Computational Statistics and Data Analysis 117 (2018) 32–44

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Network linear discriminant analysis☆

Wei Cai [a], Guoyu Guan [a,b], Rui Pan [c,*], Xuening Zhu [d], Hansheng Wang [d]

[a] Key Laboratory for Applied Statistics of the MOE, and School of Mathematics and Statistics, Northeast Normal University, Changchun, China
[b] School of Economics, Northeast Normal University, Changchun, China
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[d] Guanghua School of Management, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 15 September 2016
Received in revised form 18 July 2017
Accepted 22 July 2017
Available online 9 August 2017

Keywords:
Classification
Linear discriminant analysis
Misclassification rate
Network data

ABSTRACT

Linear discriminant analysis (LDA) is one of the most popularly used classification methods. With the rapid advance of information technology, network data are becoming increasingly available. A novel method called network linear discriminant analysis (NLDA) is proposed to deal with the classification problem for network data. The NLDA model takes both network information and predictive variables into consideration. Theoretically, the misclassification rate is studied and an upper bound is derived under mild conditions. Furthermore, it is observed that real networks are often sparse in structure. As a result, asymptotic performance of NLDA is also obtained under certain sparsity assumptions. In order to evaluate the finite sample performance of the newly proposed methodology, a number of simulation studies are conducted. Lastly, a real data analysis about Sina Weibo is also presented for illustration purpose.

1. Introduction

Linear discriminant analysis (LDA) is one of the most widely-used classification methods. Due to its simplicity and efficiency, LDA has attracted great attention in a number of fields, such as biomedical studies, face recognition, earth science and many others (Yu and Yang, 2001; Hand, 2006; Guo et al., 2007). To deal with different kinds of data, extensions of LDA widely exist in the past literatures. For instance, functional LDA is proposed where the predictive variables are curves or functions (James and Hastie, 2001). Various penalized LDA methods have been developed for multi-class classification (Witten and Tibshirani, 2011; Clemmensen et al., 2012). Recent studies of LDA mainly concentrate on high dimensional data, including feature screening (Fan and Fan, 2008; Pan et al., 2016) and sparse estimation (Shao et al., 2011).

All aforementioned classification methods are mainly developed under the assumption that all the individuals are mutually independent. With the rapid advance of information technology, relational information among individuals can be easily collected (e.g., friendship, kinship and common interest). As one typical relational information, network data are

☆ The research of Wei Cai and Guoyu Guan is supported in part by National Natural Science Foundation of China (NSFC, 11501093, 11690012), China Postdoctoral Science Foundation Funded Project (Grant No. 2015M581378), and the Fundamental Research Funds for the Central Universities (Grant Nos. 2412015KJ028, 2412017FZ030). The research of Rui Pan is supported in part by National Natural Science Foundation of China (NSFC, 11601539). The research of Xuening Zhu and Hansheng Wang is supported in part by National Natural Science Foundation of China (NSFC, 71532001, 11525101) and Center for Statistical Science at Peking University. The research of all the authors is supported by the Fundamental Research Funds for the Central Universities (Grant Nos. 130028613, 130028729), and National Natural Science Foundation of China (NSFC, 11631003).

* Corresponding author.
E-mail address: panrui_cufe@126.com (R. Pan).

http://dx.doi.org/10.1016/j.csda.2017.07.007
0167-9473/

becoming increasingly available. A network refers to a group of individuals and the corresponding relationships among them. In the past decades, there are abundant literatures on model-based statistical analysis of network data. These models include but are not limited to the ER model (Erdös and Rényi, 1959), the p_1 model (Holland, and Leinhardt, 1981; Wasserman and Pattison, 1996; Robins et al., 2007), the stochastic blockmodel (Holland et al., 1983; Nowicki and Snijders, 2001; Karrer and Newman, 2011), and the latent space model (Hoff et al., 2002; Sewell and Chen, 2015). As one can see, existing statistical classification methods are no longer appropriate for network data since individuals are correlated with each other.

For network data, classification methods firstly arise in the field of machine learning, where *collective classification* (CC) is popularly used (Neville and Jensen, 2000; Taskar et al., 2002; McDowell et al., 2007). The spirit of collective classification is to make use of the information collected from one's neighbors when predicting one particular individual's class label. As a result, network structure can be taken into consideration and the resulting prediction performance can be enhanced. However, collective classification has three main disadvantages. First of all, due to its complex model setup, the results of collective classification are lack of interpretation. Secondly, although there are a number of applications of collective classification, its theoretical properties are not clear. Thirdly, most collective classification methods rely on iterative algorithms which lead to high computational cost. This motivates us to develop a novel statistical classification model for network data.

In this paper, we propose a new methodology called network linear discriminant analysis (NLDA). It makes use of both the predictive variables and network structure. As a result, relational information can be incorporated, which leads to improved prediction performance compared with traditional LDA. Furthermore, certain sparsity assumptions are imposed for large-scale network data. This makes the computation of NLDA feasible when the network size is huge. At the same time, the newly proposed NLDA method possesses excellent theoretical properties in terms of misclassification rate. Under mild assumptions, the method of NLDA outperforms that of LDA for different kinds of network structures. Lastly, the classification ability with information only from network structure is also investigated.

The rest of this article is organized as follows. In Section 2, we introduce the NLDA model and establish its theoretical properties. In Section 3, sparse networks are considered and the corresponding asymptotic results are derived. A number of numerical studies are conducted in Section 4 to demonstrate the finite sample performance of our newly proposed methodology. A real data analysis is also presented for illustration purpose. Some concluding remarks are given in Section 5. All the technical proofs are left in Appendix.

2. Network linear discriminant analysis

2.1. Model and notations

To describe the network structure, define an adjacency matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{n\times n}$, where $a_{i_1 i_2} = 1$ if the i_1th node follows the i_2th node, and $a_{i_1 i_2} = 0$ otherwise. We follow the tradition and let $a_{ii} = 0$ for $1 \le i \le n$. In addition, let (Y_i, X_i) be the observation collected from the ith node, where $Y_i \in \{0, 1\}$ is the binary class label with $P(Y_i = k) = \pi_k$, and $\pi_0 + \pi_1 = 1$. Furthermore, $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor. Given $Y_i = k \in \{0, 1\}$, X_i is assumed to follow a p-dimensional multivariate normal distribution with mean $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p\times p}$. For convenience, we write $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ and $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$. Theoretically, the adjacency matrix is assumed to be random and generated according to some probability distribution. Specifically, we assume that conditional on $\mathbb{Y}$ and $\mathbb{X}$, different edges (i.e., $a_{i_1 i_2}$ s) are mutually independent with

$$P(a_{i_1 i_2} = 1|\mathbb{Y}, \mathbb{X}) = P(a_{i_1 i_2} = 1|Y_{i_1} = k_1, Y_{i_2} = k_2) = \omega_{k_1 k_2}, \tag{1}$$

where $\omega_{k_1 k_2} \in (0, 1)$ is the link probability from class k_1 to class k_2. In addition, let $\omega = (\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top \in \mathbb{R}^4$.

Traditional LDA predicts the class label of node i by maximizing the posterior probability $P(Y_i = k|X_i)$. It can be easily proved that this probability is proportional to $\pi_k \exp(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i)$. However, given the network structure A, we are able to employ not only the nodal information from node i but also the information from its connected nodes. Then, the corresponding prediction problem becomes maximizing $P(Y_i|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$, where $\mathbb{Y}_{(-i)} = (Y_{i'} : i' \ne i)^\top \in \mathbb{R}^{n-1}$. By assuming (1), i.e., the network structure A and the nodal covariates $\mathbb{X}$ are conditionally independent given the class labels $\mathbb{Y}$, we have

$$\begin{aligned} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A) &= P(Y_i = k)P(X_i|Y_i = k)P(A|\mathbb{Y}_{(-i)}, Y_i = k) \propto \pi_k \exp\left(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i\right) \\ &\times \prod_{j\ne i}\prod_{l}\left\{(\omega_{lk})^{a_{ji}}(1-\omega_{lk})^{1-a_{ji}}(\omega_{kl})^{a_{ij}}(1-\omega_{kl})^{1-a_{ij}}\right\}^{I(Y_j=l)}, \end{aligned} \tag{2}$$

where some constants independent of k are ignored and $I(\cdot)$ is the indicator function. We denote the optimal prediction of Y_i as $Y_i^* = \arg\max_{k\in\{0,1\}} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$. Regarding (2), we have the following three remarks.

Remark 1. Note that the right hand side of (2) consists of two components. One is in proportion to the posterior probability $P(Y_i = k|X_i)$, i.e., $\pi_k \exp(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i)$. If the network information is ignored, this is just the result derived from the traditional LDA. The other component is due to the network information.