METHODS OF REGULARITIES SEARCHING BASED ON OPTIMAL PARTITIONING

Oleg Senko, Anna Kuznetsova

Abstract: The purpose of discussed optimal valid partitioning (OVP) methods is uncovering of ordinal or continuous explanatory variables effect on outcome variables of different types. 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. Statistical validity associated with revealed regularities is estimated with the help of permutation test repeating search of optimal partition for each permuted dataset. Method for output regularities selection is discussed that is based on validity evaluating with the help of two types of permutation tests.

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

Conference: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009

Introduction

In present paper the optimal valid partitioning (OVP) approach to data analysis is discussed. The OVP procedures calculate the sets of optimal partitions of one-dimensional admissible intervals of single variables or two-dimensional admissible areas of pairs of variables and evaluate statistical validity of regularities associated with these partitions. It must be noted that applying standard techniques (F-test, Chi-square and others) for assessing validity by the same datasets which previously has been used for boundaries calculating come across problem of multiple testing (`see [Mazumdar, 2000]). So validity estimates appeared to be too optimistic. One of the ways to calculate adequate estimate is randomized splitting of initial data on two subsets. The first one is used for the boundaries calculating and the second one is used for evaluating of statistical validity. But such approach leads to significant loss of both boundaries exactness and validity levels due to decrease of observations numbers in two datasets. The another way to verify nonrandom character of differences between dependent variable levels in groups of observations formed by partitions is using permutation tests. Discussed below technique that is based on random permutations allows using the same dataset for both purposes: boundaries search and evaluating statistical significance. One more advantage of permutation tests is absence of necessity for any suppositions about variables distribution or any restrictions on groups sizes. Today rather many examples of successful use of permutation technique in different types of tasks [O'Gorman, 2001], [Abdolell, 2002]. Variants of OVP methods using search of optimal partitions inside families of different complexity levels was previously considered by [Senko,1998], [Kuznetsova,2000], [Senko,2003]. Suppose that we study dependence of variable Y on explanatory variables $X_1,\ldots,X_n$ by some empirical dataset $\tilde{S}_0$. Various types of dependent variable are admissible: Y may be continuous variables that are directly observed, vectors of probabilities of several types of events at points in X space, survival curves and so on. The observations from data set $\tilde{S}_0$ must include vectors of independent variables $\mathbf{x}$ and information $\mathcal{Y}$ related to dependent variable

Y . Existence of some common procedure is supposed for evaluating mean values of *Y* by sets of observations . In case Y is directly observed continuous variable $\mathcal{Y}$ is simply value of Y and abovementioned evaluating procedure is reduced to calculating of normal means, evaluating procedure is also reduced to calculating of normal means (fractions of events types) when Y is probabilities vector and $\mathcal{Y}$ is binary vector indicating type of events, in case Y is survival curve $\mathcal{Y}$ is pair including time of last observation and binary indicating if patient is alive. In the last case the Kaplan-Mayer technique is the example of evaluating procedure. The variant of OVP for this type of tasks will be referred to as standard OVP or simply OVP.

But sometimes tasks occur where training set does not contain direct $\mathcal{Y}$-descriptions of single objects but includes only mutual distances between $\mathcal{Y}$-descriptions. However, OVP methods may be applied in such tasks also with the help of special quality functional.. The variant of OVP using only mutual distances between $\mathcal{Y}$-descriptions will be referred to as OVP based on mutual distances or OVPMD.

Optimal Partitioning

Let Y belongs to some set M_y. It is supposed that distance function ρ 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 $\tilde{S}_j$. Besides integral functional $F_I(\tilde{Q}, \tilde{S}_0)$ local functional $F_L(\tilde{Q}, \tilde{S}_0)$ is possible that is defined as $F_L(\tilde{Q}, \tilde{S}_0) = \max_{j=1,\ldots,r} \{\rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_j)] m_j\}$. Unlike integral functional $F_I(\tilde{Q}, \tilde{S}_0)$ local functional $F_L(\tilde{Q}, \tilde{S}_0)$ allows to pick out the most distant from remaining part of $\tilde{S}_0$ subregion of partition. The optimal value of quality

functional in dataset $\tilde{S}$ will be further referred to as $F_I^o(\tilde{S})$ or $F_L^o(\tilde{S})$. In case of OVP-MD The integral quality functional $F_I(\tilde{Q}, \tilde{S}_0)$ is defined as the sum:

$$F_I(\tilde{Q}, \tilde{S}_0) = \sum_{i=1}^{r} \{ \sum_{s_j \in \tilde{S}_i} \sum_{s_{j'} \in \tilde{S}_o \setminus \tilde{S}_i} \rho_y(s_j, s_{j'}) - \frac{m_i(m-m_i)}{m(m-1)} \sum_{s_j \in \tilde{S}_i} \sum_{s_{j'} \in \tilde{S}_i} \rho_y(s_j, s_{j'}) \},$$

where m_j - is number of observations in subset $\tilde{S}_j$. The local functional $F_L(\tilde{Q}, \tilde{S}_0)$ in case of OVP-MD is defined as $F_L(\tilde{Q}, \tilde{S}_0) = \max_{i=1,\ldots,r} \{ \sum_{s_j \in \tilde{S}_i} \sum_{s_{j'} \in \tilde{S}_o \setminus \tilde{S}_i} \rho_y(s_j, s_{j'}) - \frac{m_i(m-m_i)}{m(m-1)} \sum_{s_j \in \tilde{S}_i} \sum_{s_{j'} \in \tilde{S}_i} \rho_y(s_j, s_{j'}) \}$

Regularities validation

For validation of found optimal partitions the permutation test (PT) is used. Advantage of permutation tests is freedom from constraints on probability distribution and size of samples (Senko and Kuznetsova (2006)). The initial variant (PT-1) is based on testing basic null hypothesis that variable Y is fully independent on involved explanatory variables. The optimal value of quality functional F_*^o (it may be F_I^o or F_L^o) is used as PT-1 statistics. Let optimal partition of variable X' admissible interval was found inside families I or II or optimal partition of variables X', X'' joint admissible area was found inside family III for dataset $\tilde{S}_0 = \{(\mathcal{Y}_1, \mathbf{x}_1), \ldots, (\mathcal{Y}_m, \mathbf{x}_m)\}$. Let $F_*^o(\tilde{S}_0)$ is the optimal value of used quality functional. To evaluate statistical validity of discovered regularity set of random permutations $\{\pi_1, \ldots, \pi_N\}$ is calculated with the help of random numbers generator. Initial dataset $\{(\mathcal{Y}_1, \mathbf{x}_1), \ldots, (\mathcal{Y}_m, \mathbf{x}_m)\}$ and permutations $\{\pi_1, \ldots, \pi_N\}$ give rise to permuted datasets $\{\tilde{S}_1^r, \ldots, \tilde{S}_N^r\}$, where $\tilde{S}_j^r = \{(\mathcal{Y}_{\pi_j(1)}, \mathbf{x}_1), \ldots, (\mathcal{Y}_{\pi_j(m)}, \mathbf{x}_m)\}$. For each dataset $\tilde{S}_{\pi_j}^r$ from $\{\tilde{S}_1^r, \ldots, \tilde{S}_N^r\}$ optimal partition is searched inside the same family for the same variable (variables) and by optimizing the same quality functional that were previously used in case of $\tilde{S}_0$. Let $N_{gt}[F_*^o(\tilde{S}_0)]$ is the number of datasets in $\{\tilde{S}_1^r, \ldots, \tilde{S}_N^r\}$ for which $F_*^o(\tilde{S}_*^r) > F_*^o(\tilde{S}_0)$. The ratio $N_{gt}[F_*^o(\tilde{S}_0)]/N$ is used as estimate of PT-1 p-value for regularity discovered in $\tilde{S}_0$ with the help of optimal partitioning. .

The second variant (PT-2) is based on testing more complicated null hypothesis that variable Y is independent on involved explanatory variables only inside some apriori defined subregions of X-space. Let explanatory variables admissible region in X-space is partitioned on subregions $q_1^a, \ldots, q_p^a$. This partition produces the partition of dataset $\tilde{S}_0$ on subsets $\tilde{S}_1^a, \ldots, \tilde{S}_p^a$. The following Monte-Carlo procedure of p-values estimating was used in second PT variant. Datasets $\{\tilde{S}_1^{ar}, \ldots, \tilde{S}_N^{ar}\}$ are generated from $\tilde{S}_0$ with the help of permutations $\{\pi_1^{ar}, \ldots, \pi_N^{ar}\}$. As in the first variant only $\mathcal{Y}$-components positions are permuted and the order of X-components remains fixed. Unlike permutations $\{\pi_1^r, \ldots, \pi_N^r\}$ from the first variant permutations $\{\pi_1^{ar}, \ldots, \pi_N^{ar}\}$ do not include transpositions between $\mathcal{Y}$-components of observations belonging to different

subsets from $\{\tilde{S}_1^a,\ldots,\tilde{S}_p^a\}$. The procedure of p-values calculating by generated datasets $\{\tilde{S}_1^{ar},\ldots,\tilde{S}_{\mathcal{N}}^{ar}\}$ completely coincides with the procedure of p-values calculating in the first variant. The p-values evaluating the independence of Y inside subregions $q_1^a,\ldots,q_p^a$ and calculated by PT-2 will be referred to as $p_2(q_1^a,\ldots,q_p^a)$-values.

Forming set of output regularities

The set of output regularities is selected from the set of found optimal partitions using calculated p-values. To simplify the discussion we shall not differ further between regularity and describing it optimal partition. The first and simplest way is selecting in output set only regularities with calculated p-values less than previously defined threshold p_{thr}. The OVP procedures using this way of selecting will be referred to as OVP-CIS (complexity independent selecting). But series of experiments at simulated data [Senko, 2006] demonstrated that OVP-CIS procedure resuts to falling into output set of so called partially false "regularities" with high validity according PT-1. But the cause of this validity actually is dependence of output only on one of variables describing found "regularity". So another variant of OVP procedure (OVP-CDS) will be discussed below. The basic idea underlying this modification of OVP method is selecting to output set only those optimal partitions from more complicated families II , III or IV where variations between induced groups of observations can not be explained from the viewpoint of previously found regularities from simplest family I. In other words selecting of partitions from complicated families in OVP-CDS (complexity dependent selecting) is based on testing if Y is independent on explanatory variable (variables) inside subregions belonging to simple regularities involving these explanatory variable (variables). So OVP-CDS includes different selecting modes for optimal partitions from family I and optimal partitions from more complicated families. Selecting of partitions from family I in OVP-CDS always precede selecting of optimal partitions from families II and III. Then the second variant of permutation test is used to evaluate the validity of the last. Assume that uncovered regularities from family I involving variables X' and X'' are contained in the output set. The first from these simple regularities includes subregions q_1', q_2' and second regularity includes subregions q_1'', q_2''. Then optimal partition from family II involving variable X' is put to the output set only if $p_2(q_1', q_2')$-values is less than threshold $p_{thr.}$. Optimal partition from families III or IV involving variables X' and X'' is placed to the output set only if both inequality $p_2(q_1', q_2') < p_{thr.}$ and $p_2(q_1'', q_2'') < p_{thr.}$ are satisfied. In case output regularities from family I do not involve variables used in optimal partitions from more complicated families II and III the selecting procedure for the last partitions are the same as in OPV-CIS.

Examples

Example 1 .The task of utera mioma relapse predicting from immunological parameters. The group of 6 patients with relapse is compared with 15 patients for which relapse took place before 2 years after operation. Univariate regularity with two boundary point is represented at Fig. 1.

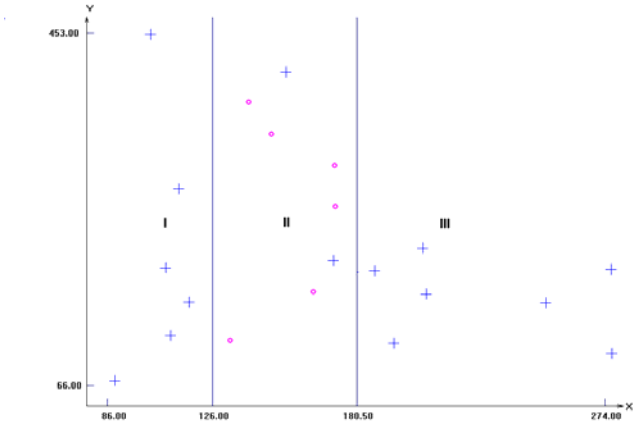

Fig. 1 – Optimal 1-dimensional regularity with two boundary points related to dependence of relapse occurrence on variable. Var. 1 correspond to X, var. 2 correspond to Y, .Quadrant I – number of patients without relapse(+) -6, number of patients with relapse (o) – 0;Quadrant I I– without relapse -2, with relapse – 6;Quadrant III – without relapse -7, with relapse 0;It is seen from figure 1 that variable 1 values in patients with relapse are concentrated inside middle interval: 126.0 <var1<180.5.

Figure 1

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.672450	>0.1	0.755497	0.013 (PF-II,PT-1)

Example 2 . Group of 23 territorial units in Russian Federation with positive migration balance is compared with group of 53 territorial units with negative migration balance. Two-variate regularity with two boundary point related to Task 1.

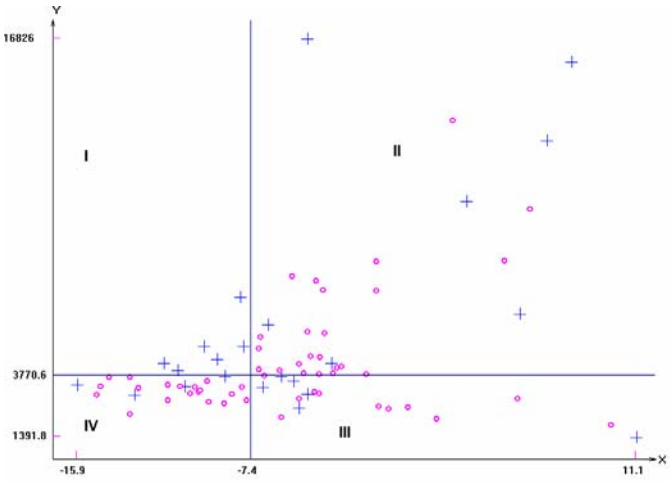

Fig. 2 – Optimal 2-dimensional regularity related to dependence of migration balance on variables 8 and 9 Var. 8 correspond to X, var. 9 correspond to Y, .Quadrant I – number of regions with positive balance (+) -6, number of regions with negative balance(o) – 0; Quadrant I I– positive balance -7, negative balance – 24;Quadrant III – positive balance -6, negative balance – 10;Quadrant IV – positive balance -4, negative balance – 19.

Figure 2

It is seen from figure 1 strong dependence of migration balance on variable 3 in case var2<-7.4, but in case var2>-7.4 a distinct dependence of migration balance on variable 3 is not observed. Statistical validity of regularity according PT-1 is p=0.014

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value var 2	0.686	p>0.1	0.768	0.46 (PF-I, PT-1)
p-value var 3	0.0398	P>0.1	0.062889	0.17(PF-I, PT-1)
2- variate p-value	0.109	-	-	0.014(PF-III, PT-1)

ANOVA F-test reveals valid (p=0.0398) difference between two groups of regions by variable 3 This difference may be related to group of 4 regions in quadrant II with positive balance and high values of variable 3. All univariate tests did not discover any difference between groups of regions by variable 2. No difference was indicated also by 2-variate ANOVA.

Conclusion

The new method for uncovering empirical regularities in data was represented. The method allows to find out regularities related to effect of ordinal or continuous explanatory variables on outcome. Method may be used in tasks with different types of dependent variables; binary scalar outcome, scalar or vector continuous variable, survival curve. Besides method may be used when outcome is not described directly but data contains mutual distances between outcome descriptions for different objects. Method is based on validity estimates with the help of permutation tests. These estimates are free from constraints on probability distribution and sample size. Using of permutation test modification (PT-2) allows to select only regularities with statistically founded inclusion of all constituents (features or boundaries).

Bibliography

[Abdollel, 2002] Abdolell M., LeBlanc M., Stephens D., Harrison R.V. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. //Statistics in Medicine, 2002, 21:3395-3409.

[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.

[Mazumdar, 2000] Mazumdar, M., Glassman, JR. Tutorial in Biostatistics. Categorizing a prognostic variable: review of methods, coding for easy implementation and applications to decision making about cancer treatment. Statistics in Medicine.2000, 19:113-132.

[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 18th 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

Authors' Information

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

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