Expert Systems With Applications 42 (2015) 9326–9332

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Penalized logistic regression with the adaptive LASSO 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

Keywords:
Adaptive LASSO
Penalized logistic regression
Cancer classification
Gene selection

ABSTRACT

An important application of DNA microarray data is cancer classification. Because of the high-dimensionality problem of microarray data, gene selection approaches are often employed to support the expert systems in diagnostic capability of cancer with high classification accuracy. Penalized logistic regression using the least absolute shrinkage and selection operator (LASSO) is one of the key steps in high-dimensional cancer classification, as gene coefficient estimation and gene selection simultaneously. However, the LASSO has been criticized for being biased in gene selection. The adaptive LASSO (APLR) was originally proposed to overcome the selection bias by assigning a consistent weight to each gene. In high-dimensional data, however, the adaptive LASSO faces practical problems in choosing the type of initial weight. In practice, the LASSO estimator itself has been used as an initial weight. However, this may not be preferable because the LASSO is inconsistent in itself. To address this issue, an alternative initial weight in adaptive penalized logistic regression (CBPLR) is proposed. The effectiveness of the CBPLR is examined on three well-known high-dimensional cancer classification datasets using number of selected genes, area under the curve, and misclassification rate. The experimental results reveal that the proposed CBPLR is quite efficient and feasible for cancer classification. Additionally, the proposed weight is compared with APLR and LASSO and exhibits competitive performance in both classification accuracy and gene selection. The proposed CBPLR has significant impact in penalized logistic regression by selecting fewer genes with high area under the curve and low misclassification rate. Thus, the proposed weight could conceivably be used in other research that implements gene selection in the field of high dimensional cancer classification.

1. Introduction

Cancer is a term that refers to uncontrolled cellular division, growth and spread of abnormal cells. It can occur in all body parts. According to the world health organization, cancer is a disease that threatens human lives and causes the second highest rate of death globally. In cancer treatment or therapy, the classification of normal and abnormal patterns of the cells is one of most important and significant processes during the diagnosis of cancer. Recently, the use of expert classifier systems in cancer diagnosis is increasing (Akay, 2009). One of the major goal of these expert systems is to extract the useful knowledge from past diagnosis database. With the fast development and widely used of the DNA microarray technology in cancer research, a highly accurate expert classifier system is needed (Du, Li, Li, & Fei, 2014; Zheng, Chong, & Wang, 2011). DNA microarray technology allows producing of thousands of genes. Dealing with all produced genes by an expert classifier system is a challenging and time consuming task. Therefore, selecting irrelevant genes is an important part in order to support the expert classifier system in high-dimensional cancer classification.

One of the properties of microarray data is that the number of genes, p, exceeds the number of tissues (patients), n (Alonso-González, Moro-Sancho, Simon-Hurtado, & Varela-Arrabal, 2012; Cui, Zheng, Yang, & Sha, 2013; Kalina, 2014; Ma & Huang, 2008). Dealing with the situation $p > n$, which is commonly known as high-dimensional data, poses a challenging task in the application of the statistical classification methods (Piao, Piao, Park, & Ryu, 2012). Overfitting and multicollinearity are the most common problems that arise in high-dimensional data when applying statistical classification methods. These issues make statistical microarray classification methods very difficult (Chen, Wang, Wang, & Angelia, 2014; Pang, Havukkala, Hu, & Kasabov, 2007; Peng, Fu, Liu, Fang, & Jiang, 2013).

From the biological perspective, only a small subset of genes is strongly indicative of a targeted disease, and most genes are irrelevant to cancer classification. The irrelevant genes may introduce noise and decrease the classification accuracy (Chandra & Gupta, 2011). Moreover, from the statistical perspective, too many genes may lead to overfitting and can negatively influence the classification

* Corresponding author. Tel.: +60197007779; fax: +6075566162.
E-mail addresses: zak.sm_stat@yahoo.com (Z.Y. Algamal), mhl@utm.my, hisyamlee@gmail.com (M.H. Lee).

http://dx.doi.org/10.1016/j.eswa.2015.08.016
0957-4174/

performance (Liang et al., 2013). Due to the significance of these problems, effective gene selection methods are desirable to help to classify the different cancer types and improve prediction accuracy. Consequently, removing irrelevant and noisy genes is an important target when dealing with high-dimensional cancer classification. In principle, gene selection aims to select a relatively small set of genes from a high-dimensional gene dataset, and, therefore, achieves high classification accuracy (Lei, Yue, & Berens, 2012; Pang, George, Hui, & Tiejun, 2012). Furthermore, selecting important genes can also help in early diagnosis and drug discovery for cancer patients (Chen et al., 2014). Numerous statistical methods have been successfully applied in the area of cancer classification. Among them, logistic regression (LR) is considered as a powerful discriminative method. LR provides the predicted probabilities of class membership and easy interpretation of the gene coefficients (Liang et al., 2013). However, LR is neither applicable nor suitable for the high-dimensional cancer classification, because the Hessian matrix will not have full rank (Kastrin & Peterlin, 2010). Thus, the iteration methods such as Newton–Raphson's method cannot work (Bielza, Robles, & Larrañaga, 2011).

Recently, there has been growing interest in applying the penalized methods in high-dimensional cancer classification (Bielza et al., 2011; Bootkrajang & Kabán, 2013; Nan et al., 2012; Zou et al., 2015). To tackle both estimating the gene coefficients and performing gene selection simultaneously, penalized logistic regression (PLR) was successfully applied in high-dimensional cancer classification (Cawley & Talbot, 2006; Li & Eng Chong, 2005; Shevade & Keerthi, 2003; Zhenqiu et al., 2007; Zhu & Hastie, 2004). A PLR with different penalties can be applied. The most widely and popular penalty is the L_1-penalty, which is known as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996). The LASSO imposes the L_1-penalty to the loss function. Because of the L_1-penalty property, the LASSO can perform variable selection by assigning some gene coefficients to zero. For this reason, the LASSO obtains its popularity in high dimensional data. SLR with L_1-penalty gives a sparse solution with high classification accuracy.

Despite the advantage of the LASSO, it has three shortcomings (Wang, Nan, Rosset, & Zhu, 2011; Zheng & Liu, 2011). First, it cannot select more genes than the number of tissues. Second, in microarray gene data, there is grouping among genes, where genes that share a common biological pathway have a high pairwise correlation with each other. The LASSO tries to select only one gene or a few of them among a group of correlated genes. To overcome the first two limitations, Zou and Hastie (2005) proposed the elastic net penalty, for which the penalty is a linear combination of L_1-penalty and L_2-penalty. Last, the LASSO has a bias in gene selection, because it penalizes all the gene coefficients equally (Fan, Fan, & Barut, 2014). In other words, the LASSO does not have the oracle properties, which refer to the probability of selecting the right set of genes (with nonzero coefficients) converges to one, and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients are known in advance (Fan & Li, 2001).

In relation to the last limitation of the LASSO, the oracle properties, Zou (2006) proposed the adaptive LASSO in which the adaptive weights are used for penalizing different coefficients in the L_1-penalty. In high-dimensional classification data, however, the adaptive LASSO faces practical problems in choosing the type of initial weight. As a result, the LASSO estimator itself has been used as an initial estimator in solving the adaptive LASSO (Bühlmann & Van De Geer, 2011; Lin, Xiang, & Zhang, 2009). In fact, using the LASSO estimator in the adaptive LASSO when $p > n$ may not be preferable for two reasons. First, LASSO estimator is inconsistent in itself. In other words, this initial weight is biased in selecting genes. Second, it does not take into account the weights for all the genes in any implantation, which means, some genes will be selected and the others will be set to zero.

In this study, correlation-based weight is proposed as an alternative initial weight inside the L_1-penalty in penalized logistic regression (CBPLR). The main objective behind this new initial weight is to adjust the L_1-penalty in the PLR by improving consistent genes selection (oracle property). The main aim of this study is to show the effectiveness of the proposed weight for the gene selection in high-dimensional cancer classification. The computational effectiveness of the proposed weight is compared with the performance of the LASSO and the adaptive LASSO on three benchmark gene expression datasets. It is observed that the proposed weight outperformed the other two methods in terms of classification accuracy and the number of selected genes.

The remainder of this paper is arranged as follows: Several related papers are listed in Section 2. The methodology applied in this study is detailed in Sections 3 and 4. In Section 5, the experimental study is carried out, including a description of the dataset and a discussion of the main results. Finally, the main conclusion is drawn in Section 6.

2. Related work

Among existing expert classifier systems in high-dimensional cancer classification, PLR has demonstrated its capability in providing an easily interpretable expert system with a highly classification accuracy. This paper is developed independently, although, in some aspects, it is related to other papers (Cawley & Talbot, 2006; Li & Eng Chong, 2005; Shevade & Keerthi, 2003; Zhenqiu et al., 2007; Zhu & Hastie, 2004).

Shevade and Keerthi (2003) proposed new algorithm based on the Gauss–Seidel method in solving PLR with application in gene selection in microarrays cancer classification data. Zhu and Hastie (2004) proposed PLR as an alternative classification method to support vector machine in microarray cancer classification to take into account probability estimation. Li and Eng Chong (2005) combined two dimension reduction methods, singular value decomposition and partial least squares, with PLR to enhance the classification accuracy and computational speed. Fort and Lambert-Lacroix (2005) proposed to combine the partial least squares and ridge PLR. The classification performance is illustrated on leukemia, colon and prostate datasets. An extension of PLR was proposed by Kim, Kwon, and Heun Song (2006) to deal with multi-class microarrays cancer classification. Cawley and Talbot (2006) proposed to use PLR with Bayesian regularization in gene selection for cancer classification data. Zhenqiu et al. (2007) proposed a novel method that combine the PLR with non-convex penalty in cancer classification data.

Bielza et al. (2011) presented a new PLR method based on the evolution of the regression coefficients using estimation of distribution algorithms. The main contribution is to avoid the determination of the penalization term in gene selection. An improvement of GLMNET algorithm for L_1-PLR was proposed by Yuan, Ho, and Lin (2012) to address some theoretical and implementation issues of the GLMNET.

Liang et al. (2013) proposed and investigated a novel PLR with $L_{1/2}$ penalty for gene selection in cancer classification data. Bootkrajang and Kabán (2013) utilized PLR to detect mislabeled arrays using Bayesian regularization. Vincent and Hansen (2014) proposed new algorithm to solve penalized group LASSO using multinomial logistic regression to deal with multi-class classification.

3. Penalized 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,