Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper PUBLISHERS key to the Researchers

Opinion
Volume 9 Issue 1 - December 2018
DOI: 10.19080/BBOAJ.2018.09.555751

Biostat Biometrics Open Acc J
Copyright © All rights are by Li-Pang Chen

Multiclassification to Gene Expression Data with Some Complex Features

Li-Pang Chen*

Department of Statistics and Actuarial Science, University of Waterloo, Canada

Submission: December 10, 2018; **Published:** December 20, 2018

***Corresponding author:** Li-Pang Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Abstract

Classification is usually an important topic which mainly classifies subjects to their classes. Many methods, including machine learning theory, have been fully developed. In the era of big data, however, we may encounter the more complex dataset, and the conventional methods may either fail to solve problems or ignore some key perspectives. In this paper, we mainly focus on gene expression data and present some features which may appear in the dataset. In addition, we also outline some research directions and discuss possible solutions to deal with these problems.

Keywords: Classification; Machine learning; Network structure; Ultrahigh-dimension

Abbreviations: SVM: Support Vector Machine; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; OVA: One vs. All; SIS: Sure Independent Screening

Introduction

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 *multiclassification*. In the perspective of linear model, Agresti [1] 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. [2].

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 multiclassification with the following features:

I. Network structure.

II. Ultrahigh-dimension.

III. Measurement error.

IV. Multiclassification 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. [3]. 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. [3] presented SVM with One vs. All (OVA) approach. 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 have been discussed in machine learning theory. For example, Huttenhower [4] developed nearest neighbor networks for gene expression data. Zhu et al. [5] proposed SVM with network structure. Cai et al. [6] discussed network linear discriminant 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 multiclassification 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 $p \gg n$ problem, i.e.,$16063 \gg 218$ 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 [7] 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. [8] 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 multiclassification 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.

References

1. Agresti A (2012) Categorical Data Analysis. Wiley, New York, USA.
2. Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, USA.
3. 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(26): 15149-15154.
4. 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: 50.
5. Zhu Y, Shen X, Pan W (2009) Network-based support vector machine for classification of microarray samples. BMC Bioinformatics 10(Suppl 1): S21.
6. Cai W, Guan G, Pan R, Zhu X, Wang H (2018) Network linear discriminant analysis. Computational Statistics and Data Analysis, 117: 32-44.
7. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Statist Soc B 70: 849-911.
8. Carroll RJ, Ruppert D, Stefanski L A, Crainiceanu CM (2006) Measurement Error in Nonlinear Model. CRC Press, New York, USA.

DOI: 10.19080/BBOAJ.2018.09.555751

How to cite this article: Li-Pang Chen. Multiclassification to Gene Expression Data with Some Complex Features. Biostat Biometrics Open Acc J. 2018; 9(1): 555751. DOI: 10.19080/BBOAJ.2018.09.555751.