Statistics Seminar

Department of Mathematical Sciences

DATE:	Thursday, October 8, 2015
TIME:	**1:00pm to 2:30pm** (note the different time and longer duration)
LOCATION:	WH 100E
SPEAKER:	Qiyi Lu, Binghamton University
TITLE:	Learning Partially Labeled Data in the High-dimensional, Low-sample Size Setting

Abstract

High-Dimensional, Low-Sample Size (HDLSS) is a very challenging data setting in statistics and statistical learning, including regression, classification, clustering, etc. The HDLSS data appear in many applied areas such as gene expression micro-array analysis, facial recognition, medical image analysis and text classification. In the context of classification, in many real applications, it is costly to manually place the class labels on observations; as a consequence, often only a small portion of labeled data is available while a large number of observations are left without labels. Such partially labeled data are very difficult to analyze in the HDLSS setting. In this dissertation, we study the HDLSS partially labeled data in two aspects. We push forward the frontier of knowledge by creating a new classification method and a significance analysis tool for the partially labeled data.

Classification is an important tool with many useful applications. Among the many existing classification methods, Fisher's Linear Discriminant Analysis (LDA) is a traditional model-based approach which makes use of the distributional information such as the covariance of the features. However, in the HDLSS setting, LDA cannot be directly deployed because the sample covariance is not invertible. While there are modern methods designed to deal with the high dimensionality, it is difficult to obtain good performance for the partially labeled data when the analysis is based on the labeled data alone, due to the scarcity of the data. In order to overcome the difficulty, and to fully make use of the seemingly useless unlabeled data, we propose a semi-supervised sparse LDA classifier in this dissertation. Our method combines LDA, a method-based approach, with some machine learning oriented components. The extra components help to extract useful information from the unlabeled data which can boost the classification performance in some situations.

Before learning a data set, a natural question to ask is whether the predefined classes are really different from one another (in the context of classification), or whether clusters are really there (in the context of clustering). Such a question may be answered by significance tests. Even in the challenging HDLSS setting, there have been some recent developments. However, a significance analysis tool for the partially labeled data has not been developed in the HDLSS setting. In this dissertation, we propose a significance analysis approach for the HDLSS partially labeled data. Our method makes use of the whole data and tries to test the class difference as if all the labels were observed. Compared to a testing method that ignores the label information, our method provides a greater power, meanwhile, maintaining the size.

In studying both aspects of the partially labeled data, we provide theoretical justifications to the methods proposed. In particular, our theoretical study has emphasized on the HDLSS setting, shedding light on the usefulness of the proposed methods. Lastly, comprehensive

simulation and data examples have illustrated the effectiveness of the methods.

From:
https://www2.math.binghamton.edu/ - **Department of Mathematics and Statistics, Binghamton University**

Permanent link:
https://www2.math.binghamton.edu/p/seminars/stat/10082015

Last update: **2015/09/06 03:29**