsilon$), then proceed the next cluster, otherwise: build for the q-th cluster instances a partial multidimensional nonlinear model of the output feature y dependence on the entire set of input features based on two-layer perceptron (5) [3, 4]:

$$f^q = \psi^{q,(2,1)} \left(w_0^{q,(2,1)} + \sum_{i=1}^{N_{q,1}} w_i^{q,(2,1)} \psi_i^{q,(2,1)} \left(w_0^{q,(1,i)} + \sum_{j=1}^{N} w_j^{q,(1,i)} x_j^s \right) \right), \tag{5}$$

where $\psi^{q,(\eta,i)}$ is the activation function of i-th neuron of η-th layer of the neural network of q-th model, $w_j^{q,(\eta,i)}$ is the weight coefficient of j-th input of i-th neuron of the η-th layer of the neural network of q-th model, $N_{q,1}$ is the number of neurons in the first layer of neural network of q-th model. The model is trained on the basis of the Levenberg-Marquardt method [20] using the error back-propagation technique [21]. Estimate the error obtained by the partial q-th model E^q. If the error E^q is acceptable (for example, if $E^q(S^q/S) \le \varepsilon$), then proceed to the next cluster, otherwise, either continue by analogy to increase the number of layers in the partial model, or accept as a partial model of the q-th cluster the most accurate of the constructed partial models.

4. The model simplification stage. For all clusters for which multidimensional regression models are constructed, sequentially iterate over combinations of features, removing k of the least individually informative features (k = 1, 2, ..., N–1), setting their weights in the partial model of the q-th cluster equal zero, as long as the error of the corresponding partial model remains acceptable. For all clusters for which multidimensional neural network models are constructed, perform contrasting of weights [22], as long as the errors of partial neural network models are acceptable. If some of the original features are excluded from all partial models, then remove them from the sample.

5. The model synthesis stage. Based on the obtained set of cluster centers and constructed partial models, synthesize a cluster-regression, neural network or neuro-fuzzy model [14], or a regression tree of a special type as described below.

5 The method of regression tree synthesis based on a cluster-regression approximation

The method of regression tree synthesis based on the built cluster-regression approximation assumes that each leaf of the tree (the node that has no descendants) is considered a cluster. At the same time, in contrast to the known methods of regression tree building, where the each leaf contains only the average value of the output feature for the instances that hit into this leaf (cluster), the proposed method in each leaf contains the function of the partial model. The proposed method consists of the following steps.

1. The pseudo-sampling stage. Firstly form pseudo-instances based on the centers of selected clusters $\{C^q\}$, taking the coordinates of the cluster centers $C^q=\{C^q_j\}$ as the coordinates of the input features of pseudo-instances, and as the output taking the average output of the partial model of the corresponding cluster (6):

$$\bar{y}^q = \frac{1}{S^q} \sum_{s=1}^{S} \{y^s \mid q^s = q\}, \tag{6}$$

where S^q is a number of instances hit into q-th cluster.

2. Stage of the decision tree building. For the sample of pseudo-instances $<C, \bar{y}>$, $C = \{C^q\}$, $\bar{y} = \{\bar{y}^q\}$, construct regression trees based on the known methods of regression tree constructing [7-9]. Then select best model from them in the sense of accuracy. This will be the tree providing the smallest error.

3. Stage of partial model construction. Regarding the best constructed tree, recognize instances of the training set, determining their hit in the leafy nodes of the tree. For the instances of each corresponding q-th cluster leaf obtain the partial model f^q performing the third stage of the cluster-regression approximation method described above.

4. The stage of the regression tree reduction. Looking through the resulting tree from the bottom to up (from the leafs to the root):

– if the current q-th node is a leaf, then skip it by going to the next node;

– if the current q-th node is not a leaf and one of its descendants is also not a leaf, then skip it by going to the next node;

– if the current q-th node is not a leaf, and all its descendants are leafs, then for instances of nodes-descendants of this node build a partial regression model f^q regarding the third stage of the above-described method of cluster-regression approximation and evaluate model error E^q for instances of these nodes. If the model f^q provides an acceptable error ($E^q(S^q/S) \le \varepsilon$), then add it to the current q-th node, and remove all children of the q-th node and then use the q-th node as a leaf.

The constructed tree can be used as an independent model, as well as for constructing the neural network model based on a cluster-regression approximation.

6 The method of neural network model constructing based on a regression tree

The method of a neural network model building based on a regression tree can be presented as follows.

1. The sample clustering stage. Recognize the given training sample $<x, y>$ using the constructed regression tree and determine belonging q^s of each instance x^s, $s = 1, 2, .., S$, to leaf nodes (clusters) of the tree.

2. The cluster centers determining stage. For instances of each q-th cluster (leaf of the tree) find the coordinates of its center $C^q = \{C^q_j\}$ using (7):

$$C_j^q = \frac{1}{S^q} \sum_{s=1}^{S} \{x_j^s \mid q^s = q\}, \; j = 1, 2, ..., N, \; q = 1,2, ..., Q. \tag{7}$$

where S^q is a number of instances hit to the q-th cluster (tree leaf).

3. The model construction stage. On the basis of the found coordinates of the cluster centers $\{C^q\}$ and the partial models of the nodes-clusters $\{f^q\}$, construct a cluster-regression model in the form of a neural or neuro-fuzzy network [14].

The proposed set of methods makes possible to synthesize a cluster-regression model on the basis of a given sample, on the basis of which a regression tree can be constructed, which can also be transformed into a neural network model.

7 Model selection and comparison criteria

The choice of a model for a specific task is determined by the presence of properties important to the user. To compare the models, we need to consider their most important properties and to define the indicators that quantitatively characterize them.

The key indicator of model properties is its error E. Also, the model is characterized by the number of adjustable parameters N_w, and by the number of used features N'.

The model quality information criteria known from the literature [23–25] such as the Hannan-Quinn Criterion, Bayesian Information Criterion, Minimum Description Length, Shortest Data Description, Akaike's Information Criterion (AIC), Corrected AIC, Unbiased AIC, Consistent AIC, and Mallow Criterion depend on the model error, the training sample dimensionality (the number of instances and the number of features), the number of the model adjustable parameters and the maximum permissible number of model parameters [26, 27].

Since the models comparing, as a rule, is assumed that training samples are identical, it is reasonable to exclude the sample size from the comparison criteria. Since regression trees and neural networks are graphs, when determining the number of configurable parameters of such models, it seems appropriate to take into account the fact that link is considered to be absent if its weight is zero. Thus, it is needed to exclude weights equal to zero from the number of configured model parameters.

At the same time, various synthesized models can use not all of the features presented in the sample. Therefore, the number of features used in the model should be considered as an important peculiarity when comparing them.

Based on these considerations, we define the integral information criterion (8):

$$I_{IC} = \left(1 - \frac{N'}{N}\frac{\left(N_w - N_{w=0}\right)}{N_w^{\max}}\right)\mathrm{e}^{-\bar{E}/\bar{E}_{\max}}, \tag{8}$$

where $N>0$, $N' \le N$, $N_w^{\max} > 0$, $N_w \le N_w^{\max}$, $N_{w=0} \le N_w$, $N_{w=0}$ is a number of adjustable model parameters equal to zero, $N_w^{\max}$ is a maximum possible number of adjustable model parameters in the set of compared models, $\bar{E}_{\max}$ is a maximal model error in the set of compared models ($\bar{E}_{\max}>0$), $\bar{E}$ is an average model error per instance ($\bar{E} \ge 0$) calculated by the formula (9):

$$\bar{E} = \frac{1}{S}\sum_{s=1}^{S}\frac{\left|y^s - f(w, x^s)\right|}{\left|\max\limits_{p=1,\ldots,S}(y^p) - \min\limits_{p=1,\ldots,S}(y^p)\right|}. \tag{9}$$

The I_{IC} criterion will take values in the range from zero to one. The smaller will be its value, the worse will be the model, and the more will be it value, the better will be the model.

8 Experiments and results

The proposed methods and indicators characterizing their quality were implemented as software and experimentally investigated in solving the problem of modeling the indicator of children's health.

In the unfavorable environmental situation in large industrial centers, it is relevant to study the effect of environmental pollution on the health of the population and, above all, children, since they are more susceptible to the adverse effects of environmental factors compared with adults. Since Zaporizhzhia is one of the most anthropogenically polluted large industrial cities of Ukraine, it is possible on its example to study the influence of various factors on the state of children's health.

The initial sample was collected as a set of instances each of which was a set of values of diagnostic features characterized environmental, medical-genetic, and social factors for the corresponding child [28]. The list of input features (commonly related for ecological and social conditions) is shown in the Table 1. The fragment of collected data sample is presented in the Table 2.

The index of regulatory mechanisms characterizing the degree of centralization of heart rhythm control was used as an indicator of children's health (model output *y*). It was computed based on features characterizing medical state of patients (electrocardiographic data) using the formula (10):

$$y = \frac{AMo}{2VAR \cdot Mo}, \tag{10}$$

where *AMo* (%) is a mode of amplitude – the proportion of R-R intervals that corresponds to the mode value (it reflects the stabilizing effect of the centralization of cardiac rhythm control, which is mainly due to the influence of the sympathetic part of the vegetative nervous system), *Mo* (ms) is a mode that characterizes the values of the duration of R-R intervals that are most typical and correspond to the most probable the level of functioning of the sinusoidal node of the blood circulation system at the moment (the physiological value of the index is to determine the activity of the humoral channel of regulation of the cardiac rhythm), *VAR* (ms) is a variation scale that characterizes the difference between the largest and the smallest duration of R-R intervals (it reflects the overall effect of regulating the rhythm of the vegetative nervous system and indicates the maximum amplitude of fluctuations of R-R intervals and also is substantially related with the state of the parasympathetic part of the vegetative nervous system, although under certain conditions, with a significant amplitude of slow waves, more depends on the state of subcortical nerve centers than on the tone of the parasympathetic vegetative nervous system).

Table 1. The input features for a children health indicator modeling

Feature	Description of a children feature
x_1	a city area code (on pollution increasing)
x_2	a child age (years)
x_3	a child gender (1 - male, 2 - female)
x_4	a child height (cm)
x_5	did the child attend preschool? (1 - yes, 2 - no)
x_6	an age at what the child attend preschool
x_7	an average duration of school lessons
x_8	a average duration of homework preparation
x_9	a time that child spend daily on the street
x_{10}	a time that child watch TV daily
x_{11}	a duration of a child's nightly sleep
x_{12}	how many times a child eat meat products per day
x_{13}	how many times a child eat fish products per day
x_{14}	how many times a child eat dairy products per day
x_{15}	how many times a child eat vegetables and fruits per day
x_{16}	does the child play sports (1 - yes, 2 - no)
x_{17}	a mother's age at child birth (years)
x_{18}	a mother's education at the time of child birth
x_{19}	professional hazards of mother's work before the child was born (1 - yes, 2 - no)
x_{20}	a family type (1 - full, 2 - incomplete)
x_{21}	a number of family members (1 - 2, 2 - 3, 3 - 4, 4 - 5, 5 - 6 and more)
x_{22}	a number of children in a family
x_{23}	a pregnancy course (1 - without complications, 2 - with complications)
x_{24}	did the mother breastfeed (1 - no, 2 - up to 4 months, 3 - more than 4 months)
x_{25}	does the mother suffer from chronic diseases (1 - yes, 2 - no)
x_{26}	professional hazards of father's work before the child was born (1 - yes, 2 - no)
x_{27}	whether the father suffers from chronic diseases (1 - yes, 2 - no)
x_{28}	whether the mother smoked during the period of pregnancy (1 - yes, 2 - no)
x_{29}	a frequency of alcohol consumption by the father
x_{30}	a frequency of alcohol consumption by the mother
x_{31}	whether the father smokes (1 - yes, 2 - no)
x_{32}	whether the mother smokes (1 - yes, 2 - no)
x_{33}	a type of housing
x_{34}	a number of people living in a dwelling
x_{35}	an average income per 1 family member per month

Since in this case it is important not only to build a model of quantitative dependence, but also to ensure its subsequent use for analyzing the dependency and the significance of each feature, the most rational approach is to use the cluster-regression approximation method, which allows to synthesize logically transparent structured neural models.

On the basis of the training sample [28] the children's health indicator dependency models were constructed using various methods. The results of the experiments are presented in the Table 3.

The results of the conducted experiments have confirmed the operability and practical applicability of the developed methods and the software.

Table 2. The fragment of a training sample for a children health indicator modeling

Feature	Instance (*s*)							
	1	2	…	310	311	…	953	954
x_1	1	1	…	2	2	…	3	3
x_2	7	7	…	7	7	…	10	10
x_3	1	1	…	1	1	…	2	2
x_4	7	3	…	8	7	…	10	10
x_5	1	1	…	1	1	…	1	1
x_6	3	2	…	4	3	…	3	3
x_7	5	3	…	2	2	…	4	3
x_8	4	1	…	3	3	…	4	2
x_9	2	1	…	2	1	…	0	1
x_{10}	2	3	…	2	3	…	2	3
x_{11}	3	3	…	3	3	…	2	3
x_{12}	2	3	…	1	2	…	1	3
x_{13}	3	3	…	3	3	…	2	3
x_{14}	1	2	…	1	2	…	2	4
x_{15}	2	2	…	1	2	…	1	2
x_{16}	1	1	…	2	2	…	1	1
x_{17}	27	27	…	17	18	…	18	28
x_{18}	3	3	…	4	3	…	2	1
x_{19}	2	2	…	2	2	…	2	2
x_{20}	1	2	…	1	2	…	1	2
x_{21}	3	1	…	4	1	…	3	3
x_{22}	2	1	…	3	1	…	1	2
x_{23}	2	2	…	1	1	…	2	1
x_{24}	2	2	…	2	2	…	1	2
x_{25}	1	1	…	2	2	…	2	2
x_{26}	1	2	…	2	2	…	2	2
x_{27}	2	1	…	1	2	…	2	2
x_{28}	2	2	…	2	2	…	2	2
x_{29}	2	3	…	2	1	…	1	3
x_{30}	1	2	…	2	1	…	3	2
x_{31}	3	4	…	1	3	…	1	4
x_{32}	2	1	…	2	1	…	1	2
x_{33}	2	2	…	2	3	…	2	2
x_{34}	5	3	…	5	3	…	4	4
x_{35}	120	120	…	300	200	…	50	100
y	1073	430	…	178	83	…	331	106

The results of the experiments presented in the Table 3 show that the proposed cluster regression approximation method allows to obtain models that are comparable in accuracy with models constructed by other methods. Also, the proposed indicator allows to compare regression, neural network and cluster regression models, as well as models based on regression trees.

The highest level of error has linear regression and two-layer feed-forward neural network with a small number of nodes. This can be explained by that such models have lack of adjusted parameters to extract knowledge from the given data sample.

Table 3. Results of experiments

Model type and training method	N'	N_w	$N_{w=0}$	$\overline{E}$	I_{IC}
Multidimensional linear regression (first order polynomial), Levenberg-Marquardt method [20]	35	36	0	0.0634	0.360233
Multidimensional linear regression (second order polynomial), Levenberg-Marquardt method [20]	35	1296	9	0.0391	0.138667
Regression tree, Breiman method [8]	34	1732	0	0.0358	0.016244
Single-layer perceptron, Widrow-Hoff method [9]	35	36	0	0.0630	0.362513
Two-layer feed-forward neural network with 10 neurons in a hidden layer, Levenberg-Marquardt method [20]	35	371	0	0.0634	0.289078
Two-layer feed-forward neural network with 20 neurons in a hidden layer, Levenberg-Marquardt method [20]	35	741	1	0.0122	0.47249
Two-layer feed-forward neural network with 30 neurons in a in hidden layer, method [20]	35	1111	3	0.0094	0.310632
Cluster-regression model [14]	33	825	58	0.0131	0.473733
Regression tree based on a cluster-regression (proposed in the paper)	33	825	116	0.0129	0.500991

The lowest level of error has feed-forward neural networks with 20 and 30 neurons in a hidden layer. This can be explained by that such models have enough number of adjusted parameters to extract knowledge from the given data sample.

The lowest level of information criterion has a regression tree constructed by the Breiman's method [8]. This can be explained by that a classic regression tree is a very rough method, which replaces regression with classification.

The highest level of information criterion has a regression tree based on a cluster-regression approximation. This can be explained by that final model contains a set of partial regression models for clusters, so it is more accurate in comparison with classical Breiman regression trees and has abilities similar to neural network models. But in comparison with feed-forward networks the proposed cluster-regression approximation model is slightly less accurate. This may be explained by that the proposed method seeks to reduce the number of features and simplify partial models, as well as to make their weights more contrast. This inevitably leads to the loss of information. However, a slight loss of accuracy leads to a simpler and more convenient model for subsequent analysis.

In contrast to the traditional methods of regression model constructing [1, 2], which build a model based on a function of a single form for the entire feature space, the proposed method forms a hierarchical combination of partial models.

In contrast to the known methods of regression tree constructing [7, 8], the leaf nodes of which contain average values of the output feature for clusters, the proposed method forms a tree consisting of partial models for clusters, which allows to provide the greater accuracy of the model.

In contrast to traditional methods of neural network model building based on feed-forward layered networks [3, 4], which build a single model for the entire feature space, the proposed method forms a hierarchical combination of partial models.

This allows to recommend the proposed methods and the software implementing them for use in practice for solving the model building problems of quantitative dependencies on precedents.

The quality indicators of models presented in this paper do not take into account the properties of training samples [26-27, 29-31]. Therefore, in further studies, it seems appropriate to study the effect of sample properties on model quality indicators, as well as to develop model quality indicators, universal for a wide class of computational intelligence models, taking into account the properties of training samples. Also in the process of constructing a cluster regression approximation, it seems appropriate in further studies to study the choice of the number of clusters and the cluster analysis method taking into account the quality indicators of training samples.

9 Conclusion

The problem of quantitative dependence model building based on precedents is addressed in the paper.

A tree-cluster-regression approximation method is proposed. It for a given training sample builds a tree for hierarchical clustering of instances, the leaf nodes of which correspond to clusters, for each cluster the method builds a partial model of dependency on learning sample instances, that fell into the cluster, trying to provide the smallest complexity of the model and uses a set of the most informative features of the smallest length. This allows to ensure acceptable accuracy of the model, high levels of interpretability and data generalization, to reduce the complexity of the model, and to simplify its implementation.

The indicator allowing to quantitatively characterize the quality for models of different types (neural and neuro-fuzzy networks, regression models, regression trees and cluster-regression models) has been proposed The proposed indicator allows to compare different models of dependencies, as well as to form on its basis the criteria for the model quality that can be used for training and simplification, as well as for the model selection.

The software implementing the proposed methods has been developed and studied at the children health index modeling problem solving. The conducted experiments have confirmed the performance of the developed software and allow to recommend it for use in practice.

The prospects for further research are to test the proposed methods on a wider set of applied problems, to study the dependence of the speed and accuracy (error) of methods work on the sample volume and the feature number in the original sample.

References

1. Newbold, P.: Statistics for business and economics. Prentice-Hall, New York (2007)
2. Afifi, A. A., Azen, S. P.: Statistical Analysis: A Computer Oriented Approach. Academic Press, London (1979)
3. Kruse, R., Borgelt, C., Klawonn, F. et. al.: Computational intelligence: a methodological introduction. Springer-Verlag, London (2013)
4. Ruan, D. (ed.): Intelligent hybrid systems: fuzzy logic, neural networks, and genetic algorithms. Springer, Berlin (2012)
5. Ivakhnenko, A.G., Müller, J.A.: Parametric and nonparametric selection procedures. Experimental Systems Analysis. Systems Analysis, Modelling and Simulation (SAMS), 1992, vol.9, pp. 157-175 (1992)
6. Madala, H. R., Ivakhnenko, A.G.: Inductive learning algorithms for complex systems modeling. CRC Press, Boca Raton (1994)
7. Clarke, B., Fokoue, E., Zhang, H. H.: Principles and theory for data mining and machine learning. Springer, New York (2009)
8. Breiman, L., Friedman, J. H., Stone, C. J., Olshen, R. A.: Classification and regression trees. Chapman & Hall / CRC, Boca Raton (1984)
9. Rabcan, J., Rusnak, P., Subbotin, S.: Classification by fuzzy decision trees inducted based on Cumulative Mutual Information. In: 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2018 - Proceedings , Slavske, 20-24 February 2018, pp. 208-212 (2018)
10. Rutkowski, L.: Flexible neuro-fuzzy systems : structures, learning and performance evaluation. Kluwer, Boston (2004)
11. Liu, P., Li, H.: Fuzzy neural network theory and application. Series in Machine Perception and Artificial Intelligence ; vol. 59. World Scientific, Singapore (2004)
12. Buckleya, J. J., Hayashi, Y.: Fuzzy neural networks: a survey. Fuzzy sets and systems. 66(1): 1–13 (1994)
13. Jang, J.R., Sun, C.-T., Mizutani, E.: Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Upple Saddle River (1997)
14. Subbotin, S.: Algorithms of cluster-regression approximation and their neural network interpretations. Radio Electronics, Computer Science, Control 1: 114-121 (2003)
15. Berkhin P., Dhillon, I. S.: Knowledge discovery: clustering. Encyclopedia of complexity and systems science. Springer, p. 5051-5064 (2009)
16. Abonyi, J., Feil, B.: Cluster analysis for data mining and system identification. Birkhäuser, Basel (2007)
17. Subbotin, S.: The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis. Studies in Computational Intelligence, vol. 606, pp. 215-228 (2015)
18. Subbotin, S.: Neural network modeling of medications impact on the pressure of a patient with arterial hypertension. In: IDT 2016 - Proceedings of the International Conference on Information and Digital Technologies 2016, 5-7 July 2016, pp. 249-260 (2016)
19. Widrow, B., Lehr, M. A.: 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. Proceedings of the IEEE. 78(9):1415–1442. (1990) doi:10.1109/5.58323
20. Ravindran, A., Ragsdell, K. M., Reklaitis, G. V.: Engineering optimization: methods and applications. John Wiley & Sons, New Jersey (2006)
21. Rumelhart, D. E., Hinton, G. E., Williams, R. J.: Learning representations by back-propagating errors. Nature, vol 323, pp. 533–536. (1986) doi:10.1038/323533a0

22. Gorban, A. N. Mirkes, Eu. M., Tsaregorodtsev, V. G.: Generation of explicit knowledge from empirical data through pruning of trainable neural networks. Proceedings of International Joint Conference on Neural Networks (IJCNN'99), Washington, July 1999. IEEE, Los Alamitos, vol. 6, p. 4393-4398 (1999)
23. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19(6): 716–723 (1974)
24. Schwarz, G. E.: Estimating the dimension of a model. Annals of Statistics. vol. 6 (2): 461–464 (1978)
25. Hannan, E. J., Quinn, B. G.: The determination of the order of an autoregression. Journal of the Royal Statistical Society, serie B, 41, pp. 190–195 (1979)
26. Subbotin, S.A.: The neural network model synthesis based on the fractal analysis Optical Memory and Neural Networks (Information Optics) 26: 257-273 (2017) https://doi.org/10.3103/S1060992X17040099
27. Subbotin, S.: Methods of data sample metrics evaluation based on fractal dimension for computational intelligence model buiding. 4th International Scientific-Practical Conference Problems of Infocommunications Science and Technology, PICS and T 2017 - Proceedings, 10-13 Oct. 2017, pp. 1-6 (2018)
28. Subbotin, S. A., Kirsanova, E. V.: Synthesis of a multi-layer neural network based on cluster-regression approximation in the task of modeling the children's health indicator. In: XII All-Russian Workshop on neuroinformatics and its application - proceedings, 1-3 October 2004, Krasnoyarsk: ICM SB RAS, 2004, pp. 136–137 (2004)
29. Subbotin, S.A.: Methods of sampling based on exhaustive and evolutionary search Automatic Control and Computer Sciences 47(3): 113-121 (2013)
30. Subbotin, S.A.: The sample properties evaluation for pattern recognition and intelligent diagnosis In: DT 2014 - 10th International Conference on Digital Technologies 2014, Zilina, 9-11 July 2014, pp. 321-332 (2014) doi: 10.1109/DT.2014.6868734
31. Subbotin, S.A.: The training set quality measures for neural network learning Optical Memory and Neural Networks (Information Optics) 19(2):. 126-139 (2010)