Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

# New Trends
# in
# Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria



**ISBN 978-954-16-0042-9**

C\o Jusautor, Sofia, 2010

# METHODS FOR EVALUATING OF REGULARITIES SYSTEMS STRUCTURE

**Irina Kostomarova, Anna Kuznetsova, Natalia Malygina, Oleg Senko**

***Abstract****: The new method for analysis of regularities systems is discussed. Regularities are related to effect of explanatory variables on outcome. At that it is supposed that different levels of outcome correspond to different subregions of explanatory variables space. Such regularities may be effectively uncovered with the help of optimal valid partitioning technique. The OVP approach is based on searching partitions of explanatory variables space that in the best way separate observations with different levels of outcomes. Partitions of single variables ranges or two-dimensional admissible areas for pairs of variables are searched inside corresponding families. Output system of regularities is formed with the help of statistical validity estimates by two types of permutation tests. One of the problems associated with OVP procedure is great number of regularities in output system in high-dimensional tasks. The new approach for output system structure evaluating is suggested that is based on searching subsystem of small size with possibly better forecasting ability of convex combination of associated predictors. Mean error of convex combination becomes smaller when average forecasting ability of ensemble members becomes better and deviations between prognoses associated with different regularities increase. So minimization of convex combination mean error allows to receive subsystem of regularities with strong forecasting abilities that significantly differ from each other. Each regularity of output system may be characterized by distances to regularities in subsystem.*

***Keywords****: Optimal partitioning, statistical validity, permutation test, regularities, explanatory variables effect, complexity*

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

## Introduction

In many applied forecasting or data analysis possibility of exact prognoses is connected with existing of subregions in explanatory variables $X_1,\dots,X_n$ space where distributions of outcome variable $Y$ significantly differ from its distributions in neighboring subregions or in whole data set. Variety of techniques exists now for revealing of such subregions: linear discriminant functions, classification or regression trees. It must be noted that these techniques are focused mainly on constructing single optimal forecasting algorithm. However in majority of applied tasks it is important not only to construct optimal predicting algorithm but also to receive most complete and valid empirical description of dependences of $Y$ on $X_1,\dots,X_n$. Some approaches that allow to receive empirical descriptions of dependencies were developed during last decade. Method for searching systems of complete or partial logical regularities in pattern recognition tasks may be considered as an example[Ryazanov,2003]. The another approach is constructing of optimal valid partitions (OVP) of $X$ variables space, that allow to achieve optimal separation of observation with different levels of $Y$.

The OVP models were previously developed in [Senko,1998], [Kuznetsova,2000], [Senko,2003]. The OVP procedures allow to calculate the sets of optimal partitions of one-dimensional admissible intervals of single variables or two-dimensional admissible areas of pairs of variables and to estimate statistical validity of regularities associated with these partitions. At that statistical validity is evaluated using two types of permutation tests. Unlike traditional statistical criteria (Chi-square or ANOVA for example) permutation tests allow to evaluate statistical significance by the same dataset which was previously used for boundaries searching. One more advantage of permutation tests before alternative statistical technique is absence of necessity for any suppositions about variables distribution or any restrictions on groups sizes. OVP models may be applied for different types of outcome variables.

One of the problems associated with OVP procedure is great number of regularities in output system in high-dimensional tasks. In typical biomedical tasks with several tens independent variables and several hundreds cases number of discovered statistically valid regularities may achieve several hundreds. In such situations manual analysis of regularities systems is difficult. So development of new computer methods that would allow to evaluate interrelations between regularities, to reveal groups of similar regularities and recover internal structure of regularities system. The approach that is discussed in paper is based on calculating of mutual distances between regularities. Let $r_1$ and $r_2$ is pair of regularities that were found with the help of OVP method. Then two forecasting functions $Z_1$ and $Z_2$ may be put in correspondence to regularities as it is described below. Distance function $\mathrm{P}(r_1, r_2)$ between regularities $r_1$ and $r_2$ is defined as mathematical men of squared difference between $Z_1$ and $Z_2$ or $\mathrm{P}(r_1, r_2) = \mathrm{P}(Z_1, Z_2) = E_\Omega (Z_1 - Z_2)^2$. Distance functions between regularities may be easily estimated by training information.

Various cluster analysis methods may be used for revealing clusters of similar regularities in case distance function $\mathrm{P}$ is defined. However clusterization methods are based only on distances and do not take into account prognostic ability of regularities. Besides previous experiments demonstrated that clusterization techniques tend to form classes that strongly differ from each other by size. So an alternative method was suggested that allow to select from system small number of regularities with possibly better forecasting ability of collective convex predictor. This subset that will be further referred to as basic subset or $\breve{Q}_B$. It was shown previously ([A. Krogh and J. Vedelsby, 1995], [O.V.Senko,2009]) that forecasting ability of collective convex predictor depends both on forecasting ability of single ensemble members and mutual distances $\mathrm{P}(Z_i, Z_j)$ between predictors. At that mean error of convex combination decrease when average forecasting ability of ensemble improves and deviations between prognoses associated with different regularities grows. So minimization of convex combination mean error allows to receive a basic subset $\breve{Q}_B$ with following properties: a) $\breve{Q}_B$ consists of regularities with relatively strong forecasting abilities, b) regularities in $\breve{Q}_B$ significantly differ from one another in terms of distance $\mathrm{P}$. Structure of output system may be characterized by distances to regularities from $\breve{Q}_B$. In other words t distances may be considered as new “coordinate” of regularities from initial set of regularizes that have been found by OVP method. Thus structure of initial system of regularities may be evaluated.

## Optimal Partitioning

Let $Y$ belongs to some set $M_y$. It is supposed that distance function $\rho$ defined on Cartesian product $M_y \times M_y$ satisfies following conditions:

a) $\rho(y', y'') \geq 0$, b) $\rho(y', y'') = \rho(y'', y')$, c) $\rho(y', y') = 0 \quad \forall y', y'' \in M_y$.

The OVP methods are based on optimal partitioning of independent variables admissible regions. The partitions that provide for best separation of observations from dataset $\tilde{S}_0$ with different levels of dependent variable are searched inside apriori defined families by optimizing of quality functional.

***Partitions families.*** The partition family is defined as the set of partitions with limited number of elements that are constructed by the same procedure. The unidimensional and two-dimensional families are considered. The unidimensional families includes partitions of admissible intervals of single variables. The simplest Family I includes all partitions with two elements that are divided by one boundary point. The more complex Family II includes all partitions with no more than three elements that are divided by two boundary points. The two-dimensional Family III includes all partitions of two-dimensional admissible areas with no more than four elements that are separated by two boundary lines parallel to coordinate axes. Family IV includes all partitions of two-dimensional admissible areas with no more than two elements that are separated by linear boundary with arbitrary orientation relatively coordinate axes.

***Quality functionals.*** Let consider at first standard OVP. Let $\tilde{Q}$ is partition of admissible region of independent variables with elements $q_1,\ldots,q_r$. The partition $\tilde{Q}$ produces partition of dataset $\tilde{S}_0$ 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 of subsets $\tilde{S}_j$ is denoted as $\hat{y}(\tilde{S}_j)$. The integral quality functional $F_I(\tilde{Q},\tilde{S}_0)$ is defined as the sum: $F_I(\tilde{Q},\tilde{S}_0)=\sum_{j=1}^{r}\rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_j)]m_j$, where $m_j$ - is number of observations in subset $\tilde{S}_j$.. The optimal value of quality functional in dataset $\tilde{S}$ will be further referred to as $F_I^o(\tilde{S})$.

***Regularities validation.*** For validation of found optimal partitions two types of permutation test is used. The first variant PT1 is used to test null hypothesis about independence of outcome on explanatory variables related to considered regularity. PT1 is used in cases when: a) significance of simplest regularities associated with partitions from family I is evaluated, b) significance of more complicated regularities is evaluated and no simplest valid regularities were previously discovered for related variables. The second variant PT2 is used to evaluate if more complicated partitions models are needed instead of simplest one to describe existing regularities. It tests null hypothesis about independence of outcome on explanatory variables inside suregions of explanatory variables space that are elements of partitions associated with simplest regularities.

## Forecasting associated with regularities

Examples of partitions describing regularities are given at figures 1 and 2. Difference between distributions of various biomedical parameters was evaluated in groups of patients in light and severe stage of encephalopathy Sparse diagram 1 represent regularity associated with relationship between encephalopathy stage on age group number (axe X). Statistics for each quadrant are give at the left part of sparse diagram. It is seen that fraction of light forms decreases as group number increases. Then probability of light form for patient s from quadrant j is calculated by forecasting function $Z(s,j)=\frac{n^+(j)}{n^+(j)+n^o(j)}$. It is seen from sparse diagram 1 that for for quadrant I number of patients in light stage $n^+$ is equal 32 and number of patients in severe stage $n^o$ is equal 11 and $Z(s,I)=0.744$

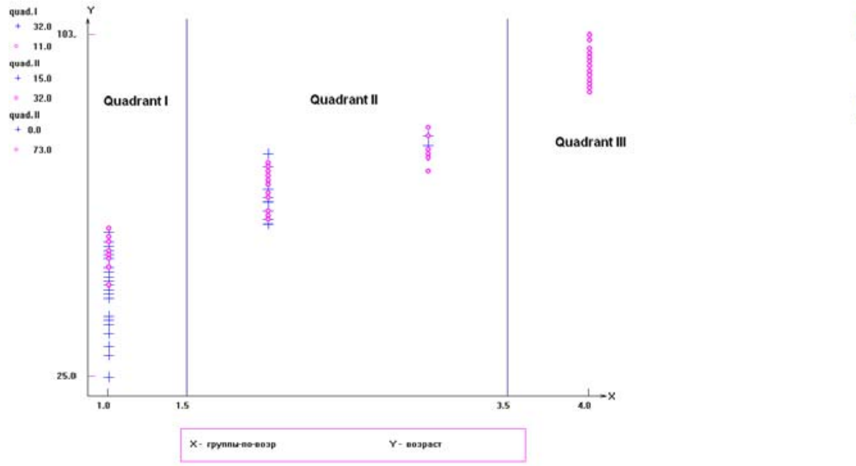


Fig. 1 .Optimal partition describes dependence of encephalopathy form on age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y). Partition belongs to the family III. Patients with severe form are denoted by It is seen that

Sparse diagram 2 represent regularity associated with relationship between encephalopathy stage and pair of input variables: age (axe X) and existing of severe disorder of cerebral blood circulation (axe Y). It is seen that light stage of encephalopathy predominates over severe stage only in quadrant IV corresponding to patients younger 70.5 years without severe disorder of cerebral blood circulation: number of patients in light stage $n^{+}$ is equal 43 , number of patients in severe stage $n^{o}$ is equal 2 and $Z(s, IV) = 0.955$ .

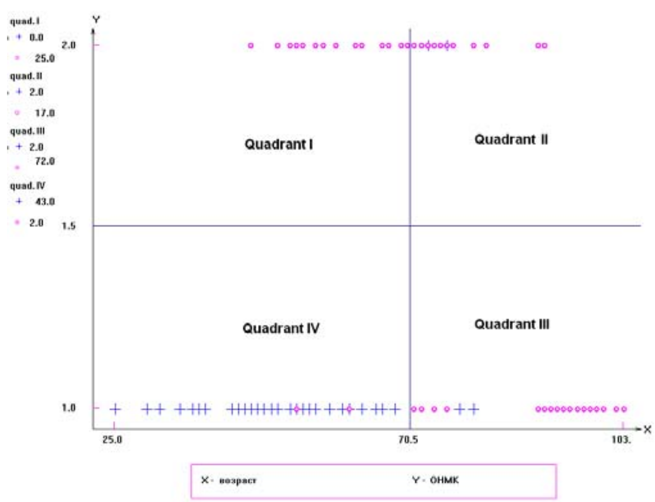


Fig. 2 .Optimal partition describes relationship between encephalopathy stage and pair of variables: age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y). Partition belongs to the family III. Patients with severe form are denoted by o and patients with light form are denoted by +.

## Generalized Error Functional

Let $\tilde{Z} = \{Z_1, \ldots, Z_l\}$ is set of prognostic variables (predictors) forecasting outcome variable $Y$ for objects that are elements of some probability space $\Omega$. The convex corrector $\hat{Z}_{ccp}(\mathbf{c})$ at $\tilde{Z}$ is defined as $\hat{Z}_{ccp}(\mathbf{c}) = \sum_{i=1}^{l} c_i Z_i$, where $\sum_{i=1}^{l} c_i = 1$ and $c_i \geq 0, i = \overline{1,l}$. The squared error of forecasting at general set will be denoted as $\delta$ . Let distance between two predictors The squared error for $\hat{Z}_{ccp}(\mathbf{c})$ may be represented as

$$\delta[\hat{Z}_{ccp}(\mathbf{c})] = \sum_{i=1}^{l} c_i \delta(Z_i) - \tfrac{1}{2} \sum_{i'=1}^{l} \sum_{i'=1}^{l} c_{i'} c_{i''} \mathrm{P}(Z_{i'}, Z_{i''}) \qquad (1)$$

where $\mathrm{P}(Z_{i'}, Z_{i''}) = E_\Omega (Z_{i'} - Z_{i''})^2$ .

It is seen from (2) that $\delta[\hat{Z}_{ccp}(\mathbf{c})]$ is always lower than $\sum_{i=1}^{l} c_i \delta(Z_i)$ and difference between them increases as increase distances between prediction. In case when contributions of all predictors (1) take form

$$\delta[\hat{Z}_{ccp}(\mathbf{u}_l)] = \tfrac{1}{l} \sum_{i=1}^{l} \delta(Z_i) - \tfrac{1}{2} \tfrac{1}{l^2} \sum_{i'=1}^{l} \sum_{i'=1}^{l} \mathrm{P}(Z_{i'}, Z_{i''}) \qquad (2)$$

where $\mathbf{u}_l = (\tfrac{1}{l}, \ldots, \tfrac{1}{l})$ is *l* - dimensional vector.

## Optimal Subset Selecting

At the initial stage optimal system of regularities $\tilde{Q}_o$ is searched using described previously optimal partitioning method. Then optimal subset selecting (OSS) procedure may be used to find $\breve{Q}_B$.

Step 1. At initial step squared error $\delta$ is evaluated for each predictor associated with regularity from system $\tilde{Q}_o$. At that leave-one-out cross validation technique is used. Regularity $q_1^{best} \in \tilde{Q}_o$ corresponding to predictor with minimal error $\delta_1^{\min}$ is added to optimal subset.

Step 2. . The squared error functional (1) is calculated for pairs of predictors associated with pairs of regularities from set $\{(q_1^{best}, q) \mid q \in \tilde{Q}_o \setminus q_1^{best}\}$. Let minimal value of functional (1) $\delta_2^{\min}$ is achieved for pair $(q_1^{best}, q_2^{best})$. In case $\delta_2^{\min} < \delta_1^{\min}$ regularity $q_2^{best}$ is added to optimal subset.

Step k. . The squared error functional (1) is calculated for sets of predictors associated with series of regularities from $\{(q_1^{best}, \ldots, q_{k-1}^{best}, q) \mid q \in \tilde{Q}_o \setminus \{q_1^{best}, \ldots, q_{k-1}^{best}\}\}$. Let minimal value of functional (1) $\delta_k^{\min}$ is achieved for series $(q_1^{best}, \ldots, q_k^{best})$. In case $\delta_k^{\min} < \delta_{k-1}^{\min}$ regularity $q_k^{best}$ is added to optimal subset. Otherwise forming of optimal regularities subset is finished and series $(q_1^{best}, \ldots, q_k^{best})$ is fixed as basic subset. $\breve{Q}_B$.

## Experiment

Performance of developed method was evaluated in task of computer diagnostics of encephalopathy severity. Group of 47 patients in early (first) stage was compared with group of 116 patients in severe stage (third) by 122 input clinical or biomedical indicators. At initial stage of research OVP method was used to find statistically valid correlations between severity and levels of input variables. As a result 300 regularities belonging to families

I, II, III was revealed at statistical significance level $p<0.05$. Previously described procedure of optimal regularities subsystem was used. Basic subset $\breve{Q}_B$ consisting of 3 regularities was found:

two-dimensional regularity associated with relationship between encephalopathy stage and pair of input variables: age (axe X) and existing of severe disorder of cerebral blood circulation that is diagnosed by magnetic resonance tomography (axe Y) .

regularity associated with relationship between encephalopathy stage on age group number (axe X)

a) two-dimensional regularity associated with relationship between encephalopathy stage and pair of input variables: existing of severe disorder of cerebral blood circulation that is diagnosed by magnetic resonance tomography and MPT LO.

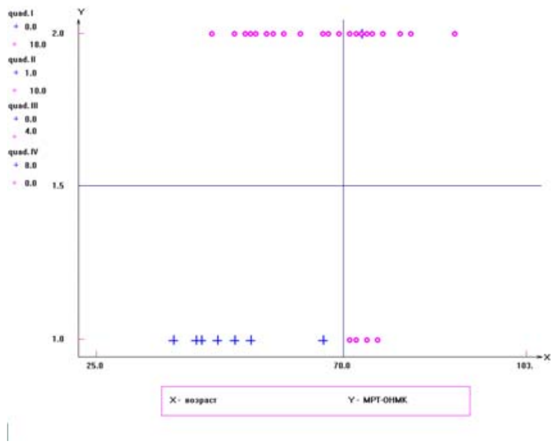


Fig. 3 .Optimal partition describes relationship between encephalopathy stage and pair of variables: age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y) diagnosed by magnetic resonance tomography. Partition belongs to the family III. Patients with severe form are denoted by o and patients with light form are denoted by +.

Forecasting errors associated with these 3 regularities and distances between them are given in next table 1.

Table 1

| | $\delta$ | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | 0.026 | 0 | 0.315 | 0.029 |
| 2 | 0.148 | 0.315 | 0 | 0.296 |
| 3 | 0.053 | 0.029 | 0.296 | 0 |

It is seen from table that prognostic abilities of regularities 1 and 3 from $\breve{Q}_B$ are much better than prognostic ability of regularity 2. But distance of regularity 2 from regularities 1 and 3 is great. Distances to regularities from $\breve{Q}_B$ were calculated for each regularity from all regularities from system $\tilde{Q}_o$, and 20 nearest neighbors were found for each element of $\breve{Q}_B$. It is appeared that regularities close to (a) and (c) are mainly related to existing severe disorder of cerebral blood circulation.

## Conclusion

The new method for analysis of regularities systems was represented. Method is based on searching of regularities subsystems $\breve{Q}_B$ in initial system $\breve{Q}_o$ with the best collective forecasting ability according expression (2). Algorithm for regularities subsystems $\breve{Q}_B$ searching in initial system $\tilde{Q}_o$ is represented. The developed technique was tested at the high-dimensional task of computer diagnostics of encephalopathy severity where using OVP procedure discovered great number of regularities at significance level $p<0.05$.

## Acknowledgment

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA ( www.ithea.org ) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine ( www.aduis.com.ua ).

## Bibliography

[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.

[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.

[Senko, 2003] Senko O.V., Kuznetsova A.V., Kropotov D.A. (2003). The Methods of Dependencies Description with the Help of Optimal Multistage Partitioning. Proceedings of the 18-th International Workshop on Statistical Modelling Leuven, Belgium, 2003, pp. 397-401.

[Sen'ko, 1998] Sen'ko O.V., Kuznetsova A.V. (1998). The use of partitions constructions for stochastic dependencies approximation. Proceedings of the International conference on systems and signals in intelligent technologies. Minsk (Belarus), pp. 291-297.

[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

[O.V.Senko,2009]. An Optimal Ensemble of Predictors in Convex Correcting Procedures// Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2009, Vol. 19, No. 3, pp. 465-468.

[A. Krogh and J. Vedelsby, 1995] Neural network ensembles, cross validation, and active learning. NIPS, 7:231–238, 1995.

## Authors' Information

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

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

***Malygina Natalia Aleksandrovna*** – *chief of laboratory in Russian Clinical-Research Center of Gerontology, principal scientific officer of laboratory of population ageing genetic, 129226, Leonova str. 16 build. 2, Moscow, Russia*

***Kostomarova Irina Victorovna*** - *Russian Clinical-Research Center of Gerontology, chief of laboratory of population ageing genetic, 129226, Leonova str. 16 build. 2, Moscow, Russia erri06@rambler.ru*