Computational Statistics and Data Analysis 92 (2015) 115–125

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Matrix completion discriminant analysis

Tong Tong Wu [a,*], Kenneth Lange [b,c,d]

[a] Department of Biostatistics and Computational Biology, University of Rochester, NY 14642, United States
[b] Department of Biomathematics, University of California, Los Angeles, CA 90095, United States
[c] Department of Human Genetics, University of California, Los Angeles, CA 90095, United States
[d] Department of Statistics, University of California, Los Angeles, CA 90095, United States

ARTICLE INFO

Article history:
Received 12 February 2014
Received in revised form 1 June 2015
Accepted 15 June 2015
Available online 2 July 2015

Keywords:
Classification
Missing observations
MM algorithm
Semi-supervised learning
Singular value decomposition

ABSTRACT

Matrix completion discriminant analysis (MCDA) is designed for semi-supervised learning where the rate of missingness is high and predictors vastly outnumber cases. MCDA operates by mapping class labels to the vertices of a regular simplex. With c classes, these vertices are arranged on the surface of the unit sphere in $c - 1$ dimensional Euclidean space. Because all pairs of vertices are equidistant, the classes are treated symmetrically. To assign unlabeled cases to classes, the data is entered into a large matrix (cases along rows and predictors along columns) that is augmented by vertex coordinates stored in the last $c - 1$ columns. Once the matrix is constructed, its missing entries can be filled in by matrix completion. To carry out matrix completion, one minimizes a sum of squares plus a nuclear norm penalty. The simplest solution invokes an MM algorithm and singular value decomposition. Choice of the penalty tuning constant can be achieved by cross validation on randomly withheld case labels. Once the matrix is completed, an unlabeled case is assigned to the class vertex closest to the point deposited in its last $c - 1$ columns. A variety of examples drawn from the statistical literature demonstrate that MCDA is competitive on traditional problems and outperforms alternatives on large-scale problems.

1. Introduction

Whenever large data sets are collected, missing responses and missing predictors occur. Missing data values are now classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Statisticians have devised a host of methods for coping with missing data, including: listwise deletion, pairwise deletion, hot deck imputation, mean substitution, regression substitution, and multiple imputation. In model fitting in general, and regression in particular, the modern tendency is to impute missing values by maximum likelihood estimates or posterior means under a Gaussian model (Little and Rubin, 2002; Schafer, 2010). Regardless of the method of imputation, the consensus among statisticians is that one should use all available information. Failures to impute data can increase bias and compromise inference.

For data presented in matrix form, a new imputation method is now available. Matrix completion aims to recover a full matrix – usually of low rank – from a subset of observed entries. During the past five years, the subject of matrix completion has captured the attention of researchers from a variety of backgrounds in statistics, applied mathematics, and computer science. Candès and Recht (2009) prove that a low-rank matrix can be almost perfectly recovered when the number of observed entries exceeds a certain level. Fortunately, their strong uniformity condition on sampled entries can be relaxed

[*] Corresponding author.
E-mail addresses: tongtong_wu@urmc.rochester.edu (T.T. Wu), klange@ucla.edu (K. Lange).

http://dx.doi.org/10.1016/j.csda.2015.06.006
0167-9473/© 2015 Elsevier B.V. All rights reserved.

(Recht, 2011). Matrix completion can be accomplished by several algorithms: subspace evolution and transfer (SET) (Dai and Milenkovic, 2009), gradient algorithms applied to the primal and dual problems (Lin et al., 2009), singular value thresholding (SVT) (Cai et al., 2010; Hu et al., 2012), fixed point and Bregman iteration (Ma et al., 2011), the alternating direction method (Chen et al., 2012; Yuan et al., 2009), modified fixed point continuation (Ma and Zhi, 2011), and alternating minimization (Jain et al., 2012).

Methods for handling missing data in discriminant analysis have lagged methods for model fitting. Most research has focused on classification trees and the nature of missingness, for example, whether missingness occurs in the testing data (Saar-Tsechansky and Provost, 2007), in the training data, or in a combination of both (Ding and Simonoff, 2010). Other relevant distinctions include missing responses in both training and testing (Ding and Simonoff, 2010) and MCAR (Feelders, 1999; Kim and Yates, 2003; Zhang et al., 2005) versus a combination of MCAR, MAR, and MNAR (Kalousis and Hilario, 2000; Twala, 2009). Farhangfar et al. (2008) evaluated a variety of imputation methods on classifiers for discrete data, while Sun et al. (2009) studied the impact of imputation on classification accuracy with DNA microarray data. A common theme among these and other papers (Acuna and Rodriguez, 2004; Luengo et al., 2012; Garcia-Laencina et al., 2010) is the deterioration of classification accuracy as the rate of missingness increases. Standard imputation methods are ill equipped to handle high rates of missingness. Two existing papers advocate matrix completion to handle missing data in transductive learning (Cabral et al., 2011; Goldberg et al., 2010). However, these papers deal with multi-label classification rather than multi-category classification and employ a fixed point continuation algorithm rather than an MM algorithm.

Our new combination of matrix completion with Vertex Discriminant Analysis (VDA) (Lange and Wu, 2008; Wu and Lange, 2010; Wu and Wu, 2012) extends VDA into the realm of semi-supervised learning. VDA is geometrically motivated and attuned to matrix representation. Missing observations are ubiquitous in practice, and discarding cases with missing predictors leads to less accurate classification. Instead of conducting imputation and classification sequentially, we believe that classification can inform imputation and advocate conducting them simultaneously. Our new method, which we call Matrix Completion Discriminant Analysis (MCDA), is specifically designed for data exhibiting high rates of missingness and an excess of features over cases. It is precisely in this setting of high-dimensional sparse data that matrix completion is expected to shine. Our cancer classification results validate this intuition and justify the inclusion of incomplete cases and the strategy of simultaneous imputation and classification.

2. Matrix completion discriminant analysis

2.1. Vertex discriminant analysis

VDA is a novel supervised classification method (Lange and Wu, 2008; Wu and Lange, 2010; Wu and Wu, 2012). In classification with c classes,it operates by mapping the classes to the c vertices of a regular simplex in the Euclidean space $\mathbb{R}^{c-1}$. For example in binary classification, the two classes correspond to the numbers -1 and 1 on the real line. In trinary classification, the three classes correspond to the three vertices of an equilateral triangle in the plane. The advantages of mapping categories to vertices include dimension reduction, simplification of computation, ease of interpretation, and enhancement of geometric intuition. It is impossible to situate more than c equidistant points in $\mathbb{R}^{c-1}$ (Lange and Wu, 2008).

There are several versions of VDA: VDA$_\text{R}$ (Lange and Wu, 2008), VDA$_\text{LE}$ (Wu and Lange, 2010), and VDA$_\text{K}$ (Wu and Wu, 2012), where the subscripts stand for Ridge, Lasso and Euclidean, and Kernel, respectively. The first two are linear classifiers, and the third is a nonlinear classifier. The linear VDA classifiers rely on the linear regression model $\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$, $i = 1, \dots, n$, to predict the vertex associated with case i. Here $\boldsymbol{x}_i$ is a p-dimensional predictor vector for case i, $\boldsymbol{A} = (a_{jk})$ is a $(c-1)\times p$ matrix of slopes, and $\boldsymbol{b} = (b_j)$ is a $c-1$ column vector of intercepts. The linear VDA classifiers minimized the objective function

$$R(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{b}) + \lambda P(\boldsymbol{A}), \tag{1}$$

where $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_{2,\epsilon} = \max\{\|\boldsymbol{z}\|_2 - \epsilon, 0\}$ denotes ϵ-insensitive Euclidean distance, $\boldsymbol{\theta}$ signifies the parameters $(\boldsymbol{A}, \boldsymbol{b})$, and $P(\boldsymbol{A})$ denotes the penalty imposed on $\boldsymbol{A}$. In VDA$_\text{R}$,

$$P(\boldsymbol{A}) = \sum_{j=1}^{c-1}\sum_{k=1}^{p} a_{jk}^2$$

is a ridge penalty, while in VDA$_\text{LE}$,

$$P(\boldsymbol{A}) = \sum_{j=1}^{c-1}\sum_{k=1}^{p} |a_{jk}| + \sum_{j=1}^{c-1}\sqrt{\sum_{k=1}^{p} a_{jk}^2}$$

is a mixture of lasso and Euclidean penalties. Lasso and Euclidean penalties promote spareness in the estimate of the slope matrix $\boldsymbol{A}$.

Nonlinear VDA$_\text{K}$ exploits reproducing kernel Hilbert spaces (RKHS). Every such Hilbert space of functions $\mathcal{H}_K$ is generated by a kernel function $K(\cdot,\cdot)$. If we fix a point $\boldsymbol{y}$, then the function $\boldsymbol{x} \mapsto K(\boldsymbol{x}, \boldsymbol{y})$ belongs to $\mathcal{H}_K$. In the current setting $\boldsymbol{x}$ is the