# Kernel Discriminant Analysis and information complexity: advanced models for micro-data mining and micro-marketing solutions

C. Liberati & F. Camillo
*Department of Statistics, Università di Bologna, Italy*

## Abstract

In this paper we shall consider Kernel Discriminant Analysis as an innovative tool for supervised classification in a business vision as a marketing solution. The main idea we propose is the combined use of information complexity and bootstrap process which allows the user to overcome the open problems of such a technique as the kernel function choice and at the same time check the robustness of the rule found.
*Keywords: Kernel Discriminant Analysis, information-theoretic complexity measure, bootstrap process, micro-data mining, marketing solution.*

## 1 From data mining to data base marketing

Today, more than in the past, companies understand the value of collecting customer data which try to exploit an intelligent system for extracting interesting information.

The need for a business to get knowledge from data comes from demand to monitor its own clients in order to preserve its relationship with its customers. In fact, the scenario with which a business has to face today is really complex: many customers, many products, many competitors, and little time to react, it means that customer loyalty is a thing of the past so a company needs to reinforce the value of its brand providing specific products projected "around the customers". So it is clear how in such situations the use of Data Mining (DM) in a Knowledge Discovery process (KDD) is dramatically important [1]. The role of DM therefore consists of helping a company to solve vexing issues and to address business processes to reach a good impact in the market. Data mining, on


WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060121

the other hand, extracts information from a database that the user did not know existed. Relationships between variables and customer behaviors that are non-intuitive are the richness of this approach. From this point of view Data Mining and Customer Relational Management (CRM) are inextricably linked. Today a successful marketing strategy must first identify market segments containing customers or prospects with high-profit potential and then build campaign that favorably impact the behavior of these individuals [2]. In this sense is evident how estimating patterns to describe concepts for analyzing association, for building classification and regression models to cluster data represents a fundamental step toward building a productive business marketing. The term *Database marketing* summarizes perfectly this new concept halfway between technology and analysis: such term in fact incorporates the importance to exploring data stored finalized to enrich Customer Table. It supports a variety of business processes and involves a transformation of the data base into business decisions. The integration of theses parts in an framework generates an operative improvement in term of efficiency of return on investment (ROI). From these statements is evident how necessity to get robust rules with high discriminatory power (i.e. low rate of misclassifications) lead us toward new statistical tools alternative as kernel-based methods. We think that the future of data mining should be individuate in this direction.

## 2 Kernel Discriminant Analysis and information complexity

In this paper we shall consider the Kernel Discriminant Analysis (KDA) which extend the idea of Linear Discriminant Analysis (LDA) to nonlinear feature space [3, 4]. The main idea of kernel-based methods is to map the input data to a very high space (Feature Space) by a nonlinear mapping

$$\Phi : \mathcal{R}^d \rightarrow \mathcal{F} \tag{1}$$

where the non-linear function is related to the symmetric positive definite kernel [5, 6]

$$K(x, y) = \Phi(x)\Phi(y). \tag{2}$$

which allows one to obtain a very rich representation of the data. This approach gives rise to a very powerful patterns recognition methods whose overcome the weak points of the LDA improving in a relevant way the results of the classification found.

The specification of the classification problem is the same of the linear DA: assuming we are given the input data set $\mathcal{I}_{XY} = \{(x_1, y_1), ..., (x_n, y_n)\}$ of training vectors $\mathbf{x}_i \in \chi \subseteq \mathcal{R}^d$ and the corresponding values of $y_i \in \mathcal{Y} = \{1, 2\}$ be sets of indices of training vectors belonging to the first $y = 1$ and the second $y = 2$ class, respectively. The class separability in a direction of the weights $\alpha = [\alpha_1, \ldots, \alpha_n]'$ in the *feature space* $\mathcal{F}$ is defined such that the Fisher criteria:

$$J_F(w) = \frac{\alpha' S_B^{\Phi} \alpha}{\alpha' S_W^{\Phi} \alpha}, \tag{3}$$

is maximized, and where $S_B^{\Phi}$, $S_W^{\Phi}$ are respectively the *between and within covariance matrices* in the future space [3, 7]. The kernel discriminant function f(x) of the binary classifier is a linear expansion of the training patterns.

$$f_y(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \tag{4}$$

One of the important advantages of kernel methods, including the KDA, is that the optimal model parameters are given by the solution of a convex optimization problem with a single, global optimum. However, optimal generalization still depends on the selection of a suitable kernel function and the values of regularization and kernel parameters. There are many kernel functions to choose from. The most common kernel functions which we consider in this paper are: with *Gaussian RBF* $(c \in \mathcal{R})$, *Polynomial* $(d \in \mathcal{N}, c \in \mathcal{R})$, *Sigmodial* $(a, b \in \mathcal{R})$, *Cauchy kernel* $(c \in \mathcal{R}_+)$, and *Multi-quadric* $(c \in \mathcal{R}_+)$.

Table 1: Most common kernel function.

| Name of Kernel | $K(\mathbf{x}_i, \mathbf{x}_j) =$ |
|---|---|
| *Gaussian RBF* | $\exp[-\frac{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{c}]$ |
| *Polynomial* | $((x_i.x_j) + c)^d$ |
| *Hyperbolic tangent* or *Sigmoidal* | $\tanh[a(\mathbf{x}_i.\mathbf{x}_j) + b]$ |
| *Cauchy* | $\frac{1}{1+\frac{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{c}}$ |
| *Multi-quadric* | $\sqrt{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2 + c^2}$ |

The choice of the suitable kernel function related with a specific problem is still an open problem. In the literature, presently a valid method for selecting the appropriate kernel function does not exist. Here, we propose to use the information complexity criterion of Bozdogan [8, 10, 11] as our model selection index as well as our criterion for feature variable selection. (In terms of parameters we define the model

$$\theta = (\mu_1, \mu_2, ..., \mu_K, \Sigma, \Sigma, ..., \Sigma, \pi_1, \pi_2, ..., \pi_K) \tag{5}$$

which can be formulated in term of MANOVA model:

$$Y_{gi} = \mu_g + \varepsilon_g \tag{6}$$

with $y_{gi}$ is a $(p \times 1)$ response pattern in the g-th group for the i-th individual, $\mu_g$ is the vector parameter, $\varepsilon_g$ is (i.i.d.) $\mathcal{N}_p(0, \Sigma)$ random error vector. Under p-variate

normal distribution of each group $Y_g \sim \mathcal{N}_p(\mu_g, \Sigma)$ in case of Kernel Discriminant Analysis such index has the following specification [7]:

$$ICOMP(\hat{\Sigma}_W) = np \log 2\pi + n \log |\hat{\Sigma}_W| + np + 2C_1(\hat{\Sigma}_W) \tag{7}$$

where the maximal information-based complexity of a covariance matrix $\hat{\Sigma}$ ($\hat{\Sigma} = \frac{1}{n} S_W^{\Phi}$) is defined by

$$C_1(\hat{\Sigma}) = \frac{p}{2} \ln(\frac{tr(\hat{\Sigma})}{s}) - \frac{1}{2} \ln \left|\hat{\Sigma}\right| \tag{8}$$

The contribution of the complexity of the model covariance structure is that it provides a numerical measure to assess parameter redundancy and stability uniquely all in one measure [8].

In examples coming from literature, however, it suffices just to use and score $C_1(\hat{\Sigma})$ by itself to choose the optimal kernel function in KDA as we'll show in the next sections.

## 3 Numerical results on Ripley data

In this section we illustrate our results using the Regularized binary KDA on the Ripley's (1994) two dimensional toy data of 1250 observations. Since the matrix $S_W$ is at most of rank $n-1$, it is not strictly positive and numerical problems can cause the matrix $S_W$ not even to be positive semi-definite. Therefore, we regularize it by adding a penalty function $\mu I$ to overcome the numerical problem caused by singular within-group covariance $S_W$. In this case, the criterion maximized is the follow:

$$J_F(w) = \frac{\alpha' S_B^{\Phi} \alpha}{\alpha'(S_W^{\Phi} + \mu I)\alpha} \tag{9}$$

(Our proposal consists of a method composed by two steps:

- usage of all information (whole data set) for computing complexity measure and making kernel selection
- use of bootstrap method on sub-samples for testing the robustness of the rule found.

We trained the classifier on a training data composed by 250 units obtained with a proportional stratified sampling. We obtained the training error and, employing a bootstrap process the confidence interval of such rate. (The bootstrap applied consists 100 replications of the KDA on 100 different samples.) Our results are based on a routine which computes KDA on alternative kernel functions with different ridge values included in $\mu \in [0.000001, 0.1]$ interval. Moreover, as there is not clear indication in literature about the specification of a best parameter (this method have been employed for all kernel distribution except Polynomial: we considered for it the firsts 5 degrees $d = 1, .., 5$) for the kernel distributions we

used the mean of the *Mahalanobis distances* among the subjects projected in the input space.

$$d_{ij}^2 = (x_i - x_j)S^{-1}(x_i - x_j)' \tag{10}$$

We found (working on the whole data set) that ceteris paribus ICOMP reaches the minimum value in correspondence of low level of regularization so we set $\mu = 0.000001$ and we score the index for choosing the best parameter for each kernel function. After this choice we select the best kernel via ICOMP.

We experimented our results by retaining all the singular values of $\hat{\Sigma}$ and scored the information-theoretic complexity $C_1(\hat{\Sigma})$ for each of the alternative kernel functions. The results from this experiment are summarized in Table 2 above. Looking such table we see that Cauchy kernel seems to be the better choice based on the minimum value of the complexity measure $C_1(\hat{\Sigma})$. In this work we such measure only for making selection of the kernel so it has not be related with the error rate of the rule found which only depend on classifier used. The confidence intervals (confidence interval=$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ with $\alpha = 0.05$) point out how KDA improves the misclassification error rate.

Table 2: Results KDA using different kernel functions.

| kernel | c | ridge | C1 | training error | confidence interval |
|---|---|---|---|---|---|
| CAUCHY | 1.7966457 | 0.000001 | 2.0831667 | 6.88% | 3.74%-10% |
| RBF | 1.7966457 | 0.000001 | 2.6394597 | 8.47% | 5%-11.92% |
| SIGM | 1.7966457 | 0.000001 | 2.7680317 | 4.43% | 1.88%-6.98% |
| MULTIQ | 1.7966457 | 0.000001 | 3.681339 | 8.41% | 4.97%-11.86% |
| POLY | 1 | 0.000001 | 3.700787 | 11.25% | 7.33%-15.16% |

## 4 Business case: results using KDA

The innovation of this work consists of the usage of KDA in an operative context as profiling solution of customers of a business. We considered data coming from one of the major Italian publishing company. We used a sample of 2774 subjects on which has been realized a survey about behaviors which are synthesized with 16 factors coming from correspondence analysis. The response pattern is a variable of 5 different groups. With data mining standard tools we found rules not effective because they present a high misclassification rate (over 50%). For this reason we decide to apply kernel-based algorithm. Therefore via ICOMP we selected best kernel function and regularization term. Results are shown in the Table 3.

This time a Cauchy kernel seems to be best choice based our selection on the minimum value of complexity index. (All the ICOMP values are to be multiply for $10^4$.) We employed a bootstrap process of 100 replications on training sample composed of 400 customers. It is evident how using Kernel methods the percentage

Table 3: Results KDA using different kernel functions.

| kernel | c | ridge | ICOMP | training error | confidence interval |
|---|---|---|---|---|---|
| CAUCHY | 5.386911 | 0.000001 | 1.7079901 | 34.90% | 30.39-39.40% |
| RBF | 5.386911 | 0.000001 | 2.5584114 | 38.94% | 34.33-43.54% |
| MULTIQ | 5.386911 | 0.000001 | 6.8058779 | 43.64% | 38.95-48.32% |
| SIGM | 5.386911 | 0.000001 | 8.2473884 | 17.74% | 14.12-21.35% |
| POLY | 1 | 0.000001 | 13.416055 | 61.87% | 57.27-66.46% |

of misclassification error decreases dramatically. In the column training error where are contained the mean of the misclassification rate obtained in the 100 replications is possible to observe the good performance of the KDA respect what we got with Linear Discriminant Analysis (Polynomial degree 1): the improvement observed is in terms of droop of error rate and shrinkage of confidence interval. The evaluation of how much the rule found via KDA is a good tool represents surely an innovation in the kernel literature, but such study, in an statistical vision, has to consider not only the description but also the prediction performance of the model.

## 5 The prediction in KDA: results and suggestions

The study of kernel-based methods belongs predominantly to the computer science area so it is for this reason that in literature today lacks reference about statistical modelling approach. In data mining vision the evaluation of a classification pattern present two critical aspects:

- specification of a rule found with an algorithm on a sample
- evaluation of the latter with a test data.

Refusing completely methodology as SEMMA (Sample, Explore, Modification, Model e Assess) which is massive employed in business applications because it evaluate the rule employing an unique sample, we propose the combining use of bootstrap technique on both training and test data to check the robustness of the model found. Therefore, this time we used a proportional stratified sampling for training data (400 observations) and a random sampling on the remaining subjects for test data (50 observations). In the Table 4 are shown the results of the bootstrap for 100 replications on training and test data.

Table 4: Results KDA using training and test data.

| kernel | c | training error | conf. interval tr. | test error | conf. interval tst |
|---|---|---|---|---|---|
| CAUCHY | 5.386911 | 34.90% | 30.39-39.40% | 70.02% | 60.75-79.28% |
| RBF | 5.386911 | 38.94% | 34.33-43.54% | 71% | 61.54-79.93% |
| MULTIQ | 5.386911 | 43.64% | 38.95-48.32% | 71% | 61.71-80.08% |
| SIGM | 5.386911 | 17.74% | 14.12-21.35% | 75.20% | 66.46-83.93% |
| POLY | 1 | 61.87% | 57.27-66.46% | 70.55% | 61.33-79.76% |

We have to point out that the improvements obtained with kernel machines are lost when we do prediction. In fact it is evident the increase of the misclassification rate in case of test data. That could happen because the re-allocation of the new subjects is done computing the Mahalanobis distance of such subjects from the centroids of the groups. The new observation is assigned to the group from whose has minimum distance. Such choice of re-allocation always is not suitable because it does not employ the information coming from the variance of the system. To overcome this limit we propose to hybridize the results coming from the analysis. It means that we apply a non parametric method as k Nearest Neighbors on the scores coming from a regularized KDA for the test data: that involves the exploitation of all the variance information contained in the data.

Obviously such way might be inefficient so a simulation study in which new observation projected in the discriminant space is as move away from the centroid of the group that in order to check how much robust is the rule.

Another point has to be developed regards the discriminant functions: as we illustrated in the previous sections such functions are an linear expansion of the training patterns. Therefore the numbers of the variables contained in these rules are equal to the size of the sample. That might affect the assessment of the model found because it not so rare to detect multi-collinearity problems. So using such tool it is important the use of method for sub-selection model especially if we apply them in prediction context. As answer to such issue we propose the use of Genetic Algorithms (GA) with the Information Complexity Criterion as a fitness function.

As we underlie many times such supervised classification technique (like all tools based on kernel-machines) present more than subjective choice to be applied. Our answer to these open problems as best kernel function choice and its relative parameter and the regularization term is Dr. Bozdogan's Information Complexity Index (ICOMP). The derivation of such criteria in the KDA case represents a first objective no time consuming method in literature to select kernel function.

This is an important step of the research because allows the researcher to have under control such powerful tool, empirical evidences in the Application chapter showed it. Obviously the specification of the index in case of equality of covariance matrixes could be restrictive if we analyze real data, so, as a further development we have the derivation of ICOMP in KDA for groups with both different means and covariance matrixes.

## References

[1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, 1996.

[2] Thearling, K., From data mining to database marketing, 1995.

[3] Mika, S., Rätsch, G., Weston, J., Schölkopf, B. & Müller, K.R., Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX*, eds. Y.H. Hu, J. Larsen, E. Wilson & S. Douglas, IEEE, pp. 41–48, 1999.

[4] Baudat, G. & Anouar, F., Generalized discriminant analysis using kernel approach. *Neural Computation*, 2000.
[5] Mercer, J., Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions Royal Society London*, 1909.
[6] Shawe-Taylor, J. & Cristianini, N., *Kernel Methods for Pattern Analysis*. Cambrige, 2004.
[7] Bozdogan, H., Camillo, F. & Liberati, C., On the choice of the kernel function in kernel discriminant analysis using information complexity. *Cladag proceeding 2005*, 2005.
[8] Bozdogan, H., Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, 2000.
[9] Schölkopf, B. & Smola, A.J., *Learining with Kernel*. MIT Press: Cambrige MA, 2002.
[10] Bozdogan, H., Icomp: A new model-selection criterion. *Classification and Related Methods of Data Analysis*, ed. H.H. Bock, Elsevier Science Publishers B. V.: Amsterdam, 1988.
[11] Bozdogan, H., On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics Theory and Methods*, 1990.
[12] Fisher, R., The use of multiple measurements in taxonomic problems. *Annals of Eugenic*, 1936.
[13] Burges, C., A tutorial on support vector machines for pattern recognition. Technical report, Knowledge Discovery and Data Mining, 1998.
[14] Vapnik, V., *Statistical Learning Theory*. Wiley: N.Y., 1998.
[15] Mika, S., *Kernel Fisher Discriminant*. Ph.D. thesis, University of Berlin, 2002.
[16] Friedman, J.H., Regularized discriminant analysis. *Journal of the American Statistical Association*, **84(405)**, pp. 165–175, 1989.
[17] Schölkopf, B., *Support Vector Learning*. R. Oldenbourg Verlag: Munich, 1997.
[18] Gunn, S.R., Support vector machines for classification and regression. Technical report, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 1998.
[19] Schölkopf, B., Burges, C. & Smola, A.J., *Advances in Kernel Methods*. MIT Press: Cambrige, MA, 1999.