Rivista Italiana di Economia Demografia e Statistica Volume LXVII n. 3/4 Luglio-Dicembre 2013

A PROPOSAL FOR A SEMIPARAMETRIC CLASSIFICATION METHOD WITH PRIOR INFORMATION

Luciano Nieddu, Cecilia Vitiello

1. Introduction

The aim of discriminant analysis is to determine a function (classifier) that, on the basis of a set of covariates $\underline{\boldsymbol{x}}$, best predicts a categorical variable y labeling the class c a unit belongs to.

If the data at hand have been previously classified by an expert the problem is known as supervised classification (Watanabe, 1985) and can be further classified as classification with perfect supervisor and with imperfect supervisor (Katre and Krishnan, 1989). Otherwise the problem is referred to as unsupervised classification.

A plethora of methods have been suggested in order to determine the best classifier. According to Wolpert (1996), if the interest is on the generalization performance of a classifier without making any prior assumptions on the data, then no classification algorithm is inherently superior to any other or even to random guessing.

A general taxonomy of classification rules divides the methods into parametric and non-parametric. In between those two categories the semi-parametric approach via finite mixture models is widely used in unsupervised classification but has found its way to supervised classification (Hastie and Tibshirani, 1996). The method we suggest follows the same approach but differs in some relevant aspects, namely: when finite mixtures of multinormal distribution are involved, no constraints are imposed on the covariance structure of each component (resulting in a more flexible method). Moreover we have adopted a weighted likelihood approach where weights express the information given by the expert. This can easily be extended to handle classification with imperfect supervisor.

The outline of the paper is as follows: in Section 2 the proposed method will be introduced and the parameter estimators will be derived. In Section 3 the experimental results will be presented and in Section 4 conclusions will be drawn.

2. Proposal

Let $\underline{y} = \{y_1 \dots y_G\}$ be an observable multinomial random variable with y_{ig}=1 iff unit i $(i = 1,\dots,n)$belongs to group g. We introduce an unobservable multinomial random variable $\underline{z} = \{z_1 \dots z_K\}$ and model $f(\underline{x}_i|y_{ig} = 1)$ as a K components finite mixture of multinormal $\phi(\underline{x}_i; \underline{\mu}_{gk}, \Sigma_{gk})$ with masses $p(z_{ik} = 1), k = 1, \dots K$.

$$f(x_i|y_{ig} = 1) = \sum_k^K p(z_{ik} = 1)\phi\left(\underline{x}_i; \underline{\mu}_{gk}, \Sigma_{gk}\right).$$

We then derive the likelihood in a weighted form as follows:

$$L\left(\underline{\theta}, \underline{x}, \underline{y}\right) = \prod_i^n \sum_k^K \sum_g^G y_{ig}\ p(z_{ik} = 1)p(y_{ig} = 1|z_{ik} = 1)\phi(\underline{x}_i; \underline{\mu}_{gk}, \Sigma_{gk})$$

In order to simplify notation lets write

- $\left(\underline{\mu}_{kg,}\Sigma_{kg}\right) = \theta_{kg}$
- $\phi\left(\underline{x}_i, \underline{\mu}_{kg,}\Sigma_{kg}\right) = f_i(\theta_{kg})$
- $p(y_{ig} = 1|z_{ik} = 1) = \pi_{g|k}$
- $p(z_{ik} = 1) = \pi_k$

Maximum likelihood estimates (MLE) are obtained iteratively and involve three blocks of parameters.

The first refers to the location and scale parameter θ_{kg}, for fixed π_k and $\pi_{g|k}$ which can be obtained as standard MLE solving the following:

$$\begin{cases} \sum_i^n w_{igk} \dfrac{\delta \log f_i\left(\theta_{kg}\right)}{\delta\, \theta_{kg}} = 0 & g = 1, \dots G; k = 1, \dots K \\ w_{igk} = \dfrac{y_{ig}\pi_{g|k}\pi_k f_i(\theta_{kg})}{\sum_k \sum_g y_{ig}\pi_{g|k}\pi_k f_i(\theta_{kg})}, & g = 1, \dots G; k = 1, \dots K; i = 1, \dots n \end{cases}$$

The other two blocks refer respectively to probabilities π_k and $\pi_{g|k}$ and are both constrained problems.

$$\begin{cases} \sum_i^n \dfrac{y_{ig}\pi_{g|k} f_i(\theta_{kg})}{\sum_k \sum_g y_{ig}\pi_{g|k}\pi_k f_i(\theta_{kg})} + \lambda = 0, & k = 1, \dots K \\ \sum_k \pi_k = 1, \end{cases}$$

$$\hat{\pi}_k = \sum_i^n \frac{1}{n} \frac{\sum_g y_{ig} \pi_{g|k} \pi_k f_i(\theta_{kg})}{\sum_k \sum_g y_{ig} \pi_{g|k} \pi_k f_i(\theta_{kg})} \quad k = 1, \dots K$$

And finally, for each *k=1,…K*, further constrained MLE with $\sum_g \pi_{k|g} = 1$.

$$\begin{cases} \sum_i^n \frac{y_{ig} \pi_k f_i(\theta_{kg})}{\sum_k \sum_g y_{ig} \pi_{g|k} \pi_k f_i(\theta_{kg})} + \lambda^k = 0 \; g = 1, \dots G; \\ \sum_g \pi_{k|g} = 1 \end{cases}$$

$$\hat{\pi}_{g|k} = \sum_i^n \frac{1}{nk} \frac{y_{ig} \pi_{g|k} \pi_k f_i(\theta_{kg})}{\sum_k \sum_g y_{ig} \pi_{g|k} \pi_k f_i(\theta_{kg})}$$

which are straightforward extensions of standard mixture model results.

3. Experimental Results

In this section the results of a simulation study will be presented to evaluate the performance of the proposed method when compared with well-known parametric and non-parametric methods. Then the application to a benchmark dataset from the UCI[1] repository will be shown.

3.1. Simulation study

At this stage, to easily visualize the results, only two class data in $\mathbb{R}^2$ have been considered, generating points around centroids whose locations have been randomly selected on planar curves.

Varying number of centroids have been used. For each centroid $n \in \{20; 30; 50\}$ points have been generated adding MVN noise with $\Sigma = \begin{bmatrix} 2 & 0.8 \\ 0.8 & 2 \end{bmatrix}$. Three types of curves have been used: lines, parabolas and cubic curves. For each class all the three configurations have been used, yielding 9 possible configurations.

Table 1 displays the choices for the parameters of the curves bearing the centroids and the range for the *x* variable for class 0 and class 1. The corresponding values for the *y*s of the centroids have been computed from the equations of each curve. The ranges for the *x* variable for the two classes have been chosen as only partially overlapping to guarantee a discriminant power for the variable itself.

[1] http://archive.ics.uci.edu/ml

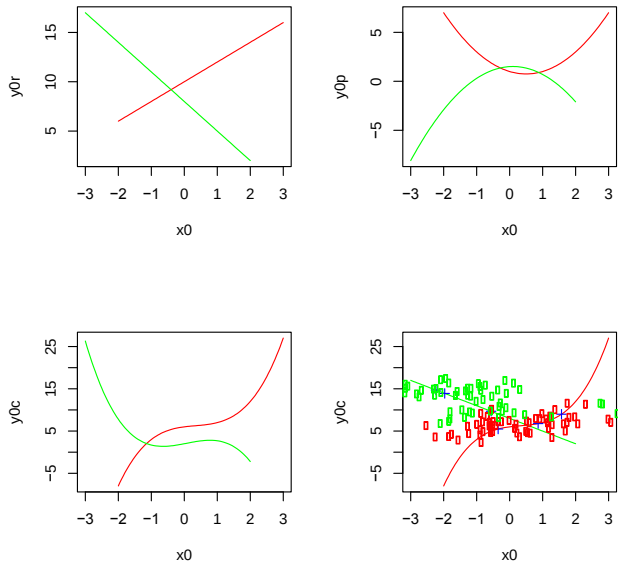
Table 1 – *Simulation setup*

	Class 0	Class 1
Range for x	$[-2;+3]$	$[-3;+2]$
Line	$y = 2x + 10$	$y = -3x + 8$
Parabola	$y = x^2 - x + 1$	$y = x^2 + 0.2x + 1.5$
Cubic	$y = x^3 - x^2 + x + 6$	$y = -x^3 + 0.2x^2 + 1.5x + 2$

In Figure 1 some configurations of curves for class 0 and class 1 and an example of a dataset have been displayed.

The proposed method has been compared with two well-known parametric methods, namely linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), and two non-parametric methods such as *k*-nearest neighbor (*k*=1) (1-NN) and Classification trees (TREE) (Hastie and Tibshirani, 2003).

Figure 1 – *Curves bearing the centroids for each class: 3 of the possible 9 configurations. In the lower right corner an example of dataset with cubic and line*

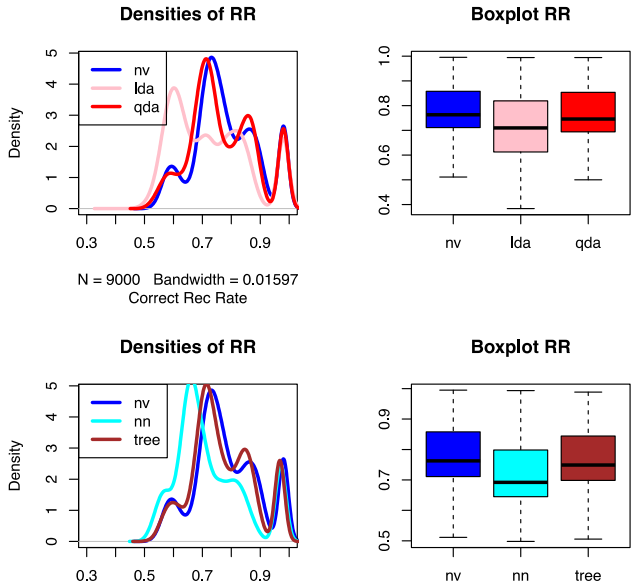


For each configuration of the curves bearing the centroids 100 trials have been carried out and, in each trial, the assessment of the performances of the various techniques have been determined using leave one out cross validation. The simulation has been carried out with the R package using the functions `rpart()`, `lda()`, `qda()` and `knn.cv()` for classification trees, linear discriminant analysis, quadratic discriminant analysis and 1-nearest neighbor respectively.

In a first phase of the simulation study, the number of centroids (nm) and the number of points (n) for each centroid have been allowed to vary. For each configuration of nm and n we have computed the number of times that the proposed method (NV in the following) has bested the others or reached an equal performance. In Figure 2 the total percentages of times that the proposed method has at least performed as well as the others have been displayed (table on the left part of the figure). On the right hand side, the density estimates of the correct recognition rates (RR) for the proposed method and the other classification algorithms, over the whole simulation, have been displayed together with the boxplots of the distributions of the correct recognition rates.

Figure 2 – *Distribution of the percentage of times that the proposed methods performed at least as best as the others (left) and distributions of the correct recognition rates (RR) (densities and boxplots)(right).*

Method	% NV >=
LDA	72.7
QDA	63.0
1-NN	99.3
TREE	81.2



The top part of the figure show the comparison with parametric techniques: LDA seems to be the method with the worst average performance. The performance of NV and QDA are comparable but the distribution of the proposed method is shifted to the right hand side, implying a better average performance (as also shown by the boxplots). This was to be expected since quadratic discriminant analysis has a better adaptive capacity to the data than linear discriminant analysis, which assumes equal covariance matrices for all the classes. The performance of

the proposed method, when compared with non-parametric methods, is displayed on the bottom part of Figure 2. NV works much better than classification trees and almost always bests or performs equally as well as 1-NN (99.3% of cases). This is clearly shown by the density estimate of the 1-NN recognition rate which is the leftmost density in the picture, immediately followed by the classification tree density estimate. Therefore on average the proposed method works at least as well as the other classification methods that have been considered in the simulation. It must be stressed that this results is an overall result over all the 9 configurations of bearing-centroid curves and over various number of centroids (nm=10,25,35) and points per centroid (n=20,30,50).

To determine if there is an effect of the varying number of centroids, a simulation with n=30 points per centroids and varying number of centroids (nm) has been undertaken. Once again the number of times that the proposed method has performed better or as well as the other methods has been considered. In Table 2 such percentages have been displayed together with the number of centroids for each classification algorithm:

Table 2 – *Percentage of the times the proposed method has performed at least as well as the other classification methods over varying number of centroids.*

Method	Number of centroids											
	2	4	6	8	10	12	14	16	18	20	25	35
LDA	83.0	83.4	81.9	81.0	79.0	77.6	76.4	78.9	75.7	75.5	74.4	72.6
QDA	80.4	79.6	74.2	71.8	71.0	69.9	68.2	66.8	67.5	66.5	64.9	62.9
1-NN	93.3	96.8	97.9	98.4	98.2	98.5	98.6	99.4	99.2	99.2	99.4	99.5
TREE	87.1	92.1	87.1	86.2	84.2	81.6	80.3	83.4	80.8	80.8	80.7	81.2

There seems to be a tendency of the proposed method to work at least as well as 1-nn when the number of centroids increases. Such a tendency is not shared by the others methods although the percentage of times the proposed method works as well as classification trees seems to stabilize around 80%. Such percentage seems to show a decreasing trend with the increasing number of centroids for LDA and QDA. This can be ascribed to the fact that with an increasing number of centroids with fixed number of points per centroid and considering the narrow range of the values that the x can assume, the scatter plots for the two classes seem to show a multinormal distribution which is the case where LDA and QDA work best.

3.2. Real data application

In order to test the performance of the proposed method on real data and to compare it to some widely used classification methods, a credit scoring benchmark dataset (Australian credit scoring) has been used (Murphy and Aha, 2001). The

data is composed of 690 records: 307 instances are creditworthy applicants and 383 instances are from not creditworthy applicants. Each instance contains 6 nominal, 8 numeric attributes, and 1 class attribute (accepted or rejected). To protect the confidentiality of data, the attributes names and values have been changed to meaningless symbolic data.

Considering the multinormality hypothesis of the proposed method only the 3 continuous variables out of the 6 numerical have been considered in the study. The performance of the proposed method has been once again compared to nearest-neighbor classifier (1-NN), linear and quadratic discriminant analysis (LDA, QDA) and classification trees (TREE). Leave one out has been used to assess the correct recognition rate. In Table 3 the correct recognition rates have been displayed

Table 3 – *Correct Recognition Rate for the Australian Credit Scoring data: comparing the new proposal to standard methods*

Method	Correct Recognition Rate
Proposed method (NV)	0.716
LDA	0.658
QDA	0.641
1-NN	0.601
TREE	0.674

The proposed method shows a correct recognition rate of 71.6% based only on 3 continuous attributes. The other methods show a recognition rate which is clearly lower, with 1-NN having the worst performance.

4. Conclusions

In this paper an adaptive classification method based on finite mixtures and a-priori information has been presented. The proposed methodology has been tested on a simulation study and on a real benchmark dataset. The proposed approach more than holds its own when compared with parametric and non-parametric well established methods. The proposed methodology assumes the structure of subclasses on the data at hand and the presence of prior information on the classification of the data. The method has been tested in the case of crisp prior information but can be easily extended to the case where there is fuzzy information on the prior classification of the data. Further studies are necessary to test the actual performance on the method also w.r.t. sensitivity and specificity.

Acknowledgements

The authors would like to thank Marco Alfò for the time spent with them discussing alternatives to obvious solutions.

References

WATANABE S. 1985. *Pattern recognition: human and mechanical*, New York: John Wiley & Sons, ISBN 0-471-80815-6

KATRE U. A., KRISHNAN T. 1989. Pattern recognition with an imperfect supervisor. *Pattern Recognition*. 22, 4 (August 1989), 423-431.

MURPHY P.M., AHA D.W. 2001. UCI repository of machine learning databases. Department of Information and Computer Science, University of California Irvine, CA. Available from http://www.ics.uci.edu/mlearn/MLRepository.html

QUINLAN J.R. 1986. Induction of decision trees. *Machine Learning*, 1(1), 81–106.

WOLPERT D.H. 1996. The Lack of A Priori Distinctions between Learning Algorithms, *Neural Computation*, 8, 1341-1390.

HASTIE T., TIBSHIRANI R. 1996. Discriminant analysis by Gaussian mixtures, *Journal of the Royal Statistical Society series B*, Vol. 58 pp. 158-176

HASTIE T., TIBSHIRANI R., FRIEDMAN J.H. 2003. *The Elements of Statistical Learning*. Springer, corrected edition.

SUMMARY

Classification methods are usually grouped into three main categories, ranging from unsupervised classification to supervised classification, passing through classification with imperfect supervisor. Our proposal tries to span a bridge between these two banks. The method proposed has been tested on a simulation study yielding very interesting results.

Luciano NIEDDU, Università degli Studi Internazionali di Roma LUSPIO, l.nieddu@unint.eu

Cecilia VITIELLO, "Sapienza" - Università di Roma - Dipartimento di Statistica, cecilia.vitiello@uniroma1.it