PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Marcin Pełka

THE APPLICATION OF SYMBOLIC KERNEL DISCRIMINANT ANALYSIS IN CREDIT RATING

1. Introduction

The symbolic data analysis is an extension of multivariate analysis dealing with data represented in an extended form. Each symbolic variable can contain single quantitative value, categorical value, interval, multivalued variable, and multivalued variable with weights. Besides that symbolic variables can also be taxonomic, hierarchically dependent, and logically dependent. Therefore symbolic data analysis introduces new methods and implements classical methods, where symbolic data is treated as an input. First part of this article presents aims of discriminant analysis with special focus on the non-parametric kernel density estimation method. Second part introduces terms of symbolic objects and symbolic variable. Third part shows how Bayesian discrimination rule can be adapted to deal with data of different symbolic types, using kernel intensity measures for symbolic data [1, pp. 240-242]. The last part of the article presents results of discrimination analysis for symbolic objects in credit rating and compares its results with credit decision made by a credit officer.

2. Discriminant analysis and kernel density estimation

Discriminant analysis assigns objects from test set to an existing structure of classes (training set).

We usually can't make any assumptions concerning density function of data in real life discrimination problems. To solve this problem we can [2, p. 132]:

a) approximate the unknown density function by applying one of well-known density functions as its estimator,

b) apply one of twelve functions proposed by Pearson and solve differential equation (see [6]),

c) estimate unknown density function with non-parametric methods.

One of the most commonly used non-parametric methods of an estimation of distribution density function is kernel density estimation (see: [7, p. 170]). Equation (1) represents general form of kernel density estimator [1, p. 239; 8, p. 27]:

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

where: $\hat{f}_k(z)$ – uniform kernel density estimator for object z in the k-th class,
$k = 1, 2, \ldots, g$ – number of classes,
n_k – number of objects in k-th class,
h_k – bandwidth window for k-th class (a parameter),
x_{ki} – i-th object in k-th class,
d – dimension equal to number of variables describing object,
$K\left(\frac{z - x_{ki}}{h_k}\right)$ – uniform kernel.

Uniform kernel can take various forms (see [2, p. 134]). In the simplest case its value is equal 1 if all coordinates of its arguments are smaller than 1, in other cases its value is equal to 0.

3. Symbolic objects and variables

Symbolic data unlike classical data situation are more complex than tables of numeric values, table 1 presents usual data representation with object in rows and variables (attributes) in columns with number in each cell while table 2 presents table of symbolic objects with intervals, sets of categories. In many real-life economic problems we deal with symbolic variables instead of classical ones. We get intervals instead single values (points), set of categories instead single categories and so on.

Table 1. Classical data matrix

Objects \ Variables	Income (in PLN)	Seniority (in years)	…	Other collaterals
Client 1	1000	12	…	1
Client 2	2500	1	…	1
Client 3	3000	0.5	…	2
⋮	⋮	⋮		⋮
Client m	675	1	…	3

1 – none; 2 – underwriter; 3 – mortgage.

Source: artificial data.

Table 2. Symbolic data table

Objects \ Variables	Income (in PLN)	Seniority (in years)	...	Other collaterals
Client 1	(1000; 1700)	[0; 0.5]	...	{none}
Client 2	(1500; 2200)	(0.5; 1]	...	(insurance, mortgage}
Client 3	(2000; 2700)	(1; 2]	...	{mortgage}
⋮	⋮	⋮		⋮
Client *m*	(750; 1100)	(2; 3]	...	{insurance, underwriter}

Source: artificial data.

Symbolic data analysis methods were designed to analyze more complex data that is describing either individuals, so called first-order objects, (described by symbolic variables) or groups (classes) of classical individuals, so called second--order objects [1, pp. 18-20].

4. Kernel discriminant analysis for symbolic objects

One cannot discuss the density distribution in the case of a symbolic objects space. The integral operator is not defined in this kind of space and it is not a sub-space of Euclidean space as well.

Let us consider the case where the data are symbolic objects described by seven different types of variables (for example 3 are multivalued variable with weights; 2 are quantitative of interval type; and 2 are multivalued variables). The density estimation can be generalized either using one dissimilarity measure or seven different dissimilarity measures (one for each variable) or three dissimilarity measures (one for each of variable types).

Bock and Diday [1] introduced a replacement of kernel density estimator for symbolic objects [1, p. 242; 10, pp. 127-132]:

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

where: $\hat{I}_k(z)$ – kernel intensity estimator for the object z and the k-th class,
$k = 1, 2, \ldots, g$ – number of classes,
n_k – number of objects in k-th class,
h_k – bandwidth window for k-th class (a parameter),
$j = 1, 2, \ldots, p$ – number of dissimilarity measures applied,
$K_{z,h_j}(x_{ki})$ – kernel for object z and x-th object in k-th class, defined as follows:

$$K_{z,h_j}(x_{ki})=\begin{cases}1 & \text{for } d_j(z,x_{ki})<h_j\\ 0 & \text{for } d_j(z,x_{ki})\geq h_j\end{cases}, \tag{3}$$

$d_j(z,x_{ki})$ – dissimilarity measure for symbolic objects.

Many dissimilarity measures are described in [1, pp. 166-183; 9, pp. 473-481]. Posterior probabilities of the class for z-th object are given as [1, p. 244]:

$$q_k(z)=\frac{\hat{p}_k\hat{I}_k(z)}{\sum_{i=1}^{g}\hat{p}_i\hat{I}_i(z)}, \tag{4}$$

where: $\hat{p}_k$ – prior probabilities for the k-th class,
$\hat{I}_k(z)$ – intensity estimator for the z-th object and the k-th class,
$i=1,2,\ldots,g$ – number of classes.

Prior probabilities ($\hat{p}_k$) could be equal for each class $\hat{p}_k=\frac{1}{g}$, or they can consider proportions observed in the training set $\hat{p}_k=\frac{n_k}{N}$, or they could be obtained by maximizing the EM-like algorithm $\hat{p}_k(t+1)=\frac{1}{m}\sum_{j=1}^{m}\left(\frac{\hat{p}_k\hat{I}_k}{\sum_{i=1}^{g}\hat{p}_i\hat{I}_i}\right)$ for $i=1,2,\ldots,g$ number of classes and t steps of iteration for m points to be classified [1, pp. 242-243].

5. Credit rating with application of symbolic kernel discriminant analysis

Training set contains 80 objects describing BGŻ S.A. Department in Kłodzko bank customers in year 2004. It has been divided into two classes. The first one contains 60 objects pre-classified as borrowers and the second contains 20 clients with negative credit decisions (chosen from 45 negative credit decisions). The test set contains 20 objects. Each of the objects has been characterized by fourteen variables:

1. V_1 – average account incomes – quantitative of interval type in thousands,
2. V_2 – borrowers seniority – quantitative of interval type,
3. V_3 – duration of a credit in months – quantitative of interval type,
4. V_4 – borrowers income – quantitative of interval type in thousands,

5. V_5 – applied amount of a credit – quantitative of interval type in thousands,
6. V_6 – credit record – set of categories received from BIK (credit information bureau) and MIG BR (banks list of unreliable clients),
7. V_7 – client seniority in a bank – set of categories,
8. V_8 – underwriter – set of categories,
9. V_9 – underwriters reliability rating – set of categories,
10. V_{10} – other collaterals – set of categories,
11. V_{11} – clients internal rating – set of categories,
12. V_{12} – evaluation of clients loyalty – set of categories,
13. V_{13} – credit information given by a client – set of categories,
14. V_{14} – allocation of a client to a given class – nominal.

For storing information about training set Microsoft Access 2000 has been used and for assigning object from test set to classes Symbolic Official Data Analysis Software (SODAS) modules DB2SO (extracting objects from database to SODAS), DI (distance measurement) and DKS (symbolic kernel discriminant analysis).

Table 3. Posterior probabilities for test set

No. of object in test set	Posterior probabilities for a class		Maximum probability
	Class 1	Class 2	
1	0.7219	0.2781	Class 1
2	0.4248	0.5752	Class 2
3	0.7249	0.2751	Class 1
4	0.5710	0.4290	Class 1
5	0.6357	0.3643	Class 1
6	0.5679	0.4321	Class 1
7	0.4285	0.5715	Class 2
8	0.6327	0.3673	Class 1
9	0.5872	0.4128	Class 1
10	0.6987	0.3013	Class 1
11	0.4261	0.5739	Class 2
12	0.2459	0.7541	Class 2
13	0.4225	0.5775	Class 2
14	0.4395	0.5605	Class 2
15	0.4259	0.5741	Class 2
16	0.4320	0.5680	Class 2
17	0.4329	0.5671	Class 2
18	0.3578	0.6422	Class 2
19	0.2547	0.7453	Class 2
20	0.3658	0.6342	Class 2

Source: own computation (SODAS software).

Ichino-Yaguchi non-standardized dissimilarity measure was applied in the research (see [1, pp. 166-183; 9]).

Prior probabilities have been estimated considering the proportions observed in training set: 0.75 for class 1 and 0.25 for class 2 posterior probabilities are presented in table 3.

Information from table 3 allows us to compare decision made by credit officer and decision resulting from symbolic kernel discriminant analysis. Correctness of classification is presented in table 4.

Table 4. Correctness of classification

No. of object in test set	Decision resulting from discriminant analysis	Bank's decision	Is object correctly classified?
1	Class 1	Class 1	Yes
2	Class 2	Class 1	No
3	Class 1	Class 1	Yes
4	Class 1	Class 1	Yes
5	Class 1	Class 1	Yes
6	Class 1	Class 1	Yes
7	Class 2	Class 1	No
8	Class 1	Class 1	Yes
9	Class 1	Class 1	Yes
10	Class 1	Class 1	Yes
11	Class 2	Class 1	No
12	Class 1	Class 1	Yes
13	Class 2	Class 2	Yes
14	Class 2	Class 2	Yes
15	Class 2	Class 2	Yes
16	Class 2	Class 2	Yes
17	Class 2	Class 2	Yes
18	Class 2	Class 2	Yes
19	Class 2	Class 2	Yes
20	Class 2	Class 2	Yes

Source: own computation.

By analyzing table 4 it can be said, that that 17 out of 20 objects where correctly classified, so the percentage of correct classification is 0.85. This value was reached by selecting a bandwidth parameter at average distance between all objects from training set 0.07420. This bandwidth parameter provides optimal rate of correctly classified objects. Other most used in literature bandwidth parameters (like 1 or 2) provided worse results (rate of correct classification equal to 0.384615 if $h = 1$ or 2).

6. Summary

A relatively small training sample allowed to get a high percentage of the accuracy of borrowers classification. A bigger sample might have provided even more accuracy. It is not a result sampling technique or sample characteristics nor the chosen period. For artificially generated symbolic data with no noisy variables symbolic kernel discriminant analysis gives high percentage of the accuracy (see [4]).

Clients who were denied by a bank to get a credit, would also receive a negative decision in the case of kernel discriminant analysis for symbolic objects.

The highest percentage of correctly classified clients is achieved when a bandwidth parameter h is set on a level of the average distance between the objects from training set.

Three out of four clients, who would not get a credit in the case of applying discriminant analysis for symbolic objects, had problems with the subsequent repayments of a credit.

No comparisons with classical estimators have been made because when we are dealing with symbolic data we need to transform symbolic variables to classical ones and then apply classical methods. Such comparisons are an opened issue for further research.

Literature

[1] Bock H-H., Diday E. (eds.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg 2000.

[2] Domański C., Pruska K., *Nieklasyczne metody statystyczne*, PWE, Warszawa 2000.

[3] Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Prace Naukowe Akademii Ekonomicznej nr 1021, AE, Wrocław 2004, s. 100-106.

[4] Dudek A., Pełka M., *Effectiveness of Symbolic Classification Trees vs. Noisy Variables*. Folia Oeconomica, Acta Universitatis Lodziensis, Łódź (in review).

[5] Dudek A., *Zastosowanie analizy dyskryminacyjnej obiektów symbolicznych do filtrowania poczty elektronicznej*, Folia Oeconomica, Acta Universitatis Lodziensis, Łódź 2005.

[6] Feldman W., *Kryterium wyboru krzywych Pearsona*, „Przegląd Statystyczny" 1975 nr 22/1.

[7] Hand D, Mannila H, Smyth P., *Principles of Data Mining*, MIT Press, Cambridge 2001.

[8] Härdle W., Simar L., *Applied Multivariate Data Analysis*, Springer Verlag, Berlin-Heidelberg 2003.

[9] Malerba D., Esposito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analisys*, [w:] P. Nanopoulos (ed.), *New Technics and Technologies for Statistics and Exchange of Tehnology and Know-how*, (ETK-NTTS'01) Post conference materials, s. 473-481.

[10] Rasson J.F., Lissoir S., *Symbolic Kernel Discriminant Analysis*, „Computational Statistics" 2000 issue 15, s. 127-132.

ZASTOSOWANIE JĄDROWEJ ANALIZY DYSKRYMINACYJNEJ OBIEKTÓW SYMBOLICZNYCH DO OCENY ZDOLNOŚCI KREDYTOWEJ

Streszczenie

Celem artykułu jest przedstawienie możliwości zastosowania jądrowej analizy dyskryminacyjnej obiektów symbolicznych do oceny zdolności kredytowej osób fizycznych. Artykuł pokazuje również, jak „klasyczna" analiza Bayesowska może być zaadaptowana dla różnych typów danych symbolicznych za pomocą jądrowego estymatora intensywności dla obiektów symbolicznych. W części empirycznej dokonano oceny zdolności kredytowej osób fizycznych na podstawie danych uzyskanych z roku 2004 dla banku BGŻ SA Oddział w Kłodzku.

Marcin Pełka – dr, asystent w Katedrze Ekonometrii i Informatyki Uniwersytetu Ekonomicznego we Wrocławiu – Wydział w Jeleniej Górze.