ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Andrzej Dudek[*], *Marcin Pełka*[**]

EFFECTIVENESS OF SYMBOLIC CLASSIFICATION TREES VS. NOISY VARIABLES

Abstract. In real research problems we usually deal with relevant variables and irrelevant (noisy) variables. Relevant variables sometimes can not be identified, by for example HINoV method or modified HINoV method. This paper compares effectiveness detection of known class structure with application of symbolic decision trees and symbolic kernel discriminant analysis in situation where we deal with noisy variables. This research was conducted on artificial symbolic data from a variety of models. The models contained known structure of clusters. In addition, the models contained different number of noisy variables added to obscure the underlying structure.

Key words: Classification, discrimination, symbolic data, noisy variables.

I. INTRODUCTION

Symbolic Data Analysis is an extension of multivariate analysis dealing with data represented in an extended form. Each cell in symbolic data table (symbolic variable) can contain data in form of and single quantitative value, categorical value, interval, multivalued variable, multivalued variable with weights. Due to extended data representation Symbolic Data Analysis introduces new methods and implements traditional methods that symbolic data can be treated as an input. In case of discriminant analysis two known methods can be adapted for symbolic data: Kernel Discriminant Analysis and classification tress. Article describes both methods and compares the quality of prediction in various scenarios with growing number of noisy variables in learning and test sets.

First part is an introduction to symbolic data analysis, symbolic objects, symbolic variables are described and dissimilarity measures for symbolic objects are presented. Second part shows how methods of discriminant analysis, and of kernel discriminant analysis in particular, may be adapted for symbolic objects.

Third part describes algorithm of creation symbolic classification trees. The forth part presents computational simulation comparing results of discriminant.

[*] Ph.D., Chair of Econometrics and Informatics, University of Economics, Wroclaw.
[**] Ph.D., Chair of Econometrics and Informatics, University of Economics, Wroclaw.

process with use of both methods in various scenarios with growing number of noisy variables in learning and test sets.

Finally some conclusions and remarks are given.

II. SYMBOLIC VARIABLES AND SYMBOLIC OBJECTS

Symbolic data, unlike classical data, are more complex than tables of numeric values. While Table 1 presents usual data representation with objects in rows and variables (attributes) in columns with a number in each cell, table 2 presents symbolic objects with intervals, set and text data.

Table 1. Classical data situation

X	Variable 1	Variable 2	Variable3	...
1	1	108	11.98	
2	1.3	123	–23.37	
3	0.9	99	14.35	
...	...	...	...	...

Source: own research.

Table 2: Symbolic data table

X	Variable 1	Variable 2	Variable 3	Variable 4
1	(0.9; 0.9)	{106; 108; 110}	11; 98	{Blue;green}
2	(1; 2)	{123; 124; 125}	–23;37	{light-grey}
3	(0.9; 1.3)	{100; 102; 99; 97}	14;35	{pale}
...	...	...	...	...

Source: own research.

Bock and Diday (2000) define five types of symbolic variables:

- single quantitative value,
- categorical value,
- interval,
- multivalued variable,
- multivalued variable with weights.

Variables in a symbolic object can also be, regardless of its type (Diday 2002):

- taxonomic – representing hierarchical structure,

- hierarchically dependent,
- logically dependent.

There are four main types of dissimilarity measures for symbolic objects (Malerba *et al.* (2000), Ichino and Yaguchi (1994)):

- Gowda, Krishna and Diday – mutual neighbourhood value, with no taxonomic variables implemented;
- Ichino and Yaguchi – dissimilarity measure based on operators of Cartesian join and Cartesian meet, which extend operators $\cup$ (sum of sets) and $\cap$ (product of sets) onto all data types represented in symbolic object,
- De Carvalho measures – extension of Ichino and Yaguchi measure based on a comparison function (CF), aggregation function (AF) and description potential of an object.
- Hausdorff distance (for symbolic objects containing intervals).

III. KERNEL DISCRIMINANT ANALYSIS OF SYMBOLIC OBJECTS

Most of modern discriminant methods are based on the maximum likelihood rule, which says that an object from test set should be assigned to the class of training set for which the value of distribution density function achieves maximum. In earlier discriminant methods (Altman equation, Fisher analysis) there was an assumption that objects in classes of training sets had normal distribution but in real discrimination problems we cannot make such assumption. Therefore one of main problems of modern discriminant analysis is to estimate distribution density function for each class of the training set.

There are three approaches to achieve this (Hand 1981), Goldstein (1975), Bock and Diday 2000, pp. 235–293)

- linear estimation (Fisher);
- quadratic estimation;
- non-parametric methods.

One of the most commonly used non-parametric methods of estimation of distribution density function is kernel density estimation. Equation (1) represents general form of kernel density estimator (Hand 1981)

$$\hat{f}_k(x) = \frac{1}{n_k (2h_k)^d} \sum_{i=1}^{n_k} K\left(\frac{x - x_{ki}}{h_k}\right) \qquad x \in R^d \tag{1}$$

where:

$\hat{f}_k$ – kernel density estimator,

d – dimension,
k – class number,
n_k – number of objects in k-th class,
h_k – bandwith window for k-th class (a parameter),
$K(.)$ – kernel (Gaussian, Epanechnikov etc.).

In case of symbolic objects space, density distribution is undisputable. The integral operator isn't defined in this kind of space and it's not a subspace of Euclidean space either.

Bock and Diday (2000) introduce a replacement of kernel density estimator for symbolic objects:

$$\hat{I}_k(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \prod_{j=1}^{p} K_{x,h_j}(x_{ki}) \tag{2}$$

where:
p – number of classes in the training set
k – class number,
I_k – kernel intensity estimator,
n_k – number of objects in k-th class,
h_j – window bandwidth for j-th class (parameter),
$K(...)$ – unified kernel for symbolic objects:

$$K_{x,h_j}(y) = \begin{cases} 1 & \text{dla} \quad d_j(x,y) < h_j \\ 0 & \text{dla} \quad d_j(x,y) \geq h_j \end{cases} \tag{3}$$

$d_j(x,y)$ – dissimilarity measure for symbolic objects, one of the dissimilarity measures listed in chapter II.

IV. SYMBOLIC DECISION TREES

Method of creation of decision trees for symbolic data proposed by Perinel and Lechevallier (2000) is based of construction of questions used for choosing the best split of tree. In case of ordered data (ordinal scale and intervals) the question has form

If value of variable Y_i is lover than constraint c ?

In case of nominal data the question may be stated as:

If value of Y_i belongs to set V ? (V is any not-empty subset of domain of variable).

Symbolic decision tree algorithm can be written (in main steps)

- Start
- Repeat until set of admissible nodes is not empty
  - For every admissible node t
  - For every question q
    - Split node t for two temporary terminal nodes l and r
    - Calculate sizes of l and r nodes
    - If sizes of l and r sufficiently big
- Calculate the quality of split W(t,q)
- If W(t,q) greater then threshold value
  - q become candidate-question for t
- else
  - reject question q
- If exists at least one candidate-question for t
  - chose the best question
- else
  - mark t as terminal node
  - if there is no node to split
    - STOP
  - else
    - chose the best split between all nodes

V. SIMULATION

500 (100 for each model) symbolic data sets have been generated for simulation purposes. Parameters of each model are described in table 3.

Table 3. Models of simulation

Model	Number of variables	Number of clusters	Type	Learning set
1	2	2	intervals	200
2	2	2	intervals	200
3	3	3	intervals and categorial	160
4	2	5	intervals	240
5	4	4	intervals and categorial	160

Source: own research.

Table 4 are 5 presents result of discrimination for every model with no noisy-variables, 2,3,5, and 10 noisy variables. For each scenario average error ratio is calculated for Kernel Symbolic Discriminant Analysis (KSDA) and for discrimination with use of Symbolic Discrimination Trees (SDT).

Table 4. Average error ratio (test set is 5% of learning set)

Noisy var	0		2		3		5		10	
Model	KSDA	SDT	KSDA	SDT	KSDA	SDT	KSDA	SDT	KSDA	SDT
1	0.11%	9.06%	9.34%	9.17%	14.34	9.43%	21.23%	9.85%	63.24%	10.12%
2	0.17%	8.07%	8.73%	8.14%	18.25%	8.27%	25.17%	9.05%	58.73%	11.43%
3	0.20%	4.34%	11.23%	5.06%	17.80%	5.78%	28.11%	6.12%	49.56%	8.10%
4	0.14%	9.19%	9.42%	9.56%	16.95%	9.74%	25.01%	10.12%	66.13%	11.50%
5	0.43%	8.43%	12.03%	8.66%	16.25%	9.01%	19.55%	9.67%	61.34%	12.07%

Source: own research with use of SymboliDA package written by authors in R environment.

Table 4. shows result with assumption that test set is 5% of learning set and table 5 show results of simulation in case of test set equal 20% of learning set.

Table 5. Average error ratio (test set is 20% of learning set)

Noisy var	0		2		3		5		10	
Model	KSDA	SDT	KSDA	SDT	KSDA	SDT	KSDA	SDT	KSDA	SDT
1	0,19%	9,04%	9,68%	9,41%	14,23%	8,99%	23,32%	10,12%	66,17%	10,34%
2	0,34%	8,12%	9,43%	8,43%	18,78%	9,03%	27,41%	8,97%	59,78%	11,22%
3	0,25%	4,24%	11,87%	5,76%	18,66%	5,90%	29,05%	6,03%	50.78%	7,89%
4	0,56%	9,09%	9,95%	9,29%	17,97%	9,47%	25,85%	10,01%	67,64%	11,34%
5	0,63%	8,13%	13,43%	8,43%	17,15%	9,41%	20,21%	9,88%	62,45%	12,53%

Source: own research with use of SymboliDA package written by authors in R environment.

VI. FINAL REMARKS

For artificially generated symbolic data with no noisy variables kernel discriminant analysis gives better results than discrimination with use of symbolic classification trees. But while the error ratio in first case rises rapidly when noisy variables are added to data set, in second case number of incorrect predictions is growing much slower.

An open issue for further research is development of method of removing noisy variables in initial stage of discrimination procedure of symbolic objects similar to HINOV (Carmone *et al.* 1999) method for clustering.

REFERENCES

Billard, L., Diday, E. (2006), *Symbolic data analysis. Conceptual statistics and data mining*, Wiley, Chichester.

Bock H.-H., Diday E (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin.

Carmone F.J., Kara, A., Maxwell S. (1999): HINoV: a new method to improve market segment definition by identifying noisy variables, *Journal of Marketing Research*, November, 36, 501-509.

Diday E. (2002), An introduction to symbolic data analysis and the SODAS software, *Journal of Symbolic Data Analysis*, Vol. 1.

Goldstein M. (1975), Comparison of Some Density Estimate Classification Procedures. *Journal of the American Statistical Association*, Sep75 Part I, Vol. 70 Issue 351, p666, 4p;

Hand D.J. (1981), *Kernel Discriminant Analysis*, Wiley, New York

Ichino M., Yaguchi H. (1994), Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 4, 698-707.

Malerba D., Espozito F, Giovalle V., Tamma V. (2001), Comparing Dissimilarity Measures for Symbolic Data Analysis, *New Techniques and Technologies for Statistcs* (ETK-NTTS'01), 473-481.

Verde R.(2004), Clustering Methods in Symbolic Data Analysis, *Classification, Clustering and Data Mining*, Berlin-Springer-Verlag, 299–318.

Andrzej Dudek, Marcin Pełka

SKUTECZNOŚĆ DRZEW KLASYFIKACYJNYCH DLA OBIEKTÓW SYMBOLICZNYCH A ZMIENNE ZAKŁÓCAJĄCE

W rzeczywistych problemach badawczych często oprócz zmiennych istotnych mamy do czynienia ze zmiennymi zakłócającymi (nieistotnymi). Nie zawsze można dokonać wyboru zmiennych istotnych, np. za pomocą metody HINoV, lub zmodyfikowanej metody HINoV. W artykule porównano efektywność wykrywania znanej struktury klas za pomocą drzew klasyfikacyjnych dla obiektów symbolicznych oraz jądrowej analizy dyskryminacyjnej obiektów symbolicznych w sytuacji, gdy mamy do czynienia ze zmiennymi zakłócającymi. Badanie efektywności przeprowadzono na symulowanych danych symbolicznych w różnych modelach. Każdy z modeli zawierał znaną liczbę klas. Dodatkowo do każdego modelu dodano różną liczbę zmiennych zakłócających.