Journal of Statistical Planning and Inference 176 (2016) 78–94

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Significance analysis of high-dimensional, low-sample size partially labeled data

Qiyi Lu[*], Xingye Qiao

Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY, 13902-6000, United States

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form 17 February 2016
Accepted 19 March 2016
Available online 29 March 2016

Keywords:
Classification
Clustering
High-dimensional, low-sample size data
Hypothesis test
Semi-supervised learning

ABSTRACT

Classification and clustering are both important topics in statistical learning. A natural question herein is whether predefined classes are really different from one another, or whether clusters are really there. Specifically, we may be interested in knowing whether the two classes defined by some class labels (when they are provided), or the two clusters tagged by a clustering algorithm (where class labels are not provided), are from the same underlying distribution. Although both are challenging questions for the high-dimensional, low-sample size data, there has been some recent development for both. However, when it is costly to manually place labels on observations, it is often that only a small portion of the class labels is available. In this article, we propose a significance analysis method for such type of data, namely 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, illustrated by a comprehensive simulation study. Theoretical properties of the proposed method are studied with emphasis on the high-dimensional, low-sample size setting. Our simulated examples help to understand when and how the information extracted from the labeled data can be effective. A real data example further illustrates the usefulness of the proposed method.

1. Introduction

Classification and clustering are both important tools in statistical learning. The availability of the class labels distinguishes these two main domains. In classification, class labels are provided prior to the analysis, while they are unavailable in the clustering analysis. A natural statistical question regarding their use is whether classes/clusters are really there. In a setting where binary class labels are observed, we may be interested in testing whether the two classes are from the same distribution. Though often neglected, this is an important step before applying a classification algorithm. In standard statistical textbooks, there are many significance tests, such as two-sample t-test, one-way ANOVA, Hotelling's T^2 test, and MANOVA. Among these, the two-sample t-test and ANOVA are univariate tests. The Hotelling's T^2 test and MANOVA are multivariate tests, though both can fail when the dimension d is much greater than the sample size n.

This problem of testing the difference between two classes becomes even more challenging for the high-dimensional, low-sample size (HDLSS) data. The Hotelling's T^2 test is very powerful when the dimension is smaller than the sample size. It is invariant under linear transformation. In addition, under the null hypothesis, the distribution of the statistic

[*] Correspondence to: Binghamton University, 4400 Vestal Pkwy E, Binghamton, NY 13902, USA. Tel.: +1 607 777 2147; fax: +1 607 777 2450.
E-mail addresses: qlu@math.binghamton.edu (Q. Lu), qiao@math.binghamton.edu (X. Qiao).

http://dx.doi.org/10.1016/j.jspi.2016.03.002
0378-3758/© 2016 Elsevier B.V. All rights reserved.

is known. However, the Hotelling's T^2 statistic cannot be computed in the HDLSS setting because the sample covariance matrix is not invertible. There are efforts attempting to overcome this issue, including Dempster (1960), Bai and Saranadasa (1996), Srivastava and Du (2008) and Chen and Qin (2010). These methods use diagonalized versions of the covariance or inverse covariance matrices in the Hotelling's T^2 statistic. There are many other treatments, such as Srivastava and Fujikoshi (2006), Schott (2007) and Srivastava (2007), which calibrate the distribution of some proposed statistic. In addition, the Direction–Projection–Permutation (DiProPerm) test (Wichers et al., 2007; Wei et al., forthcoming) has been proved to be very effective for testing the class difference of the HDLSS data.

Besides the difficulty brought from the high dimensionality, in many real problems, it is often the case that there are many observations with no class labels (the so-called unlabeled data portion). One reason is that it is often difficult or expensive to obtain the class label information, while it may be relatively cheap to obtain the covariate information even for many observations. In such a situation, those aforementioned testing methods which require label information cannot be applied to the whole data set to test the class difference. As a consequence, one may have to forfeit the potentially useful information that resides in the unlabeled data. For instance, many cancer patients are categorized to certain cancer subtypes by radiologists through an inspection of the medical images. However, because of the high health care cost, medical images are easier to obtain than the actual diagnostic. Before a classification algorithm is used to design a data-mining-based early-screening machine (see, for example, Land et al., 2012; Schaffer et al., 2012), a valuable question is whether the so-called subtypes, many of which may be ad hoc or based on experience, are really there.

One possible, but clearly flawed, solution to this problem is to treat all the data as unlabeled. In the unsupervised context, in the sense that there are no class labels provided for the analysis, clustering algorithms have been broadly applied in many fields. As to determine whether clusters are really there, several methods have been developed to assess the significance of clusters, including McShane et al. (2002), Tibshirani and Walther (2005), Suzuki and Shimodaira (2006) and Liu et al. (2008). However, these methods are not directly applicable for partially labeled data, unless one forfeits the potentially useful information that resides in the class labels of the labeled data portion.

Hence, there seems to be a dilemma in testing partially labeled data: to ignore the unlabeled data completely (and apply a significance test for the labeled data only), versus, to ignore the class labels in the labeled data portion (and test the significance of clustering). Although each has its own applicability domain, neither looks promising for us. This motivates us to devise a significance testing method for the HDLSS partially labeled data. When class labels are partially provided, the unlabeled data are used to better estimate the sampling distribution. In the meantime, the class labels help to effectively distinguish the two classes even if their difference is small. Our proposed method is named Significance Analysis of HDLSS Partially Labeled Data (SigPal).

To illustrate our main idea, we show a toy example in Fig. 1. The data are generated from a mixture of two Gaussian distributions with a small difference in the mean, $0.5N(-\boldsymbol{\mu}, \mathbf{I}_2) + 0.5N(\boldsymbol{\mu}, \mathbf{I}_2)$, where $\boldsymbol{\mu} = (0.5, 0)'$ and each component of the mixture distribution corresponds to one class. To ease the presentation, the example is two-dimensional, though the message applies to the HDLSS setting. We show two significance analysis methods that inspire our approach, the DiProPerm test of Wichers et al. (2007) and Wei et al. (forthcoming), and the Statistical Significance of Clustering method (SigClust) of Liu et al. (2008) and Huang et al. (2014). The DiProPerm test is applicable only if all the class labels were known (see the different colors/marker-types in the top-left panel.) In contrast, SigClust does not require a label (see the bottom-left panel.) The (empirical) p-value of the DiProPerm test turns out to be 0, which leads to a correct conclusion that the two classes are indeed from two distributions, whereas the SigClust method fails to find this important difference (p-value 0.11). However, when we consider the partially labeled data setting shown in the two right panels, the good performance of DiProPerm may not be reproduced because the labeled information is incomplete. In this case, SigPal can still give a significant conclusion with p-value 0.014. When we apply DiProPerm to the labeled subset only, the result (p-value 0.055) is not as assertive as the SigPal method (see the top-right panel). All these three methods will be introduced or revisited in the next two sections.

The rest of the article is organized as follows. In Section 2, we review the DiProPerm test and the SigClust test. Section 3 presents our proposed SigPal method. Some theoretical results are studied in Section 4 which emphasize the HDLSS setting. A comprehensive simulation study and real data case study are provided in Section 5. Section 6 gives some concluding remarks. Appendix is devoted to technical proofs.

2. DiProPerm test and SigClust test

In this section, we review two significance analysis methods, DiProPerm and SigClust. Both methods are designed for testing HDLSS data, although they may be applied to low-dimensional data as well.

2.1. DiProPerm test

In practice, permutation tests are often used for the purpose of testing the class difference, where the null distribution is mimicked by the empirical distribution of the statistic calculated from many randomly permuted data sets. However, for high-dimensional data, some distance measure with direct permutation may not work. This is because when $d \gg n$, such distance measure will be mainly driven by the error aggregated over dimensions, rather than the true mean difference between classes. To address this issue, a three-step procedure called Direction-Projection-Permutation test (DiProPerm)