Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification

Zakariya Yahya Algamal, Muhammad Hisyam Lee*

Department of Mathematical Sciences, Universiti Teknologi Malaysia 81310 Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 14 June 2015
Received in revised form
7 October 2015
Accepted 8 October 2015

Keywords:
Adaptive elastic net
Oracle property
Regularized logistic regression
Cancer classification
Gene selection

ABSTRACT

Cancer classification and gene selection in high-dimensional data have been popular research topics in genetics and molecular biology. Recently, adaptive regularized logistic regression using the elastic net regularization, which is called the adaptive elastic net, has been successfully applied in high-dimensional cancer classification to tackle both estimating the gene coefficients and performing gene selection simultaneously. The adaptive elastic net originally used elastic net estimates as the initial weight, however, using this weight may not be preferable for certain reasons: First, the elastic net estimator is biased in selecting genes. Second, it does not perform well when the pairwise correlations between variables are not high. Adjusted adaptive regularized logistic regression (AAElastic) is proposed to address these issues and encourage grouping effects simultaneously. The real data results indicate that AAElastic is significantly consistent in selecting genes compared to the other three competitor regularization methods. Additionally, the classification performance of AAElastic is comparable to the adaptive elastic net and better than other regularization methods. Thus, we can conclude that AAElastic is a reliable adaptive regularized logistic regression method in the field of high-dimensional cancer classification.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Recently, molecular biology and genetics research has been transformed from the study of individual genes to the exploration of the whole genome. DNA microarrays technology is one such technique to measure the expression levels of thousands of genes in a single experiment [1–4]. Cancer classification based on microarray gene expression data has become one of the most active research topics in biomedical research, which is suitable for comparing the gene expression levels in tissues under different conditions, such as normal versus abnormal [5,6].

However, cancer classification with DNA microarray data is a challenging issue because of its high dimensionality and the small samples size. Typically, the number of genes is more than thousands from a hundred or less tissue samples [7,8]. Due to the high dimensionality and the small sample size, gene selection is an important issue for cancer classification and has been extensively studied in recent years. The application of gene selection methods allows the identification of a small number of important genes that can be used as biologically relevant genes of the appropriate cancer [9–11]. From the viewpoint of biologists, gene selection can increase the classification accuracy of the classification method by removing irrelevant and noisy genes [12–14].

Many gene selection methods have been proposed to select a subset of genes that can have high classification accuracy for cancer classification. Recently, regularization methods, which are capable of conducting efficient gene selection and model estimation simultaneously, have gained popularity [15,16]. From the statistical perspective, regularization methods can control the effects of the overfitting and multicollinearity [17]. Numerous statistical methods have been successfully applied in the area of cancer classification. Among them, logistic regression (LR) is considered to be a powerful discriminative method. LR provides predicted probabilities of class membership and easy interpretation of the gene coefficients [17]. However, LR is neither applicable nor suitable for high-dimensional cancer classification because the design matrix is singular. Thus, the iteration methods, such as Newton–Raphson's method cannot work [18]. Regularized logistic regression (RLR) has been successfully applied in high-dimensional cancer classification [6,19–23]. The benefits of RLR are that (a) the classification accuracy can often be improved by shrinking the regression coefficients, and (b) selecting a small subset of genes that exhibits the strongest effects provides a classification model with easy interpretation.

* Corresponding author. Tel.: +60 7 5534236; fax: +60 7 556 6162.
E-mail addresses: zak.sm_stat@yahoo.com (Z.Y. Algamal), mhl@utm.my (M.H. Lee).

http://dx.doi.org/10.1016/j.compbiomed.2015.10.008
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

An RLR with different regularization terms can be applied. The most widely and popular regularized term is the least absolute shrinkage and selection operator (LASSO) [24]. LASSO imposes the ℓ_1-norm regularization to the loss function. Because of the ℓ_1-norm property, LASSO can perform variable selection by assigning some genes coefficients to zero. For this reason, LASSO has gained popularity in high-dimensional data.

Despite the advantage of LASSO, it has three shortcomings [25,26]. First, LASSO has a biased gene selection, which means it is an inconsistent gene selection method because it regularizes all gene coefficients equally [27]. In other words, LASSO does not have the oracle property, which refers to the probability of selecting the right set of genes (with nonzero coefficients) converges to one, and that the estimators of the nonzero coefficients have asymptotically normal distribution with the same means and covariances as if the zero coefficients are known in a prior [28,29]. Related to this limitation of LASSO, concerning the oracle property, Zou [30] proposed the adaptive LASSO in which adaptive weights are used for regularizing different coefficients in the ℓ_1-norm regularization. Second, it cannot select more genes than the number of samples. Last, in the microarray gene data, there is grouping among genes, where genes that share a common biological pathway have a high pairwise correlation with each other. LASSO tries to select only one gene or a few of them among a group of correlated genes. To overcome the last two limitations, Zou and Hastie [26] proposed the elastic net regularization, for which the regularization is a linear combination of ℓ_1-norm and ℓ_2-norm. Similar to LASSO, elastic net lacks the oracle property even though it outperforms LASSO. Zou and Zhang [31] proposed adaptive elastic net to handle grouping effects and enjoy the oracle property simultaneously.

In high-dimensional classification data, however, the adaptive elastic net faces practical problems where a maximum likelihood estimate (MLE), which is usually proposed as an initial weight, is simply infeasible, and, hence, the adaptive elastic net is no longer applicable. Zou and Zhang [31] proposed using the elastic net estimates as an initial weight in adaptive elastic net; however, using this weight may not be preferable for three reasons: First, it is well known that gene selection by elastic net can be inconsistent [31,32]. In other words, this initial weight is biased in selecting genes. Second, elastic net exhibits difficulties when a group of genes is nearly linearly dependent, because it does not take into account the correlation structure among genes [33]. Last, the elastic net does not perform well when the pairwise correlations between genes are not extremely high; El Anbari and Mkhadri [34] stated that if the absolute correlation between genes is slightly less than 0.95, the elastic net may be slightly less reliable.

In this study, a new initial weight inside ℓ_1-norm regularization in adaptive elastic regularized logistic regression is proposed, which is defined as the ratio of the standard error of the ridge regression estimator to the ridge regression estimator. The main objective behind this new initial weight is to adjust the ℓ_1-norm regularization in regularized logistic regression by improving the gene selection consistency while still maintaining the grouping effects. To evaluate the effectiveness of the new initial weight, we applied three DNA microarray datasets of cancer classification. Moreover, a comparison is made with other regularization terms and initial weights.

The rest of this paper is arranged as follows: Section 2 displays the regularized logistic regression, the adaptive regularized logistic regression, and the proposed method. While Section 3 covers the real data application results. Finally, the conclusion is covered by Section 4.

2. Methods

2.1. Regularized logistic regression

Logistic regression is a statistical method to model a binary classification problem. The regression function has a nonlinear relation with the linear combination of the genes. In cancer classification, the response variable of the logistic regression has two values either 1 for the tumor class or 0 for the normal class. Assume that we have n observations and p genes. Let $y_i \in \{0,1\}$ be the response variable value for observation i, $i=1,2,...,n$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ be the i^{th} gene vector of the gene matrix $\mathbf{X}$. Then, the response variable is related to genes by

$$\pi_i = p(y_i = 1 | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \beta)}{1+\exp(\mathbf{x}_i^T \beta)}, \quad i=1,2,...,n \tag{1}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is a $p \times 1$ vector of unknown gene coefficients. The log-likelihood function of the logit transformation of Eq. (1) is defined as

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\}. \tag{2}$$

Regularized logistic regression adds a nonnegative regularization term to the negative log-likelihood function, $\ell(\beta)$, such that the size of gene coefficients in high-dimension can be controlled. Several regularization terms have been discussed in the literature [23,24,26,35]. The ℓ_1-norm regularization, proposed by Tibshirani [36], is one of the popular regularization terms. The ℓ_1-norm regularization performs gene selection and estimation simultaneously by constraining the negative log-likelihood function of gene coefficients. Thus, the RLR is defined as:

$$RLR = -\ell(\beta) + \lambda P(\beta). \tag{3}$$

The estimation of the vector β is obtained by minimizing Eq. (3)

$$\hat{\beta}_{RLR} = \text{argmin}_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i\} + \lambda\, P(\beta) \right], \tag{4}$$

where $\lambda\, P(\beta)$ is the regularization term that regularized the estimates. The penalty term depends on the positive tuning parameter, λ, which controls the tradeoff between fitting the data to the model and the effect of the regularization. In other words, it controls the amount of shrinkage. For the $\lambda = 0$, we obtain the MLE solution. In contrast, for large values of λ the influence of the regularization term on the coefficient estimate increases. Choosing the tuning parameter is an important part of the model fitting. If the focus is on classification, the tuning parameter should find the right balance between the bias and variance to minimize the misclassification error. Without loss of generality, it is assumed that the genes are standardized, $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1})\sum_{i=1}^{n} x^2{}_{ij} = 1, \quad \forall j \in \{1,2,...,p\}$. As a result, the intercept β_0 is not regularized. The estimation of the vector β using the LASSO (ℓ_1-norm regularization) is defined as:

$$\hat{\beta}_{LASSO} = \text{argmin}_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i\} + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right], \tag{5}$$

where λ is a tuning parameter. It reduces to the MLE estimator when $\lambda = 0$. On the other hand, if $\lambda \to \infty$, the regularization term forces all the gene coefficients to be zero. In practice, the value of λ is often chosen by a cross-validation procedure. Eq. (5) can be efficiently solved by the coordinate descent algorithm [37,38].

Elastic net is a regularization method for gene selection, which is introduced by Zou and Hastie [26] to deal with the first two drawbacks of LASSO. Elastic net tries to combine the ℓ_2-norm

Download English Version:

https://daneshyari.com/en/article/504855

Download Persian Version:

https://daneshyari.com/article/504855

Daneshyari.com