Knowledge-Based Systems 99 (2016) 79–91

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Hybrid kernel density estimation for discriminant analysis with information complexity and genetic algorithm

Seung H. Baek[a,*], Dong-Ho Park[b], Hamparsum Bozdogan[c]

[a] Division of Business Administration, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, South Korea
[b] Korea Institute for Defense Analyses, Seoul, South Korea
[c] Department of Business Analytics & Statistics, University of Tennessee, Knoxville, USA

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 29 January 2016
Accepted 31 January 2016
Available online 10 February 2016

Keywords:
Hybrid kernel density estimation approach
Bandwidth selection
Information theoretic measure of complexity
Genetic algorithm
Model selection

ABSTRACT

A new hybrid approach is proposed which is computationally effective and easy to use in selecting the best subset of predictor variables in discriminant analysis (DA) under the assumption that data sets do not follow the normal distribution. The proposed approach integrates kernel density estimation for discriminant analysis (KDE-DA) and the information theoretic measure of complexity (ICOMP) with the genetic algorithm (GA). The ICOMP plays an important role in finding both the best bandwidth matrix for KDE-DA and the best subset of predictor variables which discriminate between the groups. The genetic algorithm (GA) is introduced and used within KDE-DA as a clever stochastic search algorithm. To show the working of this new and novel approach, six benchmark real data sets are considered and the results are compared with results of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and k-nearest neighbor discriminant analysis (k-NNDA) to choose the best fitting model. The experimental results show that the proposed hybrid kernel density estimation approach outperforms LDA, QDA, and k-NNDA.

1. Introduction

When data conform to the normal distribution for classification purposes with multi-class data, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA) can be used. Both QDA and LDA are popular and show good performance when data are normally distributed. LDA assumes that the population covariance matrices of each group are the same. Based on this assumption, it calculates the posterior probability of group membership of each observation, and assigns an observation to a group where the posterior probability of group membership is the greatest. Thus, LDA performs well in homoscedastic cases. On the other hand, QDA assumes that the population covariance matrices of each group are different. Based on this assumption, it calculates the posterior probability of group membership of each observation, and assigns an observation to a group where the posterior probability of group membership is the greatest. As a result, QDA works well in heteroscedastic cases. However, LDA and QDA have two major drawbacks, as well. One of the drawbacks is due to small sample size within a group when the data is high-dimensional. In this case, when there are not enough samples, the within-class scatter matrix S_w can be singular. Another problem occurs when each group does not follow the Gaussian distribution (see, e.g., [30]). If each group is not Gaussian, both LDA and QDA are not effective in maximizing the correct classification of group membership, or minimizing the probability of misclassification error rate. In many real world problems, it is very unlikely that data conform to the normal distribution. Data on one variable may be skewed while data on another variable may have the approximate lognormal distribution, and so forth. QDA and LDA are not effective in dealing with data which is following a non-normal distribution. There are several approaches to deal with this problem in DA. One of the popular approaches to handle the problem of non-normal distributions is k-nearest neighbor discriminant analysis (k-NNDA). In k-NNDA, the posterior probability of an observation x_i belonging to group k is given by:

$$P(k|x_i) = \frac{\pi_k m_k}{\sum_{k=1}^{K} \pi_k m_k}, \tag{1}$$

where m_k means the number of observation that are in the neighborhood of the x_i that belong to group k, and π_k is the prior probability of group k. Qiu & Wu [30] proposed a new feature extraction method, called a stepwise k-NNDA. k-NNDA does not depend on the non-singularity of the within-class scatter matrix, and it does not assume any particular density function. They found that k-NNDA outperforms existing LDA methods, and it was also

* Corresponding author. Tel.:+82 31 400 5646; fax: +82 31 400 5591.
E-mail address: sbaek4@hanyang.ac.kr (S.H. Baek).

http://dx.doi.org/10.1016/j.knosys.2016.01.046
0950-7051/

very efficient, accurate and robust. However, they did not study whether their new method could find an optimal solution. Another popular approach using nonparametric density estimation is the kernel density estimation for discriminant analysis (KDE-DA). It uses kernel density instead of normal density assumption in calculating class conditional probability distributions. Lin et al. [25] compared kernel based discriminant analysis with LDA to predict advanced, regular, and remedial placement levels. They found that kernel based discriminant analysis performed better than LDA. It is widely known that the performance of a kernel density estimator is primarily determined by the choice of a bandwidth, and only in a minor way by the choice of a kernel function [22]. In the literature, there is not much work done to choose the optimal bandwidth selection for multivariate kernel [22]. This is primarily due to computational difficulty in finding a data adaptive optimal bandwidth matrix. One approach to find the optimal bandwidth matrix is to use cross-validation methods to minimize misclassification rates for different bandwidth matrices. Sain et al. [31] compared the performance of the biased cross validation method, the least-squares cross-validation method, and the bootstrap method for bandwidth selection in multivariate density estimation. They found that the biased cross-validation method performed well compared to the other two methods. However, they also found that the problem of selecting an optimal bandwidth matrix in kernel density estimation grew in complexity with the dimensionality of data. Cross-validation methods sometimes find multiple values of bandwidth to minimize misclassification rates, from which it is difficult to identify the optimal bandwidth [17]. Hyndman et al. [22] proposed using Markov chain Monte Carlo (MCMC) algorithms. They treated the elements of the bandwidth matrix as parameters whose posterior density can be obtained through the likelihood cross-validation criterion. They found that the MCMC algorithm generally performed better than the bivariate plug-in algorithm of Duong & Hazelton [13] and the normal reference rule discussed in Bowman & Azzalini [5]. Yet, they also mentioned that the computational time for higher dimensional data did increase. Increased computational time for high-dimensional data makes its application to discriminant analysis impractical especially. Bensmail & Bozdogan [2] proposed eight forms of the bandwidth matrix, and they used ICOMP [6–9] as a criterion to choose the optimal bandwidth matrix. With that said, our paper focuses on data with non-normal distribution.

The contribution of this paper is the development of the new hybrid approach for KDEDA. For the development, the new ICOMP expression is derived for KDEDA, and the driven ICOMP, a GA, and KDE are combined for KDEDA. ICOMP is used for objective function of a GA, and the GA identifies a model with the minimum ICOMP value as the best model. This approach enables researchers to find both an optimal bandwidth matrix for KDE and the best model from several competing models, which was a severe obstacle for researchers to apply KDE for discriminant analysis.

This paper focuses on data with non-normal distributions. To handle the problem of non-normal distributions, the multivariate Gaussian kernel density estimate will be utilized. Gaussian kernel density estimation is a popular and a well-known non-parametric approach and is computationally simpler and faster. It essentially superposes kernel functions placed at each observations or datum. The Gaussian KDE gives oversmoothed fits with large bandwidths and undersmoothed fits with small bandwidths for data sets. Based on the selection of proper bandwidths, it can be closer to the data sets [5]. In this paper, to choose the optimal bandwidth matrix for the multivariate Gaussian kernel density estimate, eight forms of the bandwidth matrix and ICOMP are used with the genetic algorithm (GA). Further, the purpose of this paper is to apply ICOMP as a model selection criterion in KDE-DA and to develop an alternative approach in DA for dealing with problems of non-normal distributions and high-dimensional data. ICOMP will be used to: (1) find the optimal bandwidth matrix in KDE-DA, and (2) find the best subset of predictor variables which discriminate between the groups using the GA.

The rest of the organization of this paper is as follows. In Section 2, the information complexity (ICOMP) criterion to choose the optimal bandwidth and to select the best model is introduced. Section 3 discusses the Genetic Algorithm (GA) approach to search for the best model. In Section 4, the background of Kernel density estimation and the corresponding bandwidth matrices are described. In Section 5, numerical examples are provided to study the efficiency of the proposed approach with six real benchmark data sets. Additionally, the proposed approach and three existing approaches are compared. This paper is concluded by Section 6.

2. Information theoretic measure of complexity

Model selection and variable selection in discriminant analysis are critical issues. The model selection problem occurs when a researcher needs to choose the best model from several competing potential models. According to Forster (2000) [16], model selection is a bias versus variance trade-off and considers the statistical principle of parsimony in the model selection process. Inference under models with too few variables can be biased, while models with too many variables may provide a poor precision or identification of effects that are, in fact, incorrect. Under the principle of parsimony, researchers prefer a simple model which captures most of the information in the data. Moreover, this simple subset model can reduce computational time in subsequent data analysis and reduce undesirable results such as overfitting problem and multicollinearity. The addition of extra variables usually increases model complexity and positively biased R^2. In discriminant analysis, according to Huberty & Olejnik [21], the increase in the number of variables has different effects compared to multiple regression analysis. Those effects are as follows:

- *First, unlike regression, it may very well happen that as the number of variables (p) is increased, the hit rates (separate-group or total-group) will be decreased. This is particularly true if the variables to be added do not contribute substantially to the intergroup difference.*
- *Second, similar to regression, as the number of variables (p) is increased, the positive bias of the internal hit rates (correct classification) is also increases.*

Thus, it is desirable to find the best subset model to develop a rule to increase classification accuracy. The best model without redundant variables may reduce the misclassification error rate and overcome the overfitting problem. In this section, information theoretic measure of complexity called ICOMP [6–9] as a model selection criterion is introduced. ICOMP is based on the generalization of the covariance complexity index originally introduced by van Emden [34] and was motivated in part by AIC. ICOMP shows better performance than AIC-type criteria, and it has been applied to multivariate non-normal regression models [26], threshold autoregressive models [24], neural networks and support vector machines [27]. To assist readers in understanding ICOMP, the details of ICOMP criteria are described in subsequent subsections.

According to Blahut [[4], p. 250], the joint entropy $H(x) = H(x_1, x_2, \cdots\cdots, x_p)$ with arbitrary mean and covariance matrix Σ is given by:

$$H(x) = H(x_1,\ x_2,\ \ldots,\ x_p) = \frac{p}{2}\log(2\pi) + \frac{p}{2} + \frac{1}{2}\log|\Sigma|$$
$$= \frac{p}{2}[\log(2\pi) + 1] + \frac{1}{2}\log|\Sigma| \qquad (2)$$