ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Andrzej Dudek**

DISCRIMINATION OF SYMBOLIC OBJECTS

Abstract. 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 single quantitative value, categorical value, interval, multivalued variable, multivalued variable with weights. Variable can be taxonomic, hierarchically dependent, logically dependent. Due to extended data representation Symbolic Data Analysis introduces new methods and also implements traditional methods that symbolic data can be treated as an input. Article shows how "classical" Bayesian discrimination rule can be adapted to deal with data of different symbolic types, presents kernel intensity measures for symbolic data and methods of obtaining probabilities of belongings to the classes. The example of using symbolic discriminant analysis for electronic mail filtering is given.

Key words: discrimination, symbolic object, Kernel density estimators.

1. INTRODUCTION

Bayesian discriminant analysis is a well-known method, which is often used in multivariate data analysis. However this method has recently found an unexpected usage in computer science and is used to filter unsolicited electronic mail (spam). This paper describes a computational example of discriminant analysis of symbolic objects representing e-mails.

Discriminant analysis goals and the methods of estimating distribution density functions for each class are described in first part of the article with special focus on non-parametric kernel density estimation method.

The second part introduces notions of symbolic objects and symbolic variable and describes main dissimilarity measures for symbolic objects.

* Ph.D., Department of Econometrics and Computer Science, University of Economics in Wrocław.

The third part shows how methods of discriminant analysis, and of kernel discriminant analysis in particular, may be adapted for symbolic objects.

Finally, the described methods are used for filtering electronic mail. The procedure assigns two symbolic objects, each with seven variables, to two classes, one containing 17 messages pre-classified as spam and one containing 13 legitimised mails.

The paper finishes with conclusions, including suggestions for future areas of research.

2. DISCRIMINANT ANALYSIS AND KERNEL DENSITY ESTIMATION

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

Most of 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. This rule is equivalent to the Bayesian rule, which defines misclassification cost in terms of *a priori* and *a posteriori* probabilities.

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 (see: Hand 1981; Goldstein 1975; Bock, Diday 2000, p. 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)

$$\hat{f}_k(\mathbf{x}) = \frac{1}{n_k (2h_k)^d} \sum_{i=1}^{n_k} K\left(\frac{\mathbf{x} - \mathbf{x}_{k_i}}{h_k}\right), \quad x \in R^d \tag{1}$$

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.

Kernel can obtain various forms. In the simplest case its value equals 1 if all coordinates of its argument all smaller than 1; in other cases it is equal to 0.

3. SYMBOLIC OBJECTS AND SYMBOLIC VARIABLES

3.1. Symbolic data table

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

Table 1

Classical data situation

X	Variable 1	Variable 2	Variable 3
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.

H.-H. Bock and E. 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 (D i d a y 2002):

- taxonomic – representing hierarchical structure,
- hierarchically dependent,
- logically dependent.

3.2.Dissimilarity measures for symbolic objects

Because of the structure of symbolic objects, usual measures like Manhatan distance, Euclidean distance, Canbererra distance or Minkowski metrics cannot be used. With symbolic data, other measures must be used.

D. M a l e r b a *et al.* (2001) define three main types of dissimilarity measures for symbolic objects:

- 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 $\cup$ (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.

Table 3 compares the formulas of these measures.

Table 3

Dissimilarity measures for symbolic data

No.	Dissimilarity measure for variables	Dissimilarity measure for objects
1	$D^{(j)}(A, B) = D_n(A, B) + D_S(A, B) + D_C(A, B)$	$d(O_1, O_2) = \sum_{j=1}^{p} D^{(j)}(A_j, B_j)$
2	$\phi(A, B) = \lvert A \oplus B \rvert - \lvert A \otimes B \rvert + \gamma(2 \cdot \lvert A \oplus B \rvert - \lvert A \rvert - \lvert B \rvert$	$d_q(O_1, O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} \varphi(A_i, B_i)^q}\right)$
3	$(A, B) = \frac{\varphi(A, B)}{\lvert Y \rvert}$	$d_q(O_1, O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} \psi(A_i, B_i)^q}\right)$
4	$(A, B) = \frac{\varphi(A, B)}{\lvert Y \rvert}$	$d_q^1(O_1, O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} [w_j \psi(A_i, B_i)]^q}\right)$

Table 3 (cd.)

No.	Dissimilarity measure for variables	Dissimilarity measure for objects
5	$d_i(A, B)\ i = 1, 2, ..., 5$	$d_q(O_1, O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} [w_j d_i(A_i, B_i)]^q}\right)$
6	$(A, B) = \frac{\phi(A, B)}{\mu(A \otimes B)}$	$d'_q(O_1, O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} \frac{1}{p} [\psi'(A_i, B_i)]^q}\right)$
7		$d'_1(O_1, O_2) = [\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))]$
8		$d'_2(O_1, O_2) = [\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))]/\pi(O_1^E)$
9		$d'_3(O_1, O_2) = [\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))]/\pi(O_1 \oplus O_2)$
10	$d_i(A, B), i = 1, 2, ..., 5$	$d''_q(O_1, O_2) = \left(\sqrt[q]{\frac{\sum_{i=1}^{p} [w_j d_i(A_i, B_i)]^q}{\sum_{j=1}^{p} \delta(j)}}\right)$

O_1, O_2 – represent symbolic objects with variables (Aj, Bj).

Source: own researched based on: Bock, Diday 2000; Diday 2002; Gatnar 1998; Malerba *et al.* 2001.

4. DISCRIMINANT ANALYSIS OF SYMBOLIC OBJECTS

4.1. Kernel density estimation for symbolic objects

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.

H.-H. Bock and E. Diday (2000) introduce a replacement of kernel density estimator for symbolic objects

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

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

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

$d_j(x, y)$ – dissimilarity measure for symbolic objects, one of the dissimilarity measures from Tab. 3.

4.2. Finding *a posteriori* probabilities for kernel intensity estimators

An algorithm of finding post-probabilities of belonging to classes of training sets for each object in the test set is iterational. Starting from equal probabilities for each class, it determines the probability in t-th step according to the following formula (B o c k, D i d a y 2000):

$$\hat{p}_{kt+1} = \frac{1}{m} \sum_{j=1}^{m} \left(\frac{\hat{p}_{kt} \hat{I}_k(x_j)}{\sum_{l=1}^{g} \hat{p}_{lt} \hat{I}_l(x_j)} \right) \tag{4}$$

where:

g – number of classes in a training set,
m – number of all objects in a training set,
k – class number,
I_k – kernel intensity estimator,
t – step of iteration.

5. SPAM FILTERING WITH DISCRIMINANT ANALYSIS FOR SYMBOLIC OBJECTS

In the research, the training set contained 30 objects describing electronic messages. It has been divided into two classes, one containing 17 objects classified as spam and the other containing 13 legitimised messages. Each object has seven parameters:

- length of message;
- number of attachments;
- number of receivers;
- key-words;
- title;
- sender's address;
- 1 if sender server is in Open Relay DataBase[1], 0 in other cases.

The first three variables are numerical, the fourth and fifth are multi-valued, the sixth variable is categorical and the seventh is a Boolean variable.

For storing information about messages from the training set Microsoft Access 2000 has been used, and for assigning objects from the test set to classes – Symbolic Official Data Analysis Software (SODAS) modules:

- DB2SO,
- DI,
- DKS.

The training set had two objects. Their contents are listed in Fig. 1.

Test set – object 1

Received: from unilodge.com.au (61.110.152.158)
by oscar.ae.jgora.pl with MERCUR Mailserver (v4.02.30 Mjk4NjItNjQwNS0xO-TIxMg==)
for <andrzej@e.jgora.pl>; Fri, 15 Oct 2004 05:44:28 +0200
Received: from 152.109.219.62 by smtp.sebank.se;
Fri, 22 Oct 2004 03:43:03 +0000
Message-ID: <03b801c4b7e9$8102fe47$93e852ff@unilodge.com.au>
From: "Irma Tillman" <irmatillmandn@sebank.se>
To: andrzej@ae.jgora.pl
Subject: Order Rolex or other Swiss watches online
Date: Fri, 15 Oct 2004 07:43:02 +0400
X-Envelope-To: <andrzej@ae.jgora.pl>
X-Envelope-From: – irmatillmandn@sebank.se

Heya,
Do you want a rolex watch?
In our online store you can buy replicas of Rolex watches. They look and feel exactly like the real thing.

[1] The commonly known black- and grey-lists of spammers' IP-addresses, available on http://ww.ordb.org. Many popular e-mail servers use these lists to deny access for spammers.

Test set – object 2

Received: from pop.uni.lodz.pl (212.191.64.2) by oscar.ae.jgora.pl with MERCUR Mailserver (v4.02.30 Mjk4NjItNjQwNS0xOTIxMg==) for <andrzej@oscar.ae.jgora.pl>; Thu, 21 Oct 2004 12:44:23 +0200
Received: from mail.uni.lodz.pl (212.191.64.8) by pop.uni.lodz.pl (MX V5.3 An4q) with SMTP; Thu, 21 Oct 2004 12:38:50 +0200
Received: ...
From: "Konferencja MSA" <msa@uni.lodz.pl>
To: ... <marekw@oscar.ae.jgora.pl>,
<andrzej@oscar.ae.jgora.pl>, <abak@oscar.ae.jgora.pl>,
Subject: konferencja MSA'04
Date: Thu, 21 Oct 2004 12:37:42 +0200
X-Envelope-To: <andrzej@oscar.ae.jgora.pl>
X-Envelope-From: <msa@uni.lodz.pl>

Szanowni Uczestnicy Konferencji Wielowymiarowa Analiza Statystyczna = MSA'2004!
W załączniku przesyłamy program konferencji. Referenci będą
mieli do dyspozycji rzutnik pisma (folie) oraz rzutnik multimedialny.

Fig. 1. Test set

Source: own research.

An output of kernel discriminant symbolic analysis is presented in Fig. 2.

```
**************************************************
SODAS FILE c:\sodas\spam15.sds
**************************************************
   8 VARIABLES 32 INDIVIDUALS 3 CLASSES
          * C1: 17 TRAINING OBJECTS
          * C2: 13 TRAINING OBJECTS
          * C3:  0 TRAINING OBJECTS
                ----------------------------------------
                30 TRAINING OBJECTS
                 2 OBJECTS TO CLASSIFY
SMOOTHING PARAMETER:          1.0643
LOO ESTIMATED ERROR RATE:       0%
PRIOR PROBABILITIES: C1:        0.333
                     C2:        0.333
                     C3:        0.333
```

POSTERIOR PROBABILITIES:

OBJECT		CLASSE 1	CLASSE 2	CLASSE 3	MAX	
1	31		1	0	0	1
2	32		0.143	0.857	0	2

Fig. 2. Results of discriminant analysis of objects from the test set

Source: own research. Report file from SODAS software.

Object 1 has been classified as spam with 100% probability, object 2 has been classified as non-spam with 85.7% probability. These results quite sufficiently correspond with the intuitive nature of emails described by object.

6. CONCLUSIONS

- Methods of discriminant analysis based on non-parametric distribution density estimation can be adapted to symbolic data.
- Discriminant analysis of symbolic objects can be used for filtering incoming e-mail messages and marking spam.
- The results are promising but also quite preliminary. The relatively small size of training and test sets is implicated by the fact that process of creating symbolic objects describing messages has not been automated.
- More accurate measuring of quality of filtering requires full automation of the process and can be obtained by creating a simple POP3/IMAP client combined with text parser, symbolic object generator and algorithms described in the paper. Author is currently working on such a heuristic, symbolic, Bayesian anti-spam filter and hopes to share the results in not too far a future, but for now the problem is an open issue.

REFERENCES

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

Diday E. (2002), *An introduction to symbolic data analysis and the SODAS software*, J.S.D.A., International E-Journal, http://www.jsda.unina2.it/newjsda/volumes/VOLO/Edwin.PDF.

Dudek A. (2004), *Miary podobieństwa obiektów symbolicznych. Odległość Ichino–Yaguchiego*, "Prace Naukowe Akademii Ekonomicznej we Wrocławiu", nr 1021, 100–106.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Goldstein M. (1975), *Comparison of Some Density Estimate Classification Procedures.* "Journal of the American Statistical Association", Part I, **70**, Issue 351, 666–669.
Hand D. J. (1981), *Kernel Discriminant Analysis*, Wiley, New York.
Holden S. (2004), *Porównanie serwerowych filtrów bayesowskich*, "Hakin9", **2**, 62–71.
Malerba D., Espozito F., Giovalle V., Tamma V. (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis*, "New Techniques and Technologies for Statistcs" and "Exchange of Technology and Know-how" conference materials (ETK-NTTS'01), 473–481.
SODAS. Documentation, SODAS package documentation v.1.20, available at http://www.ceremade.dauphine.fr/~touati/aidedoc/.

Andrzej Dudek

DYSKRYMINACJA OBIEKTÓW SYMBOLICZNYCH

Symboliczna analiza danych jest rozszerzeniem metod wielowymiarowej analizy statystycznej ze względu na sposób reprezentacji danych. Każda komórka w symbolicznej tablicy danych (zmienna symboliczna) może reprezentować dane w postaci liczb, danych jakościowych (tekstowych), przedziałów liczbowych, zbioru wartości, zbioru wartości z wagami. Zmienne mogą ponadto reprezentować strukturę gałęziową oraz być hierarchicznie lub logicznie zależne. Ze względu na sposób reprezentacji symboliczna analiza danych wprowadza nowe metody ich przetwarzania oraz tak implementuje metody tradycyjne, żeby dane symboliczne mogły być ich danymi wejściowymi. W artykule pokazano, jak „klasyczna" analiza Bayesowska może być zaadoptowana dla różnych typów danych symbolicznych za pomocą jądrowego estymatora intensywności dla obiektów symbolicznych. Całość jest zakończona przykładem zastosowania analizy dyskryminacyjnej obiektów symbolicznych do filtrowania przychodzącej poczty elektronicznej.