Short Communication
Volume 12:8, 2021

Journal of Biometrics and Biostatistics

ISSN: 2155-6180 Open Access

Multi Classification to Gene Expression Data with Some Complex Features

Jessica Landis*

Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Description

Classification problem is one of important topics in statistical analysis. The main purpose of classification is to classify subjects to their classes. A dataset whose response contains more than two classes, called multiclass response, is frequently considered in the past literature, and it is so-called multi classification. In the perspective of linear model, Agresti shows comprehensive discussions in the logistic regression with multiclass response. In the perspective of machine learning theory, there are several methods which have been proposed, including support vector machine (SVM), k-nearest neighbor (KNN), linear discriminant analysis (LDA), and so on. General discussions can be found in Hastie et al. [1]. Thanks to the modern technology, we can easily collect the high-dimensional data with complex features incorporated. Therefore, in the era of big data, the high-dimensional data is inevitable to encounter and it always attracts our attention. In this paper, we mainly focus on the multi classification with the following features: I. Network structure. II. Ultrahigh-dimension. III. Measurement error. IV. Multi classification with ordinal response In fact, it is expected that the conventional classification methods may not either completely solve problems or ignore those complex features mentioned above in the developments of methods. A motivated example comes from cancer classification with tumor gene expression signatures. The dataset is collected by Ramaswamy et al. [2]. Basically, this dataset contains 14 common human cancer classes and 16,063 gene expression values. The sample size in this dataset is 218. The main target of this study is to correctly classify 218 patients to 14 cancer classes by treating gene expression values as predictors. To show the classification, Ramaswamy et al. presented SVM with One vs. All (OVA) approach [3]. However, some important features listed above may not be fully considered. As a result, in the following presentation, we briefly outline key ideas to analyze this dataset. Opinion Network Structure In the gene expression data, it is expected that there exists the (pairwise) dependence structure within genes. In order to detect the pairwise dependence structure, incorporate the network and improve the prediction, we need to implement the technique in graphical model theory. Actually, the idea of implementing graphical model theory has been discussed in machine learning theory. For example, Huttenhower developed nearest neighbor networks for gene expression data. Zhu et al. proposed SVM with network structure [4]. Cai et al. discussed network linear discriminate analysis. However, those methods basically assume a common network structure for predictors of all subjects without taking into account of possible heterogeneity for different classes. That is, it is expected the network structures in different classes should be different. In addition, the most existing methods with incorporation of network structure mainly focus on the binary response. Hence, it is important to invest the multi classification with incorporation of heterogeneous network structures. Ultrahigh-dimension in our motivated example, one of the challenges is the ultrahigh-dimension in predictors. It is also well-known problem, i.e., Even though Ramaswamy et al. [3] has proposed valid procedure, there is also a large improvement. Specifically, since not all genes are relevant to the response, then a nature idea is to remove those predictors which are irrelevant to the response before developing methods or analyzing the data. The Sure Independent Screening (SIS) proposed by Fan and Lv is one of powerful methods to achieve this target. Measurement error the other important feature is the measurement error in predictors. Since the gene expression values are usually obtained by medical measurement, so it may be possible to produce the error. That is, the observed gene expression values may not exactly equal to the true gene expression values of patients. In addition, Carroll et al. pointed out that the wrong estimation or conclusion may be produced if the error effect is ignored in the analysis. Therefore, to produce a precise classification and prediction, it is necessary to carefully analyze the measurement error in predictors. Multiclassification with Ordinal Response Finally, we notice that many existing works mainly focus on the multiclass response which is free of order (i.e., nominal). In the motivated dataset, it may be interesting to consider 14 cancers with ordering. The criterion of ordering can be the rank of severe cancer or proportion among patients (e.g., lung cancer may have higher proportion or may be more severe than breast cancer). To the best of our knowledge, there is no contribution in machine learning theory which considers such setting. Conclusion In this paper, we present several interesting and important extensions to the multi classification with gene expression data. The idea in Opinion section is the author's current research work, and the remaining sections are the author's research topics in the near future.

***Corresponding Author:** Jessica Landis, Department of Molecular Biology, Princeton University, Princeton, NJ, USA, Tel: +16181734382; E-mail: landisjess@gmail.com*

Received: Aug 02, 2021; **Accepted**: Aug 16, 2021; **Published**: Aug 23, 2021

References

1. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, et al.(2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98: 15149-15154.
2. Huttenhower C, Flamholz AI, Landis JN, Sahi S, Myers CL, et al.(2007) Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. BMC Bioinformatics 8: 1-13.
3. Cai W, Guan G, Pan R, Zhu X, Wang H (2018) Network linear discriminant analysis. Comput Stat Data Anal 117: 32-44.
4. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Statist Soc B 70: 849-911.

How to cite this article: Landis, Jessica. "Multi Classification to Gene Expression Data with Some Complex Features ." *J Biom Biostat* 12 (2021) : e008.