PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **16** (2004), 299-308

PLISKA
STUDIA MATHEMATICA
BULGARICA

COMPARING SEVERAL METHODS OF DISCRIMINANT ANALYSIS ON THE CASE OF WINE DATA

Dimitar Vandev and Ute Römisch[1]

The main problem of this European wine project (WINE-DB) is the identification of the geographical origin based on chemico-analytical measurements. At first the type of data collected in preparation of this project will be analysed.

Then different procedures of Discriminant analysis are described. Our special attention will be focused to some new techniques as Support Vector Mashines (also known as Kernel Mashines) – procedures from the field of Mashine Learning.

We test traditional techniques of Linear, Quadratic and Nonparametric Discriminant Analysis as well as the Support Vector Mashines on the base of our data and comment the results.

1. Introduction

The wine controlling authorities often are confronted with products which are not correctly marked with regard to their origin, vintages and quality parameters.

In order to discover such adulterations of wines, the identification of their correct geographical origin is of great interest to wine consumers and producers.

Consequently diverse wines were measured and the possibility of recognition had to be tested statistically.

[1]Partially supported by contracts: PRO-ENBIS: GTC1-2001-43031 and WINE DB: G6RD-CT-2001-00646

2000 *Mathematics Subject Classification:* 62H30, 62J20, 62P12, 68T99

Key words: application linear quadratic discriminant analysis SVM

2. Description of data

East European wines

The first data set consisted of 144 different wine samples from five East European countries and 82 analytical parameters. The data were analysed by methods of variance, correlation, factor and discriminant analysis [Römisch et al., 2001], [Römisch et al., 2002].

Most of the data (80%) were lognormally distributed, they were transformed. The others could be assumed as normally distributed.

The number of missing values was partially high – only for 21% of observations all variables were measured. These missing values were filled with within groups means.

Some variables and cases were to be deleted:

- Variables which were calculated from other variables (6 var.),
- Variables without significant mean differences between the countries (Variance Analysis, Multiple Box- and Whisker Plots and Multiple Comparisons) (21 var.),
- Variables which were highly correlated with others (Correlation and Factor Analysis) (11 var.) and
- Variables which were considered not so important by the wine specialists (13 var.).
- Some wines were to be deleted from the data matrix which were outside the quality standards (outliers) (10 wine samples).

In this way the data matrix could be reduced to 33 variables and to 134 wine samples.

On the base of this data set, 96 % of the wines could be classified correctly by a linear discriminant analysis model with 17 variables. The models for red and white wines were different and of less stability.

The goodness of the discriminant model was tested by a new data set of ten wines. Eight of these wines could be classified correctly by the obtained model. Subsequently all wine data were used to fit a new discriminant model with 19 variables and a correct classification rate of 94 %.

Finally we came to a total data set consisting of 144 wines (73 red, 62 white and 9 rosé wines) with 33 variables from 5 countries: Bulgaria (38), Romania (28), Hungary(39), Macedonia (29) and Moldova (10). This set was used to compare different methods of Discriminant Analysis.

"Oversea" wines

The "Oversea" data set consisted of 140 wines from 5 countries with 77 analytical variables.

This set was without missing data. The data set was reduced and transformed in the same way as in the case of East European wines described. The resulted set consisted of 137 wines (89 red, 47 white and 1 rosé) from California (25), South Africa (28), Australia (28), Chile (28), Argentina (28).

This set likewise was used in the comparison of methods for discrimination.

3. Traditional Methods of DA

Here a short overview of older and well known methods working in the space of parameters will be given.

3.1. Linear and Quadratic DA

Linear and Quadratic methods are based on the assumption of Gauss distribution for the observed continuous variables. In [D. Vandev, 2003] a short overview about these methods as well as a description of the program used in our tests can be found.

3.2. Nonparametric DA

If the distributions of the observed varibles are not of Gauss type the densities can be estimated by nonparametric methods.

Indeed, recently much attention has been concentrated on the application of nonparametric methods in classification problems, including methods such as neural networks (Ripley, 1994), classification and regression trees (Breiman et al., 1984), flexible discriminant analysis (Hastie, Tibshirani and Buja (1994)) and multivariate adaptive regression splines (Friedman (1991)).

3.3. Independent Component Discriminant Analysis

[Amato et al., 2002, Alfano et al., 2002] proposed so called ICDA – a nonparametric discriminant analysis method that is a simple generalization of the model assumed by linear and quadratic discriminant analysis. This generalization relies upon a transformation of the data based on independent component analysis (ICA), a statistical method for transforming an observed multivariate vector into components that are stochastically as independent as possible from each other. ICA was proposed in [Hyvärinen, 1997] and an algorithm in [Hyvärinen, 1999].

4. Features space

This section (see [Navarrete and del Solar, 2002]) is focused on the so called features space and methods connected with its use.

4.1. Kernels approach and features space

The set of vectors $\vec{x}_1, ..., \vec{x}_n \in R^n$, (our observations) is mapped into a feature space F by a set of functions $\{\Phi_j(\vec{x}), j = 1, ..., M\}$. One considers such functions that are eigenfunctions of a given kernel (i.e. satisfy the Mercer's condition).

We suppose that $M > p$. That means that the number of parameters should increase. In fact, this is an important purpose of kernel machines in order to give a good generalization ability to the system [Vapnik, 1995].

In the most cases, the dimensionality of the feature space, $M = dim(F)$, is prohibitive for computational purposes, and it could be even infinite. For this reason, any vector or matrix that have at least one of its dimensions equals to M, is said to be uncomputable. Otherwise the vector or matrix is said to be computable.

The aim of kernel machines is to work with a set of mapped vectors: $\Phi(x_i)$. Denote the matrix composed by them $\Phi = \{\Phi(\vec{x}_1), \ldots, \Phi(\vec{x}_n)\}$. Then the correlation matrix of vectors Φ is defined as:

$$R = \frac{1}{n-1}\Phi\Phi'. \tag{1}$$

The Fundamental Correlation Problem (FCP) for the matrix R consists in solving the eigensystem:

$$Rw_k = \lambda_k w_k, \quad ||w_k|| = 1, \quad k = 1, \ldots, M. \tag{2}$$

However, R is an uncomputable matrix and then (2) cannot be solved. In this situation we need to introduce the Dual form of the Fundamental Correlation Problem for R:

$$Kv_k = \lambda_k v_k, \quad ||v_k|| = 1, \quad k = 1, \ldots, n, \tag{3}$$

where K is the so called kernel matrix:

$$K = \frac{1}{n-1}\Phi'\Phi. \tag{4}$$

This can be shown by pre-multiplying (3) by Φ, and using (4).

The kernel function $k(\vec{x}, \vec{x}')$ specifies an inner product in the feature space

$$\Phi(\vec{x}).\Phi(\vec{x})' = k(\vec{x}, \vec{x}').$$

If we want to compute the solutions for which $\lambda_k > 0, k = 1, ...,$ we can go further and write the expression:

$$w_k = \frac{1}{\sqrt{\lambda_k(n-1)}}\Phi v_k, \quad k = 1, \ldots, q. \tag{5}$$

It can be seen that the solution of a general kind of kernel machines can be rewritten in terms of K. That is why this property is called Fundamental Kernel Matrix (FKM).

Dual FCP (2) is also an ill-posed problem that requires some kind of regularization as well. For the same reason the eigenvalues of R will decay gradually to zero, and then we need some criterion in order to determine q. An appropriate criterion is to choose q such that the sum of the unused eigenvalues is less than some fixed percentage (e.g. 5%) of the sum of the entire set (residual mean square error). Then, using (5), the set of primal eigenvectors R $W \in M^{M\times q}$ can be written as:

$$W = \frac{1}{n-1}\Phi V \Lambda^{-1/2}, \tag{6}$$

where the matrix V and the diagonal matrix Λ are q-truncated.

4.2. Support Vector Classification

The support vector machine [Boser et al., 1992, Cortes and Vapnik, 1995], given labelled training data

$$\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{n}, \quad \vec{x}_i \in \vec{X} \subset \mathbf{R}^d, \quad y_i \in \vec{Y} = \{-1, +1\},$$

constructs a maximal margin linear classifier in a high dimensional feature space, $\Phi(\vec{x})$, defined by a positive definite kernel function. A common kernel is the Gauss radial basis function (RBF),

$$k(\vec{x}_1, \vec{x}_2) = e^{-||\vec{x}_1 - \vec{x}_2||^2/2\sigma^2}.$$

The function implemented by a support vector machine is given by

$$f(\vec{x}) = \left\{\sum_{i=1}^{n} \alpha_i y_i k(\vec{x}_i, \vec{x})\right\} - b. \tag{7}$$

That is: if we consider the two classes $I = \{i : y_i = 1\}$ and $\overline{I} = \{i : y_i = -1\}$ the equation (7) may be rewritten as definition of two functions ("densities"):

$$f_I(\vec{x}) \quad = \quad \left\{\sum_{i\in I} \alpha_i k(\vec{x}_i, \vec{x})\right\}$$

$$f_{\overline{I}}(\vec{x}) \quad = \quad \left\{ \sum_{i \in \overline{I}} \alpha_i k(\vec{x}_i, \vec{x}) \right\}.$$

Thus the problem is similar to nonparametric DA. The observation is classified into the class with higher "density".

To find the optimal coefficients $\vec{\alpha}$ of the expansion (7), it is sufficient to maximise the functional

$$W(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(\vec{x}_i, \vec{x}_j) \tag{8}$$

in the non-negative quadrant $0 \leq \alpha_i \leq C, i = 1, \ldots, n$, subject to the constraint $\sum_{i=1}^{n} \alpha_i y_i = 0$. C is a regularisation parameter, controlling a compromise between maximising the margin and minimising the number of training set errors.

Note that in general only a limited number of Lagrange multipliers $\vec{\alpha}$ will have non-zero values; the corresponding input patterns are known as support vectors. Let $\mathcal{I}$ be the sets of indices of patterns corresponding to non-zero Lagrange multipliers,

$$\mathcal{I} = \{i \ : \ 0 < \alpha_i^0 \leq C\}.$$

Equation (7) then can be written as an expansion over support vectors,

$$f(\vec{x}) = \left\{ \sum_{i \in \mathcal{I}} \alpha_i^0 y_i k(\vec{x}_i, \vec{x}) \right\} - b. \tag{9}$$

For a full exposition of the support vector method, see any of the excellent books [Vapnik, 1995, Vapnik, 1998, Cristianini and Shawe-Taylor, 2000].

4.3. Multiclass Strategies in SVM

For the multiclass discriminant analysis problems the two–class solution given above has to be extended.

- One-against-all: The earliest used implementation for SVM multiclass classication is probably the one-against-all method (e.g. [Bottou et al., 1994]). It constructs G SVM models where G is the number of classes. The i-th SVM is trained with all samples in the i-th class with positive labels and all samples with negative labels.

- One-against-one: For each pair of classes i and j a classification model is created. Then for the test sample the class with the largest number of votes wins.

Table 1: Comparision of discriminant analysis models of East European wine data by Interactive Stepwise Discriminant Analysis (Program: ldagui)

	Models for all wines					Models for red wines				Models for white wines			
α	0	.012	.081	.081	.105	0	0	.322	.322	0	0	.058	.058
No. of variables	33	19	17	15	14	11	15	7	12	11	14	7	11
Histamine	x									x	x		
Ethanolamine	x						x			x	x		
Ethylamine	x					x							
Tyramine	x	x	x	x	x								
Phenylethylamine	x	x	x	x	x						x		x
Putrescine	x	x	x	x			x						
Cadaverine	x	x	x	x	x	x	x		x	x	x		
Alkalinity	x					x	x		x				
K	x									x	x	x	x
Na	x	x	x	x	x	x	x	x	x				
Ca	x												
Y	x	x	x	x	x							x	x
Ce	x	x	x	x	x								
Nd	x						x						
Eu	x						x						
Yb	x	x											
Er	x									x	x		
Tb	x									x	x		x
Ethanol (D/H)I	x	x	x	x	x			x					
Ethanol (D/H)II	x	x	x	x	x		x	x	x				
Ethanol σ^{13}C	x				x	x			x				
Wine σ^{18}O	x	x	x	x	x	x	x		x			x	
Ethanol(D/H)I (af)	x					x	x		x	x	x		x
Ethanol(D/H)II (af)	x					x					x	x	x
Ethanol σ^{13}C (af)	x	x	x	x						x	x	x	x
Alcohol	x	x	x	x	x								
Invert sugar	x	x	x	x		x	x			x	x		x
Free SO2	x	x	x		x			x	x				
Total acidity	x	x											x
L-Malic acid	x		x	x			x	x	x		x		
Phosphate	x	x	x		x	x	x	x	x	x	x	x	x
Butanediol	x	x							x	x	x	x	x
Gluconic acid	x	x		x	x	x	x	x	x				
Classification (%)	4.17	0	0	0	0	0	0	0	0	0	0	0	0
Test error (%)	1.9	0.6	0.9	2.24	1.84	3.02	1.06	7.42	1.46	2.28	1.12	3.78	0.72
Leave one out (%)	631	34	25	25	43	56	12	23	15	32	11	43	29

5. Software used

- `LDAgui` for Linear and Quadratic DA (LQDA) , [D. Vandev, 2003]

- Generalised DA (GDA) [Baudat and Anouar, 2000] using PCA in the feature space.

- Support Vector Machine Toolbox (SVM) with renewed QP optimizer [Gunn, 1998]
- LS-SVM Library (LSVM) [Chang and Lin, 2001] with One–To–One strategy for combining outputs of binary classifying.

6. Results and conclusion

We present in Table 1 several models obtained using QLDA implemented in [D. Vandev, 2003]. The parameter α was chosen by hand so that all models have 0 errors when applied to the training data sets. The theoretical error is estimated using large (6000 by group) simulated according the model data sample. The Leave-One-Out (L-O-O) error was obtained as proportion (in %) of all errors to the size of training sample.

All programs were feed with exactly the same training and test data sets. While QLDA has 4-5% – errors over the test set, they achieved a minimum

Table 2: Comparision of discriminant analysis models of "Oversea" wine data by Interactive Stepwise Discriminant Analysis (Program: ldagui)

	Models for all wines			Models for red wines				Models for white wines			
α	0	0	.215	0	0	.684	.684	0	0	.405	.405
No. of variables	11	14	8	8	9	6	8	6	8	5	9
Ethanol (D/H)I	x	x	x	x	x	x	x		x		x
Ethanol (D/H)II								x	x	x	x
Ethanol σ^{13}C	x	x						x	x		x
Wine σ^{18}O										x	x
Histamine	x	x									
Ethylamine				x	x						
Phenylethylamine		x									
K	x	x	x								
Na	x	x	x	x	x	x	x	x	x	x	x
Mg									x	x	x
Y		x									
Alcohol					x	x	x				
Total extract	x	x	x					x	x		x
Invert sugar								x	x		x
Total acid	x	x	x								
Sulfate	x	x	x								
Chloride	x	x	x	x	x	x	x	x	x	x	x
Butanediol	x	x		x	x		x				
Tartaric acid				x	x						
Shikimi acid							x				
Nitrate	x	x		x	x	x	x				
Propanol-1		x	x	x	x	x	x				
Classification (%)	.73	0	0	0	0	0	0	0	0	0	0
Test error (%)	2,32	1,48	1,16	.88	.6	2,02	.88	2,7	1,16	2,88	.56
Leave one out (%)	142	14,6	4	27	2	2	1	8	10	14	6

of 17%. The reason for such unexpectedly bad result may be in the fact that the test sets were generated with a model exactly the same as the model produced by QLDA.

The pure LDA and QDA behave better especially in the case of well balanced second data set. However, the number of parameters needed to achieve 0% classification error is bigger.

REFERENCES

[Alfano et al., 2002] B. Alfano, U. Amato., et al. Segmentation of MR Brain Images through Discriminant Analysis. Tecnical Report 262, Istituto per le Applicazioni del Calcolo "Mauro Picone", 2002.
`http://www.iam.na.cnr.it/rapporti/2002/RT262.pdf`

[Amato et al., 2002] U. Amato, A. Antoniadis, et al. Independent Component Discriminant Analysis. Tecnical Report 254, Istituto per le Applicazioni del Calcolo "Mauro Picone", 2002.
`http://www.iam.na.cnr.it/rapporti/2002/RT254_02.pdf`

[Baudat and Anouar, 2000] G. Baudat, F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, **12** (2000), No 10, 2385–2404.

[Boser et al., 1992] B. Boser, I. Guyon, et al. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on computational learning theory*, ACM, Pittsburgh, 1992, 144–152.

[Bottou et al., 1994] L. Bottou, C. Cortes, et al. Comparison of classifier methods: a case study in handwriting digit recognition. In: *International Conference on Pattern Recognition*, IEEE Computer Society Press, 1994, 77–87.

[Chang and Lin, 2001] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001.
Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, **20** (1995), 1–25.

[Cristianini and Shawe-Taylor, 2000] N. Cristianini and J. Shawe-Taylor.. An Introduction to Support Vector Machines (and other kernel-based learning methods), Cambridge University Press, Cambridge, U.K, 2000.

[D. Vandev, 2003] D. Vandev Interactive Stepwise Discriminant analysis in MATLAB. *Pliska Stud. Math. Bulg.* (Eds D. Vandev, N. Yanev) Proceedings of the Seminar on Statistical Data Analysis 2003. Sozopol, 21–28.06.2003, **16**, 291–298.

[Gunn, 1998] S. Gunn. Version 2.0, Support Vector Classification. Image Speech and Intelligent Systems Group. University of Southampton, 1998. mailto:S.R.Gunn@ecs.soton.ac.uk

[Hyvärinen, 1997] A. Hyvärinen. Independent component analysis by minimization of mutual information. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP97)*, 1997, 3917–3920.

[Hyvärinen, 1999] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10** (1999), 626–634.

[Navarrete and del Solar, 2002] P. Navarrete and J. R. del Solar. Pattern Recognition with Support Vector Machines, vol. **2388** of Lecture Notes, chapter On the Generalization of Kernel Machines, Springer, 2002, 24–39. http://tamarugo.cec.uchile.cl/~jruizd/papers/svm2002.pdf

[Römisch et al., 2001] U. Römisch, D. Vandev, et al. Determination of the geographical origin of wines from east european countries by methods of multivariate data analysis. In: Seminar of the International Biometric Society, Region Oesterreich- Schweiz (ROeS). 24–27 September 2001, Mayrhofen.

[Römisch et al., 2002] U. Römisch, D. Vandev, et al. Determination of the geographical origin of wines from east european countries by methods of multivariate data analysis. In: Talk and Poster at the Intern. Biom. Conf. July 21-26, 2002, Freiburg.

[Vapnik, 1995] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995. ISBN 0-387-94559-8.

[Vapnik, 1998] V. N. Vapnik. Statistical Learning Theory. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley, New York, 1998.

D. L. Vandev
Sofia University "St.Kliment Ohridski"
Sofia, Bulgaria
e-mail: vandev@fmi.uni-sofia.bg

Ute Römisch
Technical University
Berlin, Germany
e-mail: ute.roemisch@tu-berlin.de