METHOD OF DATA ANALYSIS BASED ON CLUSTERING IN "SYNDROMES" INDICATORS SPACE

Senko Oleg, Kuznetsova Anna, Kostomarova Irina

***Abstract**: A new data analysis method is discussed that is based on calculating syndromes by training data sets. Syndrome are defined as sub-regions in feature space where mean values of target Y deviates from mean value of Y in whole data set. Described method of syndromes construction uses boundaries found with the help of modified version of optimal valid partitioning (OVP) method. The modification is based on new validation technique that allows more effectively delete redundant regularities from output set. OVP boundaries are used to find sub-regions in features space with strong deviation of target Y from its mean by whole data set. Such sub-regions further are called syndromes. Hierarchical tree method was applied to receive clusters of objects from training dataset in space of binary indices indicating if feature description of object belongs to corresponding syndrome. Such technique allows discovering sets of objects with similar syndromes. Experiments with biomedical datasets are discussed.*

***Keywords**: Optimal partitioning, statistical validity, permutation test, regularities, gerontology.*

***ACM Classification Keywords**: H.2.8 Database Applications - Data mining, G.3 Probability and Statistics - Nonparametric statistics, Probabilistic algorithms*

Introduction

Many data analysis, forecasting or recognition methods are based on searching such sub-regions in space of explanatory variables (features) $X_1,\ldots,X_n$ where levels of target variable Y deviate significantly from Y mean in data set or at least in neighboring sub-regions. Such sub-regions may be associated for example with leaves in regression trees [Breiman, 1984]. Leaves in classification trees correspond to sub-regions of feature space that contain object mainly from one of classes. Approach that is based on logical regularities must be mentioned thereupon [V.V.Ryazanov, 2003]. Logical regularities are defined as conjunctions of predicates characterizing single features. Conjunctions must be true for possibly maximal subset of one of classes in training set and must be false for objects from other classes. Logical regularity describes hyper-boxes in feature space that contain objects descriptions. At that each hyper-box contains object only from one class. Special optimization techniques allow efficiently search logical regularities. Optimal valid partitioning (OVP) ([Sen'ko, 2006; Senko,2010]) is another method that is aimed to find in feature space boundaries separating objects with different levels of target OVP implements also evaluating statistical validity of empirical regularities described by found optimal partitions with the help of permutation test. Permutation test now become popular toll to asses statistical validity ([Ernst, 2004; Gorman, 2001]). Result of OVP application in some data analysis task is set of statistically valid regularities. At section 2 new modification of OVP technique is discussed that allows eliminating from output regularities system all irredundant 2-dimensional regularities. Previous variant of OVP method allows elimination of irredundant 2-dimensional regularity R only when simple one-dimensional regularities exist for variables

relevant to R. This set of regularities may be further analyzed by experts and used in forecasting algorithms. OVP technique was used in set of biomedical tasks ([Kuznetsova, 2000; Kuznetsova, 2011; Kuznetsova, 2013]).

In this paper a new additional techniques are discussed that allow receive additional useful knowledge from system of empirical regularities that were previously found with the help of OVP technique. Developed method may be used in tasks with binary target variable. At the first step boundaries of OVP regularities are used to find sub-regions in features space with strong deviation of target Y from its mean value by whole data set. Such sub-regions further are called syndromes. It must be noted that dimension of searched syndromes may be higher than 2. Represented version allows finding syndromes of dimension 3. Second stage is aimed to discover groups of objects in training set with X-descriptions belonging to the same or to the similar syndromes. Descriptions of studied objects are generated that will be further referred to as Z-descriptions. Z-description consists of set of binary indices that indicate if feature description belongs to corresponding syndrome. The discussed method is based on hierarchical cluster analysis in Z-space. Result of cluster analysis is several groups of objects with X-descriptions belonging to the same syndromes. Thus discussed method allows to evaluated structure of dataset that is relevant to target. Experiments with biomedical datasets demonstrated that method allows to outline subgroups of patients with close syndromes and to reveal systems of syndromes that simultaneously exist in sufficiently great groups.

Optimal Valid Partitioning

Let vectors of explanatory variables $X_1,\ldots,X_n$ belong to $\mathbf{M} \subseteq \mathbf{R}^n$. The OVP method implement partitioning of $\mathbf{M}$ that provide for best separation of observations from dataset $\tilde{S}_t = \{(y_1, \mathbf{x}_1),\ldots,(y_m, \mathbf{x}_m)\}$. Partitions are searched inside apriority defined families by optimizing of quality functional. In this paper two partitions families previously described in ([Sen'ko, 2006]) were considered: the simplest Family I includes all partitions with two elements that are divided by one boundary point; two-dimensional Family III including all partitions of two-dimensional admissible areas with no more than four elements that are separated by two boundary lines parallel to coordinate axes. Let R is partition of admissible region of explanatory variables with elements $q_1,\ldots,q_r$. The partition R produces partition of dataset $\tilde{S}_t$ on subsets $\tilde{s}_1,\ldots,\tilde{s}_r$, where $\tilde{s}_j$ $(j = 1,\ldots,r)$ is subset of observations with independent variables vectors belonging to q_j. The evaluated Y mean value for subsets $\tilde{s}_j$ is denoted as $\hat{y}(\tilde{s}_j)$. The integral quality functional $F_I(R,\tilde{S}_t)$ is defined as the sum: $F_I(R,\tilde{S}_t) = \sum_{j=1}^{r} [\hat{y}(\tilde{S}_t) - \hat{y}(\tilde{s}_j)]^2 m_j$, where m_j - is number of observations in subset $\tilde{s}_j$. Partition R_o is considered optimal among partitions from family $\tilde{R}$ if inequality $F_I(R_o,\tilde{S}_t) \geq F_I(R,\tilde{S}_t)$ is true $\forall R \in \tilde{R}$.

The initial variant PT1 is used to test null hypothesis about independence of outcome on explanatory variables related to considered regularity. Estimates of validity indices (p-values) are evaluated at random datasets that are received from dataset $\tilde{S}_t$ by random permutations of target Y relatively fixed positions of $X-$ variables. It was shown in that problem of partially false regularities arise when PT1 is used for verification of regularities that are found inside more complicated models. So additional variant of permutation test (PT2) was developed. Instead of testing null hypothesis that Y is completely independent on $X-$ variables second variant implement testing of

null hypotheses that Y is independent on $X-$ variables inside sub-regions of $X-$ space related to simple regularities that were previously revealed for the same variables. However PT2 can be used for verification of regularity from family III only when simple one-dimensional valid partitions exists for at least one of two relevant explanatory variables, so additional scheme of verification was developed that will be further referred to as PT3. Suppose that R is optimal partition of explanatory variables X' and X'' admissible area that belongs to family III. Let R is described by boundary point b' for variable X' and boundary point b'' for variable X''.

Estimates of p-values are calculated separately for b' and b''.

Validity of optimal boundary $\mathbf{b}'$. To evaluate statistical validity of optimal boundary b' for variable X' we try to test null hypothesis about independence Y on X' and X'' inside subsets formed by boundary b'', At first step optimal partition $R_o(\tilde{S}_t)$ is found for initial training set $\tilde{S}_t$. Two subsets of $\tilde{S}_t$ are formed by boundary b'': subset $\tilde{S}_l = \{(y_1^l, \mathbf{x}_1^l), \ldots, (y_{m_l}^l, \mathbf{x}_{m_l}^l)\}$ includes objects from $\tilde{S}_t$ with $X' \le b''$; subset $\tilde{S}_r = \{(y_1^r, \mathbf{x}_1^r), \ldots, (y_{m_l}^r, \mathbf{x}_{m_r}^r)\}$ includes objects from $\tilde{S}_t$ with $X' > b''$. Estimate of p-value for op b' is calculated by artificial datasets that are built independently from initial datasets $\tilde{S}_l$ and $\tilde{S}_r$ by random permutation of Y values relatively fixed positions of $\mathbf{x}$ descriptions. Let generate N independent permutations of sets of numbers $\{1, \ldots, m_l\}$ and $\{1, \ldots, m_r\}$: $\{\tilde{f}_l^t = \{f_{l1}^t, \ldots, f_{lm_l}^t\},\ \tilde{f}_r^t = \{f_{r1}^t, \ldots, f_{rm_r}^t\} \mid t \in \{1, \ldots, N\}\}$. Then generated permutations are used to build random sets $\{\ \tilde{S}_{pl}^t = \{(y_{f_{l1}^t}^l, \mathbf{x}_1^l), \ldots, (y_{f_{lm_l}^t}^l, \mathbf{x}_{m_l}^l)\}, \tilde{S}_{pr}^t = \{(y_{f_{r1}^t}^r, \mathbf{x}_1^r), \ldots, (y_{f_{rm_r}^t}^r, \mathbf{x}_{m_r}^r)\} \mid t \in \{1, \ldots, N\}\}$. Optimal partition $R_o(\tilde{S}_p^t)$ is found for each $\{t \in \{1, \ldots, N\}$ by union $\tilde{S}_{pl}^t \cup \tilde{S}_{pr}^t$ that is denoted as $\tilde{S}_p^t$. Estimate of p-value is calculated as fraction of permutations with $F_I[R_o(\tilde{S}_t), \tilde{S}_t] \ge F_I[R_o(\tilde{S}_p^t), \tilde{S}_p^t]$.

Validity of optimal boundary $\mathbf{b}''$. Validation procedure for optimal boundary b'' for variable X'' is practically the same and is based test null hypothesis about independence Y on X' and X'' inside subsets formed by boundary b'.

Two-dimensional regularity from family III is considered valid at level β if to inequalities are simultaneously true: $p(b') < \beta$, $X', X'', X'''\ p(b'') < \beta$

Example. Figure 1 represents example of 2-dimensuonal regularity found with the help of technique described in this section. Regularity describes relationship between occurrences of ischemic stroke, polymorphism of gene coding lipoprotein lipase and α-lipoprotein level in patients after transient ischemic attack with chronic cerebral ischemia. The task is described in details in ([Kuznetsova,2013]).

Strong effect of α-lipoprotein level on ischemic stroke risk is seen for cases with H+H+ genotype..For genotypes H-H-,H+H- effect is not so expressed and is opposite by direction. Technique described above was used to calculate p-values to evaluate statistically contribution of polymorphism of gene coding LPL and α-lipoprotein level to considered regularity. It was evaluated by 2000 permutations that for LPL polymorphism p=0.005, for α-lipoprotein level p=0.001.

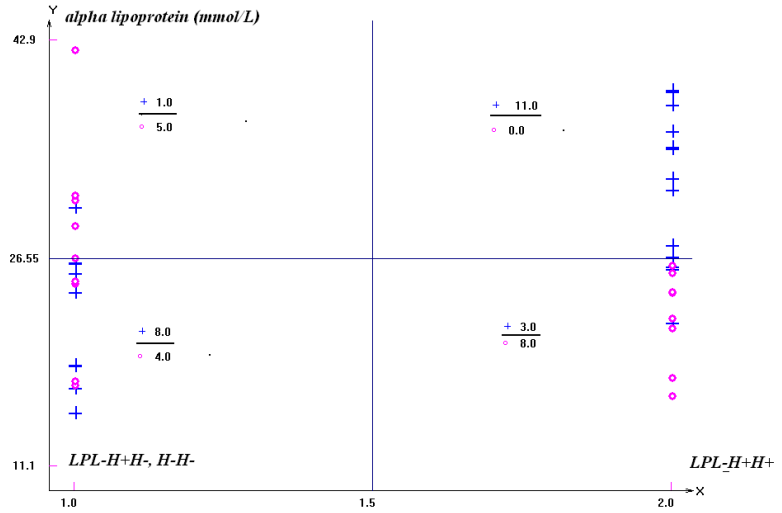

Fig. 1. Sparse diagram describing relationship between polymorphism of gene coding LPL (axis X) and α - lipoprotein level (axis Y), «O» corresponds cases after ischemic stroke, «+» corresponds cases without ischemic stroke.

Syndromes construction

OVP boundaries may be used for construction of syndromes - sub-regions in X-space where mean values of target Y significantly deviate from mean value of Y in training set $\tilde{S}_t$. Suppose that set of one-dimensional regularities $\tilde{r}^1$ and set of two-dimensional regularities $\tilde{r}^2$ were received with the help of OVP technique described in previous section. Let note that regularities from $\tilde{r}^1$ belong to family I and regularities from $\tilde{r}^2$ belong to family III.

Let optimal boundaries b', b'', b''' were found for variables X', X'', X''' with the help of OVP. Then sub-regions of $\mathbf{M}$ that are defined by inequality $X' < b'$, by pair of inequalities $X' < b', X'' < b''$ or by three inequalities $X' < b', X'' < b'', X''' < b'''$ may be examples of one-dimensional, two-dimensional or three-dimensional syndromes correspondingly.

Syndrome quality. Sub-region $q \subset \mathbf{M}$ is considered syndrome only if its quality is sufficient. At that quality of sub-region q is described with the help of functional $\mathrm{X}(q, \tilde{S}_t) = [\bar{y}(q) - \bar{Y}]^2 m(q)$, where $m(q) = |\{s_j = (y_j, \mathbf{x}_j) \in \tilde{S}_t \mid \mathbf{x}_j \in q\}|$, $\bar{y}(q) = \dfrac{1}{\mu(q)} \sum_{\mathbf{x}_j \in q} y_j$, $\bar{Y} = \frac{1}{m} \sum_{j=1}^{m} y_j$ So sub-region q is considered syndrome if $\mathrm{X}(q, \tilde{S}_t) > T_q$ where T_q is initially specified threshold.

Sets of boundaries. It must be noted that for some variables several boundaries may be calculated. Suppose that $\tilde{b}_i$ is set of boundaries for variable X_i that is relevant to regularity r_i^1 from $\tilde{r}^1$ and set of regularities $\tilde{r}_i^2$

from $\tilde{r}^2$. Set $\tilde{b}_i$ consists of boundary for X_i in regularity r_i^1 and regularities from $\tilde{r}_i^2$. Let $\tilde{I}_b$ is set of numbers of variables that are relevant to $\tilde{r}^1$ or $\tilde{r}^2$.

Let describe a necessary condition for family $\tilde{\mathbf{Q}}_l$ of l-dimensional sub-regions of $\mathbf{M}$ to be syndromes family. Suppose that $\tilde{J}$ is set of maps from $\{1,\ldots,l\}$ to $\tilde{I}_b$. Each sub-region $q \in \tilde{\mathbf{Q}}_l$ is characterized:

a) by J_q from $\tilde{J}$;

b) by vector of boundaries $\mathbf{b}(q) = [b_1(q),\ldots,b_l(q)]$, where boundary $b_i(q)$ is taken from $\tilde{b}_{J(i)}$;

c) and by vector of indices $\boldsymbol{\beta}(q) = [\beta_1(q),\ldots,\beta_l(q)]$, where $\beta_i(q) \in \{-1,1\}$.

It is considered that vector $\mathbf{x} = (x_1,\ldots,x_n) \in \mathbf{R}^n$ belongs to q if following inequalities are simultaneously satisfied:

$$x_{J_q(i)}\beta_i(q) < b_i(q),\ i = 1,\ldots,l \tag{1}$$

Structure of dependencies existing in data may be evaluated more exactly by calculating all syndromes with dimension less or equal *k*.

Necessary condition. Let subset q of $\mathbf{M}$ is *l*-dimensional syndrome. Then inequalities (1) must be simultaneously satisfied for any $\mathbf{x} \in q$. Thus simultaneous satisfying of inequalities (1) may be discussed as necessary condition for $q \in \mathbf{M}$ to be syndrome.

As it was mentioned above quality of $q \in \mathbf{M}$ must be sufficient. So inequality $\mathrm{X}(q,\tilde{S}_t) > Tr$ must be satisfied also. But demand of sufficient quality is not single.

For some syndrome with dimension greater than 1 high quality is achieved by some subset of relevant variables. At that another relevant variables are actually irredundant. So additional condition must be used that make it possible to delete irredundant multidimensional syndromes from final syndromes set. Let $\tilde{\mathbf{Q}}_1,\ldots,\tilde{\mathbf{Q}}_k$ are families of $\mathbf{M}$ sub-regions satisfying necessary condition. At that dimension of sub-regions from $\tilde{\mathbf{Q}}_l$ is equal l, $l = 1,\ldots,k$. Following conditions are sufficient for each sub-region from families $\tilde{\mathbf{Q}}_1,\ldots,\tilde{\mathbf{Q}}_k$ to be syndrome.

Sufficient conditions. Let $q_1,\ldots,q_k$ is set of syndromes, where q_l belongs to family $\tilde{\mathbf{Q}}_l$, $l \le k$. Besides $q_l \supset q_{l+1}, l = 1,\ldots,k-1$. Then inequalities

$$h_l \mathrm{X}(q_l,\tilde{S}_t) < \mathrm{X}(q_{l+1},\tilde{S}_t),\ l = 1,\ldots,k-1$$

must be simultaneously satisfied, where $h_l > 1$ is penalty multiplier.

So to find all syndromes with dimension not less than *k* it is sufficient to enumerate all sub-regions that are defined by inequalities (1) and to select sub-regions satisfying conditions (2).

Following procedure may be used of construct all possible syndromes. At initial stage penalty multipliers $h_1,\ldots,h_{l-1}$ are selected.

At first step all one-dimensional syndromes are built by enumerating of all one-dimensional sub-regions satisfying necessary condition and evaluating inequality $\mathrm{X}(q,\tilde{S}_t)>T_q$. At step $k\geq l>1$ all l -dimensional syndromes are built by enumerating of all l-dimensional sub-regions satisfying necessary condition and evaluating for each q_l if sufficient condition $h_{l-1}\mathrm{X}(q_{l-1},\tilde{S}_t)<\mathrm{X}(q_l,\tilde{S}_t)$ is true for any pair (q_{l-1},q_l) where q_{l-1} is syndrome with dimension l that was built at previous step. Search is finished when all $k-$ dimensional syndromes are found.

Clustering method

A set of syndromes $\tilde{\mathbf{Q}}$ defines map from $\mathbf{M}$ to binary hypercube $\mathbf{B}^N$ of dimension $N=|\tilde{\mathbf{Q}}|$. Let $\tilde{\mathbf{Q}}=\{q_1,\ldots,q_N\}$. Binary vector $\mathbf{z}(\mathbf{x})=[z_1(\mathbf{x}),\ldots,z_N(\mathbf{x})]$ is constructed by vector $\mathbf{x}\in\mathbf{M}$ with the help of simple rule: $z_i(\mathbf{x})=1$ if $\mathbf{x}\in q_i$ and $z_i(\mathbf{x})=0$ otherwise, $i=1,\ldots,N$. Our goal is to find groups inside $\tilde{S}_t$ with similar syndromes. Such task may be reduced to search of groups with close z-descriptions. To achieve this goal hierarchical clustering technique is used.

Let ρ is semi-metrics that is defined at $\mathbf{B}^N$. We may use for example standard Hemming metrics. In this study we use semi-metrics $\rho(\mathbf{z}',\mathbf{z}'')=|\{i\mid z_i'=1,z_i''=1,i=1,\ldots,N\}|$. In other words we is defined as fraction of z-variables that are equal 1 in descriptions Let $G'=\{\mathbf{z}_1',\ldots,\mathbf{z}_{m_1}'\}$ and $G''=\{\mathbf{z}_1'',\ldots,\mathbf{z}_{m_2}''\}$ are sets of z-descriptions of objects from $\tilde{S}_t$. Distance between G' and G'' is defined as

$$\boldsymbol{\rho}(G',G'')=\frac{1}{m'm''}\sum_{i=1}^{m'}\sum_{i'=1}^{m''}\rho(\mathbf{z}_i',\mathbf{z}_{i'}'').$$

Initially a threshold T_{cl} is chosen, At first step z-description of each object from $\tilde{S}_t$ is considered cluster. So at first step stage we have set of m clusters $\tilde{G}^0=\{G_1^0=\{s_1\},\ldots,G_m^0=\{s_m\}\}$. Mutual distances between clusters are calculated and minimal distance $\mathrm{P}_{\min}^0$ is selected. Then pair of clusters $(G_{i_b'}^0,G_{i_b''}^0\}$ satisfying equality $\boldsymbol{\rho}(G_{i_b'}^0,G_{i_b''}^0\}=\mathrm{P}_{\min}^0$ is selected. In case $\mathrm{P}_{\min}^0<T_{cl}$ new cluster $G_{i_b'}^0\cup G_{i_b''}^0$ is added to $\tilde{G}^0$ and pair of clusters $(G_{i_b'}^0,G_{i_b''}^0\}$ is removed. Thus we receive new set of clusters $\tilde{G}^1=\{G_1^1,\ldots,G_{m-1}^1\}$.

At step k-1 we have clusters $\tilde{G}^k=\{G_1^k,\ldots,G_{m-k}^k\}$. The same procedure that was used for set $\tilde{G}^0$ is repeated for set $\tilde{G}^k$. Pair of clusters $(G_{i_b'}^0,G_{i_b''}^0\}$ satisfying equality $\boldsymbol{\rho}(G_{i_b'}^0,G_{i_b''}^0\}=\mathrm{P}_{\min}^0$ is selected and new

cluster $G_{i'_b}^{k-1} \cup G_{i''_b}^{k-1}$ is added to $\tilde{G}^{k-1}$ and pair of clusters $(G_{i'_b}^{k-1}, G_{i''_b}^{k-1})$ is removed. Thus we receive new set of clusters including $G_{i'_b}^{k-1} \cup G_{i''_b}^{k-1}$.

Procedure is finished when a) at some step there are no such two clusters that inequality $\mathrm{P}_{\min}^{0} < T_{cl}$ is true for distance between them, b) all objects are put to one clusters. At that case (b) corresponds to absence of cluster structure at level T_{cl}. So it is necessary to select higher level $T'_{cl} > T_{cl}$ to asses cluster structure.

Each cluster G may be characterized by set of syndromes $\tilde{\mathbf{Q}}_G^k \subseteq \tilde{\mathbf{Q}}^k$. Subset $\tilde{\mathbf{Q}}_G^k$ is searched by a threshold T_{cov} that is selected by user. Let $q \in \tilde{\mathbf{Q}}^k$ and $G_q = \{s_j \in G \mid \mathbf{x}_j \in q\}$. Syndrome q belongs to $\tilde{\mathbf{Q}}_G^k$ only when $| G(q) | / | G | > T_{\text{cov}}$.

Let note that set $\tilde{\mathbf{Q}}_G^k$ may be characterized by set of inequalities $IP(\tilde{\mathbf{Q}}_G^k) = \{x_{J_q(i)} \beta_i(q) < b_i(q) \mid i = 1, \ldots, l(q), q \in\}$. So group $IP(\tilde{\mathbf{Q}}_G^k)$ may be considered as short description of group G.

Experiment with biomedical data

Performance of developed technique in gerontology task was evaluated. Effect of clinical and genetic factors on life duration was studied in patients from Moscow population with chronic cerebral ischemia. Two groups of patients were compared by wide set of clinical, biochemical, genetic and instrumental indices: group of 123 long-livers older than 89 (average age 91.0), group of 235 patients of middle and old age (all younger 90).

At the initial stage OVP method was used to search one-dimensional regularities from model I. Valid regularities (p<0.02) were found for 41 variables. Found boundaries and threshold $T_q = 5$ were used to calculate syndromes with dimension 1-3. Thus 56 one-dimensional, 34-two-dimensional and 506 three-dimensional syndromes were found.

Three compact clusters were outlined with the help of technique described in previous section at $T_{cl} = 0.17$.

First cluster GI includes 227 patients: 216 patients from group I and 11 patients from group II. Thus it may be considered that first cluster represents majority of patients with age <90. This cluster is characterized by set $\tilde{\mathbf{Q}}_{GI}^3$ that includes 18 syndromes selected according $T_{\text{cov}} = 0.9$. In other words each of syndromes from $\tilde{\mathbf{Q}}_{GI}^3$ exists for not less than 90% of patients from first cluster. Table 1 includes all inequalities from set $IP(\tilde{\mathbf{Q}}_{GI}^3)$ that describe at least 1 syndrome from $\tilde{\mathbf{Q}}_{GI}^3$.

Table 1.

Glucose >6,4 mmol/L	Aspartate transaminase(AST)> 15,5 units
Diastolic pressure> 72,5 mmHg	Hemoglobin > 115 g/L
Whole protein > 68,5 g/L	Cholesterol > 4,795 mmol /L

Second cluster $G\text{II}$ includes 27 patients older 89 and no patients younger 90. This cluster is characterized by set $\tilde{\mathbf{Q}}^3_{G\text{II}}$ that includes 47 syndromes selected according $T_{\text{cov}} = 1$. This cluster is characterized by such indicators as angina pectoris (II-III functional classes), coronary atherosclerosis, and third stage of chronic cerebral ischemia. It is necessary to note that more than 80% of patients from $G\text{II}$ have B1B2 and B2B2 genotypes of Cholesteryl ester transfer protein (TaqIB polymorphism).

Second cluster does not include patients with systolic arterial pressure below 164,5 diastolic pressure below 90. Patients from second cluster do not have ischemic stroke. Table 2 includes all inequalities from set $IP(\tilde{\mathbf{Q}}^3_{G\text{II}})$ that describe at least 1 syndrome from $\tilde{\mathbf{Q}}^3_{G\text{II}}$.

Table 2.

systolic pressure. < 164,5 mmHg	morbus hypertonicus –no	angina pectoris (II-III fc)- yes
diastolic pressure < 90 mmHg	Ischemic stroke-no	coronary atherosclerosis – yes
B1B2 and B2B2 genotypes of CETP	Smoking –no	CCI – III stage– yes

Third cluster $G\text{III}$ includes 7 patients from group II and systolic arterial pressure higher 164.5. All patients have H-H- genotype of lipoproteinlipaze - LPL (HindIII polymorphism).

Table 3 includes all inequalities from set $IP(\tilde{\mathbf{Q}}^3_{G\text{III}})$ that describe at least 1 syndrome from $\tilde{\mathbf{Q}}^3_{G\text{III}}$.

Table 3.

Systolic pressure. >164,5 mmHg.	General cholesterol>5,0 mmol/L	angina pectoris (II-III fc) -yes
Diastolic pressure > 90 mmHg	Ischemic stroke-no	coronary atherosclerosis – yes
genotype H-H- of LPL	Smoking - no	CCI – III stage– yes

Conclusion

Thus new method of intellectual data analysis was developed that is combination of optimal valid partitioning technique and hierarchical clustering. The method allows discovering in multidimensional feature space sub-regions corresponding to one of target classes (syndromes).

Binary descriptions of objects indicating to what syndromes initial feature descriptions belong are generated at the second stage. Hierarchical cluster analysis is used to discover compact groups in binary descriptions space. So method allows discovering groups of objects that belong to similar syndromes.

Biomedical application was discussed that is aimed to find features related to life duration in patients with chronic cerebral ischemia. It was shown that almost all patients younger 90 were put to one compact cluster in binary descriptions space. At that two clusters were revealed include patients долгожителей only. These two clusters differ by genetic parameters and systolic pressure levels. It is possible that influence of arterial pressure on life duration is also associated with polymorphism of genes that are related to lipid metabolism: gene of lipoproteinlipaze, Developed method may be used in task of biomedical data analysis.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)

Bibliography

[Breiman, 1984] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classifcation and Regression Trees, Chapman & Hall, New York

[Ernst, 2004] Ernst M, J (2004) Permutation methods: A basis for exact inference. *Statistical Science* 19,676-685

[Gorman, 2001] T.W. O'Gorman An adaptive permutation test procedure for several common test of significance. Computational Statistics & Data Analysis. 35(2001) 265-281.

[Kuznetsova, 2011] Kuznetsova A.V., A.V., Kostomarova I.V., Vodolagina N.N. Senko O.V. Study of effects of clinical and genetic factors on severity of discirculatory encephalopathy with the help of pattern recognition methods. *Mathematical biology and bioinformatics.* 2011. T. 6. № 1. C. 115–146. URL: http://www.matbio.org/2011/ Senko2011(6_115).pdf (in russian)

[Kuznetsova, 2013] Kuznetsova A.V., Kostomarova I.V., Senko O.V. Logical and statistical analysis of correlation between disturbances of cerebral circulation and clinical or laboratory indices in old patients with chronic cerebral ischemia. *Mathematical biology and bioinformatics. 2013. T. 8. № 1. C. 182–224. URL: http://www.matbio.org/2013/Kuznetsova_8_182.pdf (in russian)*

[Kuznetsova, 2000] Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabotina T.N., Korotkova O.V. The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges //J. Theor. Med., 2000, Vol. 2, pp.317-327.

[Sen'ko, 2006] Oleg V.Senko and Anna V. Kuznetsova, The Optimal Valid Partitioning Procedures . Statistics on the Internet ***http://statjournals.net/***, April, 2006

[Senko, 2010] Senko Oleg, Kuznetsova Anna, Kostomarova Irina, 2010 Methods for evaluating of regularities systems structure. In book «New Trends in Classification and Data Mining.» ITHEA. Sofia. Bulgaria. P. 40-46, 2010.

[V.V.Ryazanov, 2003] Ryazanov V.V. About some approach for automatic knowledge extraction from precendent data // Proceedings of the 7th international conference "Pattern recognition and image processing", Minsk, May 21-23, 2003, vol. 2, pp.35-40.

Authors' Information

Senko Oleg Valentinovich – Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40, e-mail: *senkoov@mail.ru*

Kuznetsova Anna Victorovna– senior researcher in Institute of Biochemical Physics of Russian Academy of Sciences, Russia, 117997, Moscow, Kosygina, 4, e-mail: *azfor@narod.ru*

Kostomarova Irina Victorovna - Department of Pirogov Russian National Research Medical University "Clinical and Research Center of Gerontology" Russian Ministry of Health, 16, 1st Leonova St., Moscow 129226; e-mail: *hla2222@yandex.ru*