Computational Statistics and Data Analysis 78 (2014) 119–134

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Efficient classification for longitudinal data

Xianlong Wang [a,*], Annie Qu [b]

[a] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA*

[b] *Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form 24 March 2014
Accepted 10 April 2014
Available online 24 April 2014

Keywords:
QIFC
Classification
Linear discriminant analysis
Longitudinal data analysis
Quadratic inference function
Quadratic distance

ABSTRACT

A new classifier, *QIFC*, is proposed based on the quadratic inference function for longitudinal data. Our approach builds a classifier by taking advantage of modeling information between the longitudinal responses and covariates for each class, and assigns a new subject to the class with the shortest newly defined distance to the subject. For finite sample applications, this enables one to overcome the difficulty in estimating covariance matrices while still incorporating correlation into the classifier. The proposed classifier only requires the first moment condition of the model distribution, and hence is able to handle both continuous and discrete responses. Simulation studies show that *QIFC* outperforms competing classifiers, such as the functional data classifier, support vector machine, logistic regression, linear discriminant analysis, the naive Bayes classifier and the decision tree in various practical settings. Two time-course gene expression data sets are used to assess the performance of *QIFC* in applications.

1. Introduction

In many longitudinal biomedical experiments, such as the gene expression microarray studies on yeast cells (Spellman et al., 1998; Eisen et al., 1998) and fruit flies (Arbeitman et al., 2002; Ma et al., 2006), the gene expressions of thousands of genes are repeatedly measured over multiple time-points. These genes are assumed to be associated with a set of pre-defined biological functions, and it is of scientific interest to identify which genes are associated with which biological functions. A classifier for longitudinal data is called for to address such a problem. In addition, the sample sizes for most longitudinal studies are small to moderate due to the cost and complexity of the longitudinal design. Hence, a desirable longitudinal classifier should also work effectively for finite sample applications. As high throughput technologies become increasingly cost-effective, longitudinal studies will be conducted in more research fields, and more features or covariates will be collected at each time point. Therefore, there is an emerging demand for longitudinal classification tools to mine such high-dimensional longitudinal data.

Classifications for single point data are well developed, but these methods might not be effective for classifying longitudinal data. For longitudinal data, Choi (1972) proposes a mixed model; Bagui and Mehra (1999) develop a multi-stage nearest neighbor classification rule; Brown et al. (2000) apply support vector machine (SVM); Liang and Kelemen (2005) propose regularized neural networks; Lee (2004), Rossi and Villa (2005, 2006) and Park et al. (2008) apply the functional SVMs; Müller (2005) uses functional principal component scores; Leng and Müller (2006) use logistic regression; De la Cruz-Mesía et al. (2007) apply semiparametric Bayesian classification based on dependent Dirichlet processes; and Schmah et al. (2010) compare several classification methods for longitudinal fMRI studies and identify the adaptive quadratic discriminant

[*] Corresponding author. Tel.: +1 2066675768.
E-mail addresses: xwan2@fhcrc.org (X. Wang), anniequ@illinois.edu (A. Qu).

http://dx.doi.org/10.1016/j.csda.2014.04.008
0167-9473/© 2014 Elsevier B.V. All rights reserved.

function and the support vector machine as the best classifiers. Functional data classifiers (Febrero-Bande and Oviedo de la Fuente, 2012) are also applicable to most longitudinal data.

We propose a new classification method, *QIFC*, for longitudinal data based on the quadratic inference function (QIF) which builds a semi-parametric model. Our approach builds a classifier by taking advantage of modeling information between responses and covariates of the subjects within each class, and assigns a new subject to the class with the shortest newly defined distance to the subject. Our approach overcomes the difficulty in estimating covariance matrices as in linear discriminant analysis (LDA) while still being able to incorporate into the classifier the correlation among multiple observations on the same subject. We use simulation to compare *QIFC* to commonly used classifiers including the functional data classifier, SVM, logistic regression, linear discriminant analysis, the naive Bayes classifier and the decision tree. The proposed classifier shows advantages for both continuous and discrete response data for various settings. We also provide asymptotic optimality theory for *QIFC*. Applications to time-course gene expression data indicate that the generalization error of *QIFC* is improved compared to other classifiers when the sample sizes are small to moderate.

The paper is organized as follows. We describe *QIFC* in Section 2, and provide the theoretical results in Section 3. Simulation studies and applications follow in Sections 4 and 5, respectively. Section 6 summarizes our results and provides a brief discussion.

2. *QIFC*

For longitudinal data, let $y_i(t)$ be a response variable and $x_i(t)$ be a $p \times 1$ vector of covariates, measured at time t, $t = t_1, \ldots, t_q$ for subject i, $i = 1, \ldots, N$. We assume that the model satisfies the first moment model assumption

$$\mu_i(t_j) = E\{y_i(t_j)\} = \mu\{x_i(t_j)'\beta\}, \tag{1}$$

where $\mu(\cdot)$ is a known inverse link function and β is a p-dimensional parameter vector. The quasi-likelihood equation (Wedderburn, 1974) for longitudinal data is

$$\sum_{i=1}^{N} \dot{\mu}_i' V_i^{-1}(y_i - \mu_i) = 0,$$

where $V_i = \mathrm{Var}(y_i)$, $y_i = (y_i(t_1), \ldots, y_i(t_q))'$, $\mu_i = (\mu_{it_1}, \ldots, \mu_{it_q})'$, and $\dot{\mu}_i = \partial\mu_i/\partial\beta$. In practice, V_i is often unknown, and the empirical estimator of V_i based on sample variance could be unreliable, especially when the sample size is small relative to the number of variance components in V_i. Liang and Zeger (1986) introduce generalized estimating equations to substitute V_i by assuming $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal marginal variance matrix and R is a common working correlation matrix, which only involves a small number of nuisance parameters. The advantage of the GEE approach is that the GEE estimator of the regression parameter is consistent, even if the working correlation R is misspecified. However, the GEE estimator is not efficient within the same class of estimating functions when R is misspecified.

Qu et al. (2000) introduced the quadratic inference function by assuming that the inverse of the working correlation can be approximated by a linear combination of several basis matrices, that is,

$$R^{-1} \approx a_1 M_1 + \cdots + a_m M_m,$$

where M_i's are symmetric matrices. We observe that the generalized estimating equation is an approximate linear combination of the components in the estimating functions,

$$\bar{g}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) = \frac{1}{N}\begin{pmatrix} \displaystyle\sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \displaystyle\sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-1/2} M_m A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}. \tag{2}$$

Hence, the advantage of this approach is that it does not require estimation of linear coefficients a_i's which can be viewed as nuisance parameters.

Since the dimension of (2) is larger than the number of parameters, we cannot set each component in (2) to be zero to solve for β. Instead we estimate β by setting $\bar{g}_N$ as close to zero as possible, in the sense of minimizing the quadratic function,

$$\hat{\beta} = \arg\min_{\beta} \bar{g}_N' \Omega^{-1} \bar{g}_N,$$

where $\Omega = \mathrm{Var}(g_i)$. In practice, Ω is often unknown, but can be estimated consistently by $\bar{W}_N = N^{-1}\sum_{i=1}^{N} g_i g_i'$. The quadratic function,

$$Q_N(\beta) = N \bar{g}_N' \bar{W}_N^{-1} \bar{g}_N, \tag{3}$$