Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2007) 452–470

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

High-dimensional pseudo-logistic regression and classification with applications to gene expression data

Chunming Zhang*, Haoda Fu, Yuan Jiang, Tao Yu

Department of Statistics, 1300 University Avenue, University of Wisconsin, Madison, WI 53706, USA

Available online 28 December 2006

Abstract

High dimension low sample size data, like the microarray gene expression levels, pose numerous challenges to conventional statistical methods. In the particular case of binary classification, some classification methods, such as the support vector machine (SVM), can efficiently deal with high-dimensional predictors, but lacks the accuracy in estimating the probability of membership of a class. In contrast, the traditional logistic regression (TLR) effectively estimates the probability of class membership for data with low-dimensional inputs, but does not handle high-dimensional cases. The study bridges the gap between SVM and TLR by their loss functions. Based on the proposed new loss function, a pseudo-logistic regression and classification approach which simultaneously combines the strengths of both SVM and TLR is also proposed. Simulation evaluations and real data applications demonstrate that for low-dimensional data, the proposed method produces regression estimates comparable to those of TLR and penalized logistic regression, and that for high-dimensional data, the new method possesses higher classification accuracy than SVM and, in the meanwhile, enjoys enhanced computational convergence and stability.

Keywords: Bayes optimal rule; Large p and small n data; Logistic regression; Loss function; Support vector machine

1. Introduction

Technological invention and information advancement have revolutionized scientific research and technological development. Many sophisticated large-scale data sets have recently been collected. These new data sets and streams pose numerous challenges to conventional statistical or data mining methods due to not only the massive size, but also the high dimensionality.

In this paper, we focus on high dimension low sample size data, the so-called large p small n data, with binary class label responses. Notable examples include clinical assessment of tumor types for microarray gene expression data, in which the number of variables (genes) far exceeds the number of samples (arrays). The traditional logistic regression (TLR) method effectively estimates the probability of class membership for large n small p data, but does not handle data sets with high-dimensional predictors. Besides, a monotone likelihood problem will occur when the predictors are fully separable (Firth, 1993). In that case, logistic regression will give unreliable estimates. See Albert and Anderson (1984) and Santner and Duffy (1986) for details.

* Corresponding author. Tel.: +1 608 262 0084; fax: +1 608 262 0032.
E-mail addresses: cmzhang@stat.wisc.edu (C. Zhang), fuhaoda@stat.wisc.edu (H. Fu), jiangy@stat.wisc.edu (Y. Jiang), yutao@stat.wisc.edu (T. Yu).

0167-9473/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.csda.2006.12.033

On the other hand, the support vector machine (SVM) has emerged as a powerful pattern classification tool for high-dimensional data. By means of the dual representation, SVM translates an optimization problem of p-variables into the counterpart of n-variables. This characteristic enables SVM to efficiently deal with high-dimensional predictors. Refer to Vapnik (1996) and Cristianini and Shawe-Taylor (2000), among many others, for details. Nonetheless, unlike the logistic regression, SVM lacks the accuracy in estimating the probability of membership for each class. Therefore, SVM is less appropriate to estimate the class probability, which is of significant importance in various scientific disciplines.

In this paper, we aim to develop a high-dimensional regression and classification method which simultaneously combines the strengths of both SVM and TLR. To achieve this goal, we bridge the gap between SVM and TLR by their loss functions. Based on our proposed new loss function, we further propose a pseudo-logistic regression (PsLR) and classification approach which integrates the classification ability of SVM and the regression capability of TLR. Simulation evaluations and real data applications demonstrate that for low-dimensional data, the proposed method produces regression estimates comparable to those of TLR and penalized logistic regression (PeLR) (Eilers et al., 2001), and that for high-dimensional data, the new method possesses higher classification accuracy than SVM and, in the meanwhile, enjoys enhanced computational convergence and stability. As will be discussed in Section 3.2, the PeLR when applied to high-dimensional data, reduces the size of the estimating equations, but could not genuinely resolve the problems of computational instability and solution non-uniqueness. In contrast, our proposed method effectively overcomes these problems.

This paper is organized as follows. In Section 2, we review TLR and SVM, and connect them by their loss functions. In Section 3, we propose the PsLR method. In Section 4, we present some property of PsLR and propose a bias correction procedure for PsLR estimates. We apply our method to simulated data in Section 5 and real data sets in Section 6. Section 7 concludes this paper by a discussion. All detailed derivations are postponed to the Appendix.

2. Logistic regression and SVM

In this section, we start by reviewing TLR and SVM. After that, we will connect these two methods by their loss functions, which motivate the proposed PsLR method.

2.1. Logistic regression

Let $Y \in \{0, 1\}$ indicate the class label of a sample and $\mathrm{X} = (X_1, \dots, X_p)^{\mathrm{T}}$ be the vector of explanatory variables. Define the conditional mean response function by $m(\mathrm{x}) = P(Y = 1|\mathrm{X} = \mathrm{x})$ and the canonical parameter by $\theta(\mathrm{x}) = \ln[m(\mathrm{x})/\{1 - m(\mathrm{x})\}]$. In TLR, it is assumed that

$$\theta(\mathrm{x}) = \beta_0 + \mathrm{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2.1}$$

where β_0 and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ are unknown parameters.

For independent samples $\{(\mathrm{x}_i, y_i)\}_{i=1}^n$ drawn from (X, Y), the maximum likelihood estimates of β_0 and $\boldsymbol{\beta}$ are obtained from minimizing the negative conditional log-likelihood function

$$\begin{aligned}\ell(\widetilde{\boldsymbol{\beta}}) &= -\sum_{i=1}^{n}[y_i \ln\{m(\mathrm{x}_i)\} + (1 - y_i)\ln\{1 - m(\mathrm{x}_i)\}] \\ &= -\sum_{i=1}^{n}[y_i \widetilde{\mathrm{x}}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \ln\{1 + \exp(\widetilde{\mathrm{x}}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})\}],\end{aligned} \tag{2.2}$$

where $\widetilde{\mathrm{x}}_i = (1, \mathrm{x}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. For computational implementation, it is customary to use the Newton–Raphson algorithm which requires the score vector

$$\frac{\partial \ell(\widetilde{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}}} = -\sum_{i=1}^{n}\widetilde{\mathrm{x}}_i\{y_i - m(\mathrm{x}_i)\},$$