Journal of Statistical Planning and Inference 147 (2014) 66–83

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Effects of unlabeled data on classification error in normal discriminant analysis

Keiji Takai [a,*], Kenichi Hayashi [b]

[a] Faculty of Commerce, Kansai University, Japan
[b] Graduate School of Medicine, Osaka University, Japan

ARTICLE INFO

Article history:
Received 14 July 2012
Received in revised form
21 June 2013
Accepted 13 November 2013
Available online 23 November 2013

Keywords:
Partially labeled data
Unlabeled data
Missing data
Asymptotic relative efficiency
Normal discrimination
Nonnormal data
Semi-supervised learning

ABSTRACT

Semi-supervised learning, i.e., the estimation of parameters based on both labeled and unlabeled data, is widely believed to be effective in constructing a boundary in classification problems. The present paper investigates whether this belief is true in the case of normal discrimination in terms of the classification error for normal and nonnormal data. For this investigation, we use the framework of missing-data analysis because data consisting of labeled and unlabeled individuals can be regarded as missing data. Based on this framework, we introduce two labeling mechanisms: feature-independent labeling and feature-dependent labeling. For each of these labeling mechanisms, we analytically derive the asymptotic relative efficiency based on the labeled data alone and based on both the labeled and unlabeled data. Numerical computations reveal that (i) under the feature-independent labeling mechanism, unlabeled data tend to contribute to the improvement of the classification error even for nonnormal data and (ii) under the feature-dependent labeling mechanism, unlabeled data from both normal and nonnormal distributions are helpful when the labeled data are informative, but unlabeled data can augment the classification error when the labeled data are not informative. Finally, we describe some future areas of research.

1. Introduction

Making a classification rule based on unlabeled data in addition to labeled data is referred to as semi-supervised learning in the context of machine learning. Semi-supervised learning has been widely used because this type of learning is intuitively believed to be able to produce better results than when unlabeled data are not used. This belief is supported by the well-accepted fact that the amount of information for classification increases as the sample size increases. However, is this belief true? The goal of the present paper is to provide an answer to this simple question. As a starting point, we restrict our discussion to normal discriminant analysis.

Suppose that an individual $\boldsymbol{x}$ is independently taken from either one of the two distinct populations Π_1 and Π_0 with probabilities $\pi_1(>0)$ and $\pi_0(=1-\pi_1)$, respectively. The two distributions are p-variate normal distributions with different means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ and with the same covariance matrix Σ. In linear discriminant analysis, an individual $\boldsymbol{x}$ is assigned to a

* Corresponding author.
E-mail address: takai@kansai-u.ac.jp (K. Takai).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.11.004

population by the Fisher linear discriminant function:

$$L(\boldsymbol{x}) = [\beta_0,\ \boldsymbol{\beta}_1'] \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix},$$

where

$$\beta_0 = \log \frac{\pi_1}{\pi_0} - .5(\boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0) \quad \text{and} \quad \boldsymbol{\beta}_1 = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{1}$$

An individual is assigned to Π_1 if $L(\boldsymbol{x}) > 0$ and to Π_0 if $L(\boldsymbol{x}) \leq 0$. The same coefficient $\boldsymbol{\beta}_1$ can be derived without specifying a distribution of variables. The derivation is formulated by several equivalent optimization problems. An example of an optimization problem is to find a weight vector on $\boldsymbol{x}$ that maximizes the ratio of the between-class variance to the within-class variance. This derivation is used in a number of papers on the semi-supervised learning problem (Cai et al., 2007; Sugiyama et al., 2010). However, we use the normal distribution model for deriving the linear discriminant function because such a distribution accommodates the stochastic structures of the labeled and unlabeled data. In order to practically apply linear discriminant analysis to real data, the unknown parameters necessary to construct $L(\boldsymbol{x})$ must be estimated by the maximum likelihood (ML) method from data of n individuals.

The ML method usually assumes that all of the individuals are labeled as belonging to either of the populations. However, in practical discriminant analysis, for various reasons, there are cases in which not all of the individuals are labeled. Consider the example from Airoldi et al. (1995) of a geneticist interested in differentiating between identical and fraternal twins. Out of 100 pairs of female twins whose physical features were measured, due to cost constraints, only 20 pairs are randomly selected and diagnosed as identical or fraternal. Consider next the case of medical screening (McLachlan and Scot, 1995). Patients with a feature vector below or above a certain threshold are further investigated in order to determine whether the patients are truly sick or are healthy, while ethics prevents further investigation of the other patients unless they choose it. In both these examples, not all of the individuals are labeled.

These two examples raise two important points concerning discriminant analysis: (i) how to choose the individuals to be labeled and (ii) how to use such labeled data (and the unlabeled data, if possible). The first example deals with the case in which the labeled individuals are randomly chosen without depending on the values of the feature vector (the external characteristics), whereas the second example deals with the case in which the labeled individuals are chosen depending on the values of the feature vector. For both cases, the classification boundary for normal discriminant analysis can be constructed either with the labeled data alone or with the unlabeled data in addition to the labeled data (partially labeled data).

A natural question then arises. Do the unlabeled data always improve discrimination for any labeling method? This question has been studied extensively, but attention has been focused on the situation in which the labeled individuals are chosen without depending on the values of a feature vector (feature-independent labeling). O'Neill (1978) compared the errors between partially labeled data and data with all of the individuals labeled, using the results of Efron (1975). Castelli and Cover (1995, 1996) showed that labeled observations are more valuable than unlabeled observations in terms of the classification error. Zhang and Oles (2000) investigated the effect of unlabeled data on the precision in the maximum likelihood estimation. Rigollet (2007) studied the effect of unlabeled data under "the cluster assumption." Under the same

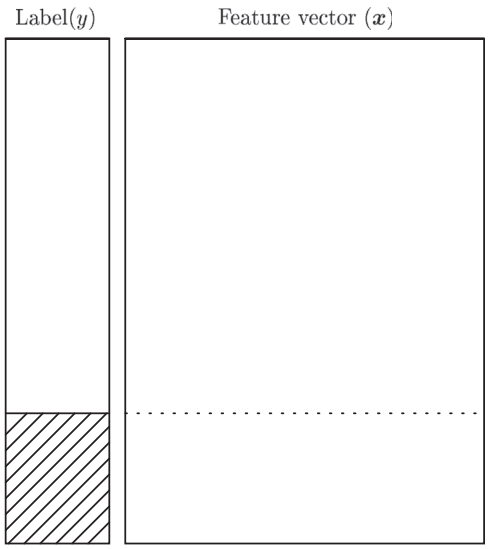

Fig. 1. Partially labeled data for classification. The shaded box in y indicates no label. Labeled data: the feature vector data above the dotted line and their labels. Unlabeled data: only the feature vector data below the dotted line with no label.