Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1643–1655
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data

Yongdai Kim[a,*], Sunghoon Kwon[a], Seuck Heun Song[b]

[a] *Seoul National University, Korea*
[b] *Korea University, Korea*

Received 22 March 2006; received in revised form 23 May 2006; accepted 5 June 2006
Available online 30 June 2006

Abstract

Monitoring gene expression profiles is a novel approach to cancer diagnosis. Several studies have showed that the sparse logistic regression is a useful classification method for gene expression data. Not only does it give a sparse solution with high accuracy, it provides the user with explicit probabilities of classification apart from the class information. However, its optimal extension to more than two classes is not obvious. In this paper, we propose a multiclass extension of sparse logistic regression. Analysis of five publicly available gene expression data sets shows that the proposed method outperforms the standard multinomial logistic model in prediction accuracy as well as gene selectivity.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Classification; Gene expression data; Multinomial logit model; One-against-all; Sparse logistic regression

1. Introduction

Constructing a classification rule for tissue samples based on gene expression profiles has received much attention recently due to emerging microarray technology. A new challenge is that the number of genes (i.e. the dimension of inputs) is much larger than the number of tissue samples, in which case standard classification methods either are not applicable or perform badly. Also, identifying a small subset of informative genes, called marker genes, which discriminate types of tumors or tumor versus normal tissues, has become an important subject. Hence, good learning algorithms with gene expression data should provide a classification rule which not only yields high accuracy but also has the ability to identify marker genes. In related literature, Guyon et al. (2002) proposed a recursive feature elimination technique with support vector machines, Li et al. (2002) introduced two Bayesian approaches with the technique of automatic relevance determination, and Shevade and Keerthi (2003) and Roth (2002) applied the sparse logistic regression, to name just a few.

* Corresponding author.
E-mail address: ydkim@stats.snu.ac.kr (Y. Kim).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.007

Among these tools, sparse logistic regression is a useful classification method for gene expression data. It gives a sparse solution with high accuracy and also it provides the user with explicit probabilities of classification apart from the class information. However, its optimal extension to more than two classes is not obvious. A standard multiclass extension of sparse logistic regression might be sparse multinomial logistic (SML) regression (Krishnapuram et al., 2004), which is a sparse version of the multinomial logit model—a popular multiclass formulation in statistics (see, for example, Agresti, 1990). SML, however, has a problem in gene selection. Since the estimates of the regression coefficients depend on the choice of the baseline class (see Section 2 for definition), and so do the selected genes. Hence, some important genes are dropped in the final model, which in turn degrades the prediction accuracies. Empirical results in Section 4 confirms this observation.

In this paper, we propose a new multiclass extension of sparse logistic regression called *sparse one-against-all logistic* (*SOVAL*) *regression*, whose main idea is to reduce a multiclass problem to multiple binary problems and to construct a classifier using the reduced multiple binary problems simultaneously. By analyzing five real data sets of gene expressions, we show that SOVAL outperforms SML in prediction accuracy as well as gene selectivity.

The paper is organized as follows. In Section 2, SOVAL as well as SML are presented. A computational algorithm based on the gradient LASSO algorithm of Kim et al. (2005) is given in Section 3. Results of numerical experiments are presented in Section 4 and concluding remarks follow in Section 5.

2. Models

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be input–output pairs of a given data set where $\mathbf{x}_i \in R^p$ is a gene expression level and $y_i \in \{1, 2, \ldots, J\}$ is a type of cancer of the ith tissue sample. Here, n is the number of tissues, p the number of genes and J the number of classes (i.e. tumor types). We first present SML and then propose SOVAL.

2.1. SML regression

SML starts with the multinomial logit model

$$\Pr(y_i = j|\mathbf{x}_i) = \frac{\exp\left(f_j(\mathbf{x}_i)\right)}{\sum_{m=1}^{J} \exp(f_m(\mathbf{x}_i))}$$

for $j = 1, \ldots, J$ where

$$f_j(\mathbf{x}_i) = \beta_0^{(j)} + \beta_1^{(j)} x_{i1} + \cdots + \beta_p^{(j)} x_{ip}.$$

For identifiability, we let $\beta_k^{(J)} = 0$ for $k = 0, 1, \ldots, p$.

Let $\boldsymbol{\beta}_0 = \left(\beta_0^{(1)}, \ldots, \beta_0^{(J-1)}\right)$, $\boldsymbol{\beta}_j = \left(\beta_1^{(j)}, \ldots, \beta_p^{(j)}\right)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1})$. For the sparse model, we estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ by maximizing the log-likelihood

$$\mathscr{L}_1(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{J} I(y_i = j) f_j(\mathbf{x}_i) - \log\left(\sum_{m=1}^{J} \exp(f_m(\mathbf{x}_i)) \right) \right] \tag{1}$$

with the constraint $\sum_{j=1}^{J-1} \sum_{k=1}^{p} \left|\beta_k^{(j)}\right| \leqslant \lambda$. Here, $\lambda > 0$ is a regularization parameter, which should be selected in advance using cross validation or any other method.

Once the regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are estimated, the classifier is constructed as follows. Let $c(i|j)$ be the cost of classifying an observation to the ith class when the true class is j. Then, a new tissue sample with gene expression $\mathbf{x}$ is classified into class $C(\mathbf{x})$ where

$$C(\mathbf{x}) = \arg\min_j \sum_{i=1}^{J} c(i|j) \Pr(y = j|\mathbf{x}).$$

If $c(i|j)$ are all equal, which is most frequent in practice, $C(\mathbf{x})$ becomes $\arg\max_j \Pr(y = j|\mathbf{x})$.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات