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,

the response variable of the logistic regression has two values either 1 for the tumor class or 0 for the normal class. Let $\mathbf{y}_i \in \{0, 1\}$ be a vector of size $n \times 1$ of tissues, and let $\mathbf{x}_i$ be a $p \times 1$ vector of genes. The logistic transformation of the vector of probability estimates $\pi_i = p(y_i = 1|\mathbf{x}_i)$ is modeled by a linear function, logit transformation:

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \sum_{j=1}^{p} \mathbf{x}_{ij}{}^T \beta_j, \quad i = 1, 2, \ldots, n, \tag{1}$$

where β_0 is the intercept and β_j is a $p \times 1$ vector of unknown gene coefficients. The log-likelihood function of Eq. (1) is defined as:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\}. \tag{2}$$

Logistic regression offers the advantage of simultaneously estimating the probabilities π_i and $1 - \pi_i$ for each class and classifying subjects. The probability of classifying the ith sample in class 1 is estimated by $\hat{\pi}_i = \exp(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_{ij}{}^T \beta_j)/1 + \exp(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_{ij}{}^T \beta_j)$. The predicted class is then obtained by $I\{\hat{\pi}_i > 0.5\}$, where $I(\cdot)$ is an indicator function.

PLR adds a nonnegative penalty term to Eq. (1), such that the size of the gene coefficients in high-dimension cancer classification can be controlled. Several penalty terms have been discussed in the literature (Hoerl & Kennard, 1970; Liang et al., 2013; Tibshirani, 1996; Zhenqiu et al., 2007). The L_1-penalty, proposed by Tibshirani (1996), is one of the popular penalty terms. The L_1-penalty performs genes selection and estimation simultaneously by constraining the log-likelihood function of gene coefficients. The penalized method for the logistic regression is obtained by adding the penalty term to the negative log-likelihood function:

$$\text{PLR} = -\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\} + \lambda P(\beta). \tag{3}$$

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

$$\hat{\beta}_{PLR} = \arg\min_{\beta}\left[-\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\} + \lambda P(\beta)\right], \tag{4}$$

where $\lambda P(\beta)$ is the penalty term that penalizes 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 penalty. In other words, it controls the amount of shrinkage. For $\lambda = 0$, we obtain the MLE solution, while for large values of λ the influence of the penalty term on the coefficient estimates increases. Choosing the tuning parameter is an important part of the model fitting. If we are interesting in classification, the tuning parameter should find the right balance between the bias and the 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 $(n^{-1}) \sum_{i=1}^{n} x^2{}_{ij} = 1, \quad \forall j \in \{1, 2, \ldots, p\}$. As a result, the intercept β_0 is not penalized. The estimation of the vector β using LASSO (L_1-penalty) is defined as:

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

where λ is a tuning parameter. It reduces to the MLE estimator when $\lambda = 0$. On the other hand, if $\lambda \to \infty$, the penalty forces all the genes to be zeros. In practice, the value of λ is often chosen by a cross validation (CV) procedure. Eq. (5) can be efficiently solved by using the coordinate descent algorithm (Friedman, Hastie, & Tibshirani, 2010; Park & Hastie, 2008).

The LASSO has an advantage in that it is computationally feasible in high dimensional classification data. On the other hand, the LASSO has three main drawbacks. First of all, if $p > n$, the LASSO selects at most n genes because of the nature of the convex optimization problem. In addition, the LASSO cannot handle the effect of grouping. When the pairwise correlations among a group of genes are very high, then the LASSO tends to select only one gene from the whole group and does not take into account which one is selected (Zeny, 2012; Zou & Hastie, 2005). Lastly, the LASSO lacks the oracle properties, as stated by Fan and Li (2001).

4. Adaptive penalized logistic regression

According to Fan and Li (2001), a good penalty term should result in an estimator with three properties: unbiasedness, sparsity and continuity. Unbiasedness means the resulting estimator has no over penalization for large parameters to avoid unnecessary modeling biases. Sparsity is another property that an estimator enjoys. In other words, the resulting estimator automatically sets insignificant parameters to zero. Lastly, continuity is the third property, meaning that the resulting estimator is continuous in data in order to avoid instability in model prediction.

One of the main reasons for the LASSO not being consistent, i.e., lacking the oracle property (Fan & Li, 2001) is that it equally penalizes all the coefficients, which over-penalizes the irrelevant genes leading it to be a biased estimator. To alleviate this drawback, Zou (2006) proposed the adaptive LASSO in which adaptive weights are used for penalizing different coefficients in the L_1-penalty. The basic idea behind the adaptive LASSO is that by assigning a higher weight to the small coefficients and lower weight to the large coefficients, it is possible to reduce the selection bias; therefore, it can consistently select the model. Furthermore, the adaptive LASSO solution is continuous from its definition, which enables it to enjoy oracle properties. The PLR using the adaptive LASSO (APLR) of β is defined by:

$$\hat{\beta}_{APLR} = \arg\min_{\beta}\left[-\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\} + \lambda \sum_{j=1}^{p} w_j |\beta_j|\right], \tag{6}$$

where $\mathbf{w} = (w_1, \ldots, w_p)^T$ is $p \times 1$ data-driven weight vector. This depends on the root n-consistent initial values of $\hat{\beta}$ and $w_j = (|\hat{\beta}_j|)^{-\gamma}$, where γ is a positive constant. The adaptive LASSO originally used MLE estimates as the initial weight (Zou, 2006). This is no longer valid in high-dimensional data. Several researchers have used the LASSO estimates as an alternative initial weight (Bühlmann & Van De Geer, 2011). However, using the LASSO estimator in the adaptive LASSO penalized logistic when $p > n$ may not be preferable for three reasons. First, the LASSO estimator is inconsistent in itself. In other words, this initial weight is biased in the selection of genes. Second, it does not take into account the weights for all the genes in any implantation which means that some genes will be selected and the others will be set to zero.

To overcome these limitations, the correlation-based weight has been proposed as a CBPLR. The idea behind using the correlation-based estimator as an initial weight is that it can give weights to all genes, which is very important to take into account all the gene information when performing gene selection in the adaptive penalized likelihood. As a result, it can adjust the L_1-penalty in PLR by

Table 1
Details of the used datasets.

Dataset type	n	Genes	Classes
Colon	62	2000	Tumor/Normal
Prostate	102	5966	Tumor/Non-tumor
DLBCL	77	7129	DLBCL/FL

improving consistent genes selection. The CBPLR is defined as:

$$\hat{\beta}_{\text{CBPLR}} = \arg\min_{\beta}\left\{-\sum_{i=1}^{n}\{y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\} + \lambda \sum_{j=1}^{p} w_{j(CB)}|\beta_j|\right\} \quad (7)$$

where $\mathbf{w}_{\text{CB}} = (w_{1(\text{CB})}, \ldots, w_{p(\text{CB})})^T$ be the CB weight vector and $w_j = [\text{abs}(\hat{\beta}_{j(\text{CB})})]^{-\gamma}, \quad j = 1, 2, \ldots, p$. Then a coordinate descent method can be used to solve Eq. (7). The $\hat{\beta}_{j(\text{CB})}$ is calculated according to the correlation-based penalty, which was introduced by Tutz and Ulbricht (2009). This penalty does not perform gene selection, but it has the capability of dealing with correlated genes. The $\hat{\beta}_{j(\text{CB})}$ is defined as:

$$\hat{\beta}_{\text{CB}} = \arg\min_{\beta}\left(-\ell(\beta_0, \beta) + \lambda \sum_{i=1}^{p-1}\sum_{j>i}\left\{\frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}}\right\}\right) \quad (8)$$

where ρ_{ij} represents the pairwise correlation between the ith and jth genes.

Ridge regression is a special case from Eq. (8) when the $\rho_{ij} = 0$. Eq. (8) is a convex function; this means that there is always a minimum local solution. Unfortunately, this penalty is no longer be a convex function when $\rho_{ij} = 1 \quad \forall i \neq j$. This reason makes Eq. (8) inapplicable when there is at least one pairwise correlation between genes equal to one. Compared to LASSO and to APLR, this drawback does not affect their solutions, because their function still convex. Furthermore, it is well known that correlation used in Eq. (8) is extremely sensitive to outliers, and can be strongly affected the results. In the other hand, LASSO and APLR tends to select more irrelevant genes when CV method used in estimating the tuning parameter (Lin et al., 2009). Interestingly, correlation-based penalty performs well when CV is used.

For practical applications, one has to decide the values of λ. Classically, CV has been widely used. However, it is computationally intensive for the CBPLR, simply because there are two tuning parameters, λ and γ. For simplicity, $\gamma = 1$ was used for the real data application. Then, CBPLR tuning parameters were reduced to only λ.

5. Results and discussion

To prove the effectiveness of the proposed initial weight, three DNA microarray datasets, with different sample sizes and number of genes, were used. First, the colon dataset (Alon et al., 1999). Second, the prostate dataset (Singh et al., 2002); and, third, the diffuse large B-cell lymphoma (DLBCL; Shipp et al., 2002). Table 1 lists the details of these datasets. To guarantee a fair comparison of the proposed method with the other two methods, two procedures were set up. First, two datasets were randomly generated from each microarray dataset: a training set with 70% of the original size and a testing set with 30%. Second, the CV method with 10-fold was conducted depending on the training set to find the optimal value of λ. All computations were carried out in the R software using the *glmnet* package.

Depending on 50 partitions of the training and testing sets, the average value of the number of selected genes, the area under the curve (AUC) for the training set, and the misclassification rate (Miss.) for the testing set were used as evaluation criteria for the classification performance.

Table 2
Classification performance results on colon cancer over 50 partitions.

Method	Number of selected genes	AUC	Misclassification rate
LASSO	17	0.915	0.263
APLR	16	0.928	0.210
CBPLR	10	0.966	0.105

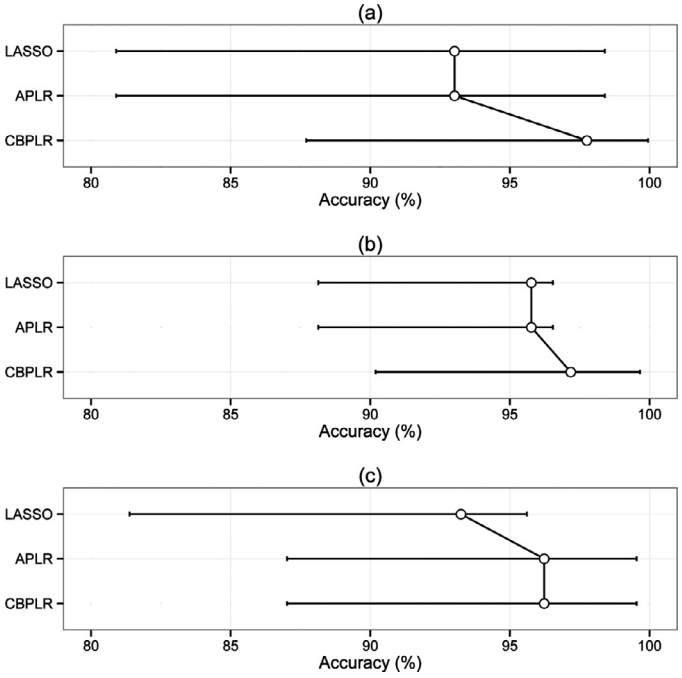

Fig. 1. Conficence interval 95% of the classification accuray average: (a) colon data, (b) prostate data, and (c) DLBCL data.

5.1. Colon results

The colon cancer dataset contains gene expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes obtained with an Affymetrix oligonucleotide array. A subset of 2000 genes with highest minimal intensity across the samples was used (Alon et al., 1999). Table 2 reports the criteria results for the CBPLR as well as for the APLR and the LASSO.

As we can see from Table 2, the CBPLR was superior to all the compared methods in terms of selected genes, AUC, and misclassification rate. Hence, the utilization of the CBPLR yielded a higher AUC. Furthermore, it can be seen that the misclassification rate of the CBPLR was about 0.105 lower than that of the APLR and the LASSO, respectively.

The number of genes selected by each method is an important factor. Methods selecting more genes tend to overfit the data. Hence, methods with a small number of selected genes are preferred. For a comparison of methods in terms of the number of selected genes, the CBPLR outperformed the other two methods. It selected 10 genes compared to 16 and 17 genes for the APLR and LASSO, respectively. Fig. 1(a) shows a 95% confidence interval for the mean of the classification accuracy. It clearly shows that the CBPLR significantly outperformed the APLR and LASSO.

Table 3
Classification performance results on prostate cancer over 50 partitions.

Method	Number of selected genes	AUC	Misclassification rate
LASSO	26	0.957	0.129
APLR	26	0.957	0.096
CBPLR	16	0.971	0.064

Table 4
Classification performance results on DLBCL cancer over 50 partitions.

Method	Number of selected genes	AUC	Misclassification rate
LASSO	29	0.941	0.208
APLR	25	0.953	0.167
CBPLR	17	0.953	0.083

Table 5
Two-way ANOVA for average classification accuracy over 25 times.

Source	df	SS	MS	*F*	*p*-value
Methods	2	6261.7	3130.8	189.1	0.000
Datasets	2	1946.9	973.45	58.8	0.000
Error	445	7366.7	16.554		
Total	449	15575.3			

Table 6
P-value of Duncan's multiple range test for average classification accuracy.

	LASSO	APLR	CBPLR
LASSO		0.036	0.000
APLR			0.004
CBPLR			

5.2. Prostate results

The original prostate dataset contains 12,600 genes for 52 prostate tumor samples and 50 non-tumor tissues. A subset of 5966 genes was adapted in the classification (Singh et al., 2002). Table 3 presents the results obtained from the evaluation criteria.

The classification performance in the training set using the AUC of the proposed method, CBPLR, was 0.971, which was better than 0.957 and 0.957 obtained by the APLR and LASSO, respectively, which indicated the better classification ability of the CBPLR than the other two methods. Depending on the test set, the CBPLR reduced the misclassification significantly in comparison with the other two methods. The reduction of misclassification using the CBPLR was 33.34% and 50.38% compared with the APLR and LASSO, respectively. In addition, the CBPLR reduced the number of original genes from 5966 to 16, the APLR selected 26 genes, and the LASSO had 26 genes. This indicated that the CBPLR outperformed the other two methods in terms of the number of selected genes.

Fig. 1(b) displays the confidence interval 95% of the mean of the classification accuracy mean. It can be concluded from Fig. 1(b) that the CBPLR has much better classification accuracy compared to the other two methods.

5.3. DLBCL results

The DLBCL dataset consists of the gene expression values for 77 samples, which were measured by high-density oligonucleotide microarrays of the two most prevalent adult lymphoid malignancies: 58 samples of the DLBCL and 19 samples of follicular lymphoma (FL). Each sample contained 7129 gene expression values (Shipp et al., 2002). The evaluation criteria results are shown in Table 4. As shown in Table 4, both the CBPLR and APLR provided similar results, followed by the LASSO, which was slightly worse based on the AUC.

Moreover, the reliability of the CBPLR was also assessed from its misclassification rate value. It ranked the CBPLR above the APLR and the LASSO. Although the CBPLR and APLR have the same AUC, the CBPLR provided a reduction in the misclassification rate of about 50.30% compared to the APLR. It is also seen from Table 4 that the CBPLR selected significantly less genes than the other two methods.

A 95% confidence interval for the mean of the classification accuracy over 50 partitions is depicted in Fig. 1(c). It can be seen from Fig. 1(c) that the CBPLR has the same mean for the classification accuracy to the APLR. On the other hand, the LASSO provided worse classification accuracy compared to the CBPLR and APLR, respectively.

5.4. Stability test for the proposed method

In the stability test for the proposed method, the CBPLR seeks to prove that it can classify high-dimensional cancer data with a high degree of accuracy compared to the other two used methods. Depending on the training dataset, a two-way analysis of variance (ANOVA) was used as a statistical test to check whether the CBPLR, APLR and the LASSO were statistically significant and if there was any significant difference between the three datasets used in terms of classification accuracy. Table 5 reports the two-way ANOVA results. From Table 5, the results showed statistically significant differences between the CBPLR and the two other used methods in terms of classification accuracy. In addition, we can see that the colon,

Table 7
Classification accuracy (%) for different splitting of the three datasets used.

	50%:50%		60%:40%		70%:30%		80%:20%	
	Train	Test	Train	Test	Train	Test	Train	Test
Colon								
LASSO	81.39	79.06	83.72	76.74	93.02	73.72	92.95	90.69
APLR	87.32	83.72	88.37	86.04	93.02	79.00	92.95	91.54
CBPLR	92.95	90.14	91.75	91.75	97.76	89.54	92.95	91.45
Prostate								
LASSO	91.75	89.75	92.75	91.75	95.77	87.15	93.02	91.54
APLR	94.36	92.95	91.54	91.54	95.77	90.47	94.36	92.95
CBPLR	95.34	92.82	96.22	95.34	97.18	93.69	94.89	94.77
DLBCL								
LASSO	91.54	91.54	92.73	91.54	93.25	79.25	91.54	91.54
APLR	92.80	92.95	94.01	92.21	96.23	83.38	94.36	93.74
CBPLR	93.22	93.34	95.12	94.24	96.23	91.75	96.22	95.34

prostate, and the DLBCL datasets had different classification accuracy values.

Furthermore, Duncan's multiple range test was used to obtain more detailed information about the differences between the CBPLR and the two other used methods. Table 6 lists the p-value of each compared pair of methods. It is apparent from Table 6 that the CBPLR showed statistical differences compared to the APLR and LASSO in terms of classification accuracy.

To further prove the stability of the results for the proposed method, the classification accuracy using the CBPLR is also consistently improved for different percentages of splitting the original dataset for each of the dataset used. The average classification accuracy over 25 times for both the training and testing partitions are shown in Table 7. As can be seen from Table 7, the CBPLR performed remarkably well compared to the APLR and LASSO. It always achieved higher classification accuracy for both the training and testing sets for each dataset. In contrast, the LASSO provided less classification accuracy in all cases.

Overall, the results demonstrated the fact that the CBPLR is effective in high-dimensional cancer classification. The CBPLR not only improved the classification accuracy but also identified a small subset of genes compared to the APLR and LASSO.

6. Conclusions

High-dimensional classification problems associated with DNA microarray data analysis constitute a very important research area in cancer classification. In the present paper, we proposed and applied a CBPLR model, CBPLR, to simultaneously estimate the gene coefficients and perform gene selection, and then improve the classification performance of the expert classifier system using high-dimensional DNA microarray data.

The proposed method, CBPLR, has been evaluated in terms of the number of selected genes, AUC, and misclassification rate through applying three high-dimensional cancer classification datasets. The experiment results consistently indicated that the CBPLR has the ability to significantly reduce the size of the relevant genes compared to APLR and LASSO. Moreover, it is observed that the CBPLR has the superiority in terms of AUC. It achieved maximum AUC of 0.966, 0.971, and 0.953 for colon, prostate, and DLBCL datasets. In addition, the misclassification rate obtained by the proposed method was the lowest for all of the three high-dimensional cancer classification datasets, as compared to APLR and LASSO. Furthermore, the stability test results confirmed the superiority of our proposed method over the different splitting values. Overall, the results demonstrated the fact that CBPLR is a very competitive method to accurately analyze high-dimensional DNA microarray data for cancer classification. For practical use of the results, the CBPLR can be applied straightforward to other types of high-dimensional classification data related to the medical field.

Although CBPLR results yielded significantly better performance, it has two limitations. Firstly, if the pairwise correlation between two genes equal to one, then CBPLR is no longer a convex function. Secondly, CBPLR depends on the pairwise correlation between genes in its calculations. It is well known that microarray dataset with many genes often contains outliers, therefore, the pairwise correlation used in CBPLR is extremely sensitive to outliers, and can be strongly affected the results. In the aspects concerning the future research, the present work can be extended to cover high-dimensional multiclass classification cancer data. Further study can be conducted with more emphasis on ultrahigh-dimensional DNA microarray data for cancer classification.

References

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*, 3240–3247.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, 96*, 6745–6750.

Alonso-González, C. J., Moro-Sancho, Q. I., Simon-Hurtado, A., & Varela-Arrabal, R. (2012). Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications, 39*, 7270–7280.

Bielza, C., Robles, V., & Larrañaga, P. (2011). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications, 38*, 5110–5118.

Bootkrajang, J., & Kabán, A. (2013). Classification of mislabelled microarrays using robust sparse logistic regression. *Bioinformatics, 29*, 870–877.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Heidelberg: Springer.

Cawley, G. C., & Talbot, N. L. C. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics, 22*, 2348–2355.

Chandra, B., & Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics, 44*, 529–535.

Chen, K.-H., Wang, K.-J., Wang, K.-M., & Angelia, M.-A. (2014). Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Applied Soft Computing, 24*, 773–780.

Cui, Y., Zheng, C.-H., Yang, J., & Sha, W. (2013). Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. *Computers in Biology and Medicine, 43*, 933–941.

Du, D., Li, K., Li, X., & Fei, M. (2014). A novel forward gene selection algorithm for microarray data. *Neurocomputing, 133*, 446–458.

Fan, J., Fan, Y., & Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics, 42*, 324–351.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fort, G., & Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics, 21*, 1104–1111.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Kalina, J. (2014). Classification methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering, 34*, 10–18.

Kastrin, A., & Peterlin, B. (2010). Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. *Expert Systems with Applications, 37*, 5178–5185.

Kim, Y., Kwon, S., & Heun Song, S. (2006). Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis, 51*, 1643–1655.

Lei, Y., Yue, H., & Berens, M. E. (2012). Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*, 262–272.

Li, S., & Eng Chong, T. (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2*, 166–175.

Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., & Zhang, H. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics, 14*, 198–210.

Lin, Z., Xiang, Y., & Zhang, C. (2009). Adaptive Lasso in high-dimensional settings. *Journal of Nonparametric Statistics, 21*, 683–696.

Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics, 9*, 392–403.

Nan, X., Wang, N., Gong, P., Zhang, C., Chen, Y., & Wilkins, D. (2012). Biomarker discovery using 1-norm regularization for multiclass earthworm microarray gene expression data. *Neurocomputing, 92*, 36–43.

Pang, H., George, S. L., Hui, K., & Tiejun, T. (2012). Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*, 1422–1431.

Pang, S., Havukkala, I., Hu, Y., & Kasabov, N. (2007). Classification consistency analysis for bootstrapping gene selection. *Neural Computing and Applications, 16*, 527–539.

Park, M. Y., & Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics, 9*, 30–50.

Peng, H., Fu, Y., Liu, J., Fang, X., & Jiang, C. (2013). Optimal gene subset selection using the modified SFFS algorithm for tumor classification. *Neural Computing and Applications, 23*, 1531–1538.

Piao, Y., Piao, M., Park, K., & Ryu, K. H. (2012). An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics, 28*, 3306–3315.

Shevade, S. K., & Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics, 19*, 2246–2253.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine, 8*, 68–74.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*, 203–209.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*, 267–288.

Tutz, G., & Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing, 19*, 239–253.
Vincent, M., & Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis, 71*, 771–786.
Wang, S., Nan, B., Rosset, S., & Zhu, J. (2011). Random Lasso. *The Annals of Applied Statistics, 5*, 468–485.
Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). An improved GLMNET for L1-regularized logistic regression. *Journal of Machine Learning Research, 13*, 1999–2030.
Zeny, Z. F. (2012). The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*, 629–636.
Zheng, C.-H., Chong, Y.-W., & Wang, H.-Q. (2011). Gene selection using independent variable group analysis for tumor classification. *Neural Computing and Applications, 20*, 161–170.
Zheng, S., & Liu, W. (2011). An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification. *Computers in Biology and Medicine, 41*, 1033–1040.
Zhenqiu, L., Feng, J., Guoliang, T., Suna, W., Fumiaki, S., & Ming, T. (2007). Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology, 6*, 1–22.
Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics, 5*, 427–443.
Zou, F., Wang, Y., Yang, Y., Zhou, K., Chen, Y., & Song, J. (2015). Supervised feature learning via l2-norm regularized logistic regression for 3D object recognition. *Neurocomputing, 151, Part 2*, 603–611.
Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association, 101*, 1418–1429.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320.