Outlier Modification and Gene Selection for Binary Cancer Classification using Gaussian Linear Bayes Classifier

Md. Hadiul Kabir
Laboratory of Bioinformatics *hadi_ru07@yahoo.com*
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

Md. Nurul Haque Mollah *mollah.stat.bio@ru.ac.bd*
Laboratory of Bioinformatics
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

Abstract

Gaussian linear Bayes classifier is one of the most popular approaches for classification. However, it is not so popular for cancer classification using gene expression data due to the inverse problem of its covariance matrix in presence of large number of gene variables with small number of cancer patients/samples in the training dataset. To overcome these problems, we propose few top differentially expressed (DE) genes from both upregulated and downregulated groups for binary cancer classification using the Gaussian linear Bayes classifier. Usually top DE genes are selected by ranking the *p*-values of *t*-test procedure. However, both *t*-test statistic and Gaussian linear Bayes classifier are sensitive to outliers. Therefore, we also propose outlier modification for gene expression dataset before applying to the proposed methods, since gene expression datasets are often contaminated by outliers due to several steps involves in the data generating process from hybridization to image analysis. The performance of the proposed method is investigated using both simulated and real gene expression datasets. It is observed that the proposed method improves the performance with outlier modifications for binary cancer classification.

Keywords: Gene Expression, Outlier Modification, Top DE Genes Selection, Binary Classification, Gaussian Bayes Classifier, Misclassification Error Rate (MER).

1. INTRODUCTION

The classification of patient samples into one of the two classes (normal/cancer) using their gene expression profile is an important task and has been attracted widespread attention [1-3]. The gene expression profiles measured through DNA microarray technology provide accurate, reliable and objective cancer classification. It is also possible to uncover cancer subclasses that are related with the efficacy of anti-cancer drugs that are hard to be predicted by pathological tests [3-5]. Previously, cancer classification has always been morphological and clinical based but they are reported to have several limitations in diagnostic ability [6-9]. The recent advent of microarray technology has allowed the simultaneous monitoring of thousands of genes, which motivated the development in cancer classification using gene expression data. For the last few years, classification problem using gene expression has been extensively studied by researcher in the area of statistics, machine learning and databases [10-15]. In order to gain a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been proposed [16-18]. A number of methods have been proposed for cancer classification with promising results based on gene expression datasets, such as the decision tree, support vector machine (SVM), linear discriminant analysis (LDA), Bayesian network [19-

21]. Though Gaussian Bayes classifier is one of the most powerful statistical approach for classification, but it is not so popular for cancer classification based on gene expression data due to the inverse problem of its covariance matrix in presence of large number of gene variables with small number of cancer patients/samples in the training gene expression dataset. To overcome these problems, our proposal is to use few informative genes/features to train the Gaussian Bayes Classifier.

A gene expression dataset is very different from any of the other datasets. It has very high dimensionality, usually contains hundred thousands of genes with very small sample size. Most genes in the dataset are irrelevant to cancer distribution. From these points of views, relevant gene selection prior to cancer classification is essential. In fact, relevant gene selection removes a large number of irrelevant genes, which improves the classification accuracy. The feature selection algorithms are considered to be an important way of identifying crucial/relevant genes for classification. There are three types of feature selection algorithms (filtering/wrapper /embedded) exist in the literature [22-28]. The disadvantage of wrapper and embedded feature selection approaches than the filtering approaches, are computationally intensive, classifier dependent selection and higher risk of over fitting. The advantages of filtering techniques [27-28] are (i) they are easily scalable to very high-dimensional datasets (ii) they are computationally simple and fast, and (iii) they are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated. Therefore, in this paper we consider filtering approach to select few top differentially expressed (DE) genes from both upregulated and downregulated groups for binary cancer classification using linear Bayes classifier in this paper, since equally expressed (EE) gene has no significant contribution to the minimization of misclassification error rate (MER). There are some filtering approaches [27,28] for selection of important features/genes from top ranked genes detected by t-test or ANOVA approaches. However, both *t*-statistic and linear Bayes classifier are sensitive to outliers. Therefore, in this paper, we would like to propose outlier modification for gene expression dataset before gene selection and cancer classification using *t*-statistic and linear Bayes classifier, respectively, since gene expression datasets are often contaminated by outliers due to several steps involves in the data generating process from hybridization to image analysis.

We organized this paper as follows. In section 2, we formulate the linear Bayes classifier and the proposed method for binary cancer classification (normal/cancer). In section 3, we described the results of the simulated and real gene expressions datasets. Finally, we end this paper with a conclusion.

2. FORMULATION OF GAUSSIAN LINEAR BAYES CLASSIFIER FOR BINARY CANCER CLASSIFICATION

Suppose we have a training gene expression dataset obtained from n_1 normal patients and n_2 cancer patients with p genes, where the column vector $\boldsymbol{x}_{jt}$ consist of expressions of p genes (t = 1, 2,…, n_j). Here the problem is to classify a new patient having the vector $\boldsymbol{x} = (x_1, x_2, ..., x_p)^T$ of expressions with p genes (known as test vector) into one of $m=2$ groups (normal/cancer) corresponding to two populations Π_1 and Π_2, respectively. Two solve this problem using Gaussian Bayes classifier, let a training data vector $\boldsymbol{x}_{jt}$ follows Gaussian density function $f_j(\boldsymbol{x}_j) = N(\boldsymbol{x}_j / \boldsymbol{\mu}_j, \boldsymbol{V}_j)$, where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{V}_j$ is the covariance matrix for this population (t = 1, 2, ..,n_j, j = 1, 2). If the test vector $\boldsymbol{x}$ originates randomly from one of this $m=2$ populations, then it follows the mixture of $m=2$ multivariate normal distributions as follows

$$f(\boldsymbol{x}) = q_1 f_1(\boldsymbol{x}) + q_2 f_2(\boldsymbol{x}) \tag{1}$$

Where q_j is the mixing proportion or prior probability of $\boldsymbol{x} \in \Pi_j$ such that $\sum_{j=1}^{m} q_j$ .Then the posterior pdf of $\boldsymbol{x} \in \Pi_j$ is given by

$$g(\Pi_j \mid \boldsymbol{x}) = \frac{q_j f_j(\boldsymbol{x})}{f(\boldsymbol{x})}, \qquad j = 1,2 \tag{2}$$

To formulate the Bayesian classifiers, the space of all observations is divided into *m* mutually exclusive regions R_j , (j = 1, . . .,*m*). The classification region R_j with the cost of misclassifying an observation from Π_j as from Π_i is defined for classifying $\boldsymbol{x}$ to the population Π_j as follows:

$$R_j : g(\Pi_j \mid \boldsymbol{x})\, C(i|j) > g(\Pi_i \mid \boldsymbol{x})\, C(j|i), \qquad i, j=1,2 \ (i \neq j)$$
$$\Rightarrow R_j : U_{ij} > \lambda_{ij} = \log \frac{[q_k C(j \mid i)]}{[q_j C(i \mid j)]} \ , \tag{3}$$

Where $C(i \mid j)$ is the cost of misclassifying an observation from Π_j as from Π_i and

$$U_{ij}(\boldsymbol{x}) = \log \frac{f_j(\boldsymbol{x})}{f_i(\boldsymbol{x})}, \qquad i = 1,2 \ (\mathrm{i} \neq \mathrm{j})$$
$$= \frac{1}{2} \log \frac{|\boldsymbol{V}_i|}{|\boldsymbol{V}_j|} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{V}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) \ , \tag{4}$$

which is known as quadratic Gaussian Bayes classifier. If the parent populations have the same covariance matrix (i.e., $\boldsymbol{V_1} = \boldsymbol{V_2} = \boldsymbol{V}$), the quadratic classifier reduces to the Gaussian linear Bayes classifier as follows

$$U_{ij}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{V}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) - \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i)^T \boldsymbol{V}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_j) \tag{5}$$

This is also simply known as linear discriminant analyzers (LDA). It is also coinciding with Fisher's linear discriminant analyzers (FLDA). If $\boldsymbol{x}$ originate from Π_i, then U_{ij} is distributed as $N(\Delta_{ij}^2 / 2, \Delta_{ij}^2)$, where, $\Delta_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. The classification regions R_j, (j=1, 2,…,m) as defined in (Eq. 3) minimize the expected cost of misclassification (ECM) defined by

$$ECM = q_1 C(2 \mid 1) \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + q_2 C(1 \mid 2) \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \tag{6}$$

When $C(j|i) = 1$ for $i \neq j$, the ECM reduces to the total probability of misclassification (TPM) and classification results can be obtained based on posterior probabilities (Eq. 2) only. If the mixing proportions q_j's and the cost of misclassifications $C(j|i)$ for $i \neq j$ are unknown, we can roughly assume q_j=1/m for all j and $C(j|i) = 1$ which implies $\lambda_{ij} = 0$ for $j \neq i$ in (Eq. 3). Also the value of the threshold λ_{ij} can be determined to sufficient accuracy by a trial-and-error method using the asymptotic distribution of $U_{ij}(\boldsymbol{x})$. For detail discussion, please see [26]. The maximum likelihood estimators (MLEs) for the Gaussian parameters $\boldsymbol{\mu_j}$ and $\boldsymbol{V_j}$ for all j are as follows

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{\mathrm{k=1}}^{\mathrm{n_j}} \boldsymbol{x_{jk}}$$
$$\hat{\boldsymbol{V}}_j = \frac{1}{n_j} \sum_{\mathrm{k=1}}^{\mathrm{n_j}} (\boldsymbol{x_{jk}} - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x_{jk}} - \hat{\boldsymbol{\mu}}_j)^T \tag{7}$$

If we assume that the parent populations have the same covariance matrices (i.e., $\boldsymbol{V_1} = \boldsymbol{V_2} = \boldsymbol{V}$), then the estimated covariance matrices $\hat{\boldsymbol{V}}_j$ are combined (pooled) to derive a single estimate of $\boldsymbol{V}$ that is used in (Eq. 5) as follows.

$$\hat{\boldsymbol{V}}_{pooled} = \frac{1}{n} \sum_{\mathrm{j=1}}^{2} n_j \hat{\boldsymbol{V}}_j. \tag{8}$$

We observe that quadratic (Eq.4) and linear (Eq.5) Bayes classifiers need to compute the inverse of group covariance matrices and pooled covariance matrix, respectively. So number of

genes/variables '*p*' should be smaller than min(n_1,n_2) and max(n_1,n_2) for quadratic and linear Bayes classifier, respectively to solve the problem of matrix inversion. However, in the gene expression dataset the number of genes '*p*' usually very much larger than both sample size n_1 and n_2. So binary cancer classification is difficult by the linear Bayes classifier using all genes as feature variables, though it is one of the most popular statistical classifier. To overcome this problem, we consider only the important features/genes to train the Bayes classifier, since classification does not depends on all feature variables. There are some discussions of feature variable selection for classification in the literature [11, 14, 22-28]. However, they did not consider the problems of outliers in the dataset. So the existing approaches sometimes produce misleading results. Therefore, in this paper we consider outlier modification and relevant gene selection for cancer classification with maximum accuracy using linear Bayes classifier as discussed in the next subsection 2.1.

2.1 Outlier Modification and Gene Selection for Binary Cancer classification using Gaussian Linear Bayes Classifier (Proposed)

Microarray gene expression datasets are often contaminated by outliers due to several steps involve in the data generating process from hybridization to image analysis. There are two types of statistical approaches for data analysis in presence outliers [29,30]. Type-I is the application of robust algorithms on the contaminated datasets and Type-II is the application of classical algorithms on the modified/reduced datasets obtained by removing outliers from the original contaminated datasets or replacing outlying components with the appropriate values. In this paper, we propose Type-II approaches by replacing outlying components with the appropriate values and application of classical *t*-test to select top DE genes for the linear Bayes classifier as follows:

(i) Select one of several approaches for detection of univariate outliers [31,32]. In our current problem, we consider inter-quartile range (IQR) rule for identification of outliers. If Q1 and Q3 are the lower and upper quartiles respectively, then IQR is defined by IQR= $Q_3 - Q_1$. Then an observation is said to be an outlier if it does not belongs to the interval $[Q_1 - \beta \times \text{IQR}, \; Q_3 + \beta \times \text{IQR}]$ for some non-negative constant β, where we usually use β =1.5.

(ii) Check the existence outliers for each gene from both patients groups (normal/cancer) separately from the training dataset using IQR rule. If outlier exist, replace outliers by their respective group medians.

(iii) Apply *t*-test in the modified training dataset to identify differentially expressed (DE) genes. Then arrange the genes from top DE genes by ranking the *p*-values of *t*-test.

(iv) Select top $k < \max(n_1, n_2)$ genes out of *p* genes from both patterns of DE genes (upregulated/downregulated) and estimate the linear Bayes classifier using the expressions of these top *k* genes.

(v) To check the existence of outlying component in the test data vector '$\boldsymbol{x}$' with respect to the top *k* genes using IQR rule, compute $\boldsymbol{d}_j = abs(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j), \; j = 1, 2.$ If there is no any outlying component in $\boldsymbol{d}_j$, then the test data vector '$\boldsymbol{x}$' is said to be usual/uncontaminated. Otherwise, it is said to be unusual/contaminated.

(vi) If the test data vector '$\boldsymbol{x}$' is not contaminated by outliers, we compute Gaussian Bayes classifier as defined in (Eq. 5) using the MLEs $\{\hat{\boldsymbol{\mu}}_j, \hat{V}\}$ of $\{\boldsymbol{\mu}_j, \boldsymbol{V}\}$, based on the modified training dataset, since the training dataset also might be contaminated by outliers. If the test data vector $\boldsymbol{x}$ is detected as a contaminated/outlying vector, our proposal is to classify it as follows. Arrange the values of $\boldsymbol{d}_j$ from (iv) in ascending order such that $d_{j(1)} \leq d_{j(2)} \leq \ldots d_{j(k)}$ for both *j=1,2*. Compute $S_j = \sum_{i=1}^{r} d_{j(i)}$ for both *j=1,2*, where $r < k$. Then classify the contaminated data vector $\boldsymbol{x}$ to the *j*-th class if

$$j = \underset{j' \in \{1,2\}}{\operatorname{argmin}} S_{j'}$$

Also classification result of $\boldsymbol{x}$ can be obtained using the posterior probability (Eq. 2) by replacing *k-r* outlying values of $\boldsymbol{x}$ corresponding to the largest *k-r* values of $\boldsymbol{d}_j$ by the corresponding estimated mean values from the mean vector $\boldsymbol{\mu}_j$. This approach can tolerate up to (*k-r*) outlying values of the data vector $\boldsymbol{x}$. For example, if we choose *r = k/2*, then this approach can tolerate up to *k-r = k/2* outlying values in the data vector $\boldsymbol{x}$.

3. SIMULATED AND REAL GENE EXPRESSION DATA ANALYSIS

To investigate the performance of outlier modification for gene selection and cancer classification by the t-test and the Gaussian linear Bayes classifier respectively, we analyzed both simulated and real gene expression datasets in both absence and presence of outliers .

3.1 Simulated Gene Expression Data Analysis

We generated three types artificial gene expression datasets using the data generating model as described in figure 1 with $\mu =$ 0.5, 1.0 and 2.0 and common variance $\sigma^2 = 1$.

Gene Expression Patterns (Number of Genes)	**Normal** ($\mathbf{n_1}$)	**Cancer** ($\mathbf{n_2}$)
Pattern 1 ($\boldsymbol{p_1}$)	$\boldsymbol{N}(+\boldsymbol{\mu}, \boldsymbol{\sigma^2})$	$\boldsymbol{N}(-\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
Pattern 2 ($\boldsymbol{p_2}$)	$\boldsymbol{N}(-\boldsymbol{\mu}, \boldsymbol{\sigma^2})$	$\boldsymbol{N}(+\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
Pattern 3 ($\boldsymbol{p_3}$)	$\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2})$	$\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2})$

FIGURE 1: Schematic drawing of artificially generated gene expression data. The generated dataset will consist of three patterns of gene expressions. Pattern 1 contains p_1 genes. For each gene, n_1 expressions are generated for normal patients with Gaussian density $N(+\mu, \sigma^2)$ and n_2 expressions for cancer patients with density $N(-\mu, \sigma^2)$. Pattern 2 contains p_2 genes, where n_1 expressions are generated for normal patients with density $N(-\mu, \sigma^2)$ and n_2 expression for cancer patients with density $N(+\mu, \sigma^2)$, for each gene. Pattern 3 consists of p_3 genes, where expressions of each gene are generated for both normal and cancer patients with density $N(0, \sigma^2)$.

Each dataset contains p =100 genes of which p_1=10 DE genes of pattern 1, p_2=10 DE genes of pattern 2 and p_3 = 80 EE genes of pattern 3. Each gene is generated with N =408 sample expressions of which N_1=204 expressions are generated from normal patients and N_2=204 expressions are generated from cancer patients. Then we construct training and test datasets from each dataset by choosing $n_1 = N_1/2$=102 random samples from N_1 =204 normal patients and $n_2 = N_2/2$=102 random samples from N_2 =204 cancer patients for the test dataset. The rest of the patients belong to the training dataset. Then we contaminated 5%-10% patients with 30% genes in both training and test datasets by outliers. Then we computed both training and test MER for both the classical and proposed methods with respect to the increasing number of top DE genes as feature variables. We repeated this procedure 200 times and calculate the average of training and test MER. Figures 2 (a1, b1, c1 and d1) represent the average training and test MER against the number of top DE genes with μ *=0.5, 1.0 and 2.0* respectively for the classical method. Figures 2 (a2, b2, c2 and d2) represent the average training and test MER against the number of top DE genes with μ*= 0.5, 1.0 and 2.0 re*spectively for the proposed method. It is observed that

the proposed method produces much smaller MER than the classical method with top DE genes for each case of μ= *0.5, 1.0* and *2.0*. It is also observed that the proposed method produces smallest test MER (almost close to 0%) with the numbers 20, 10 and 2 of top DE gens for the case of μ= *0.5, 1.0* and *2.0,* respectively. So we can consider only top two DE genes for binary cancer classification using Gaussian linear Bayes classifier when sample size is small (for example, n1=n2=3) when μ grater than *2*. Therefore, this result suggest to use $k < \max(n_1, n_2)$ top DE genes to overcome the inverse problem of Gaussian linear Bayes classifier.

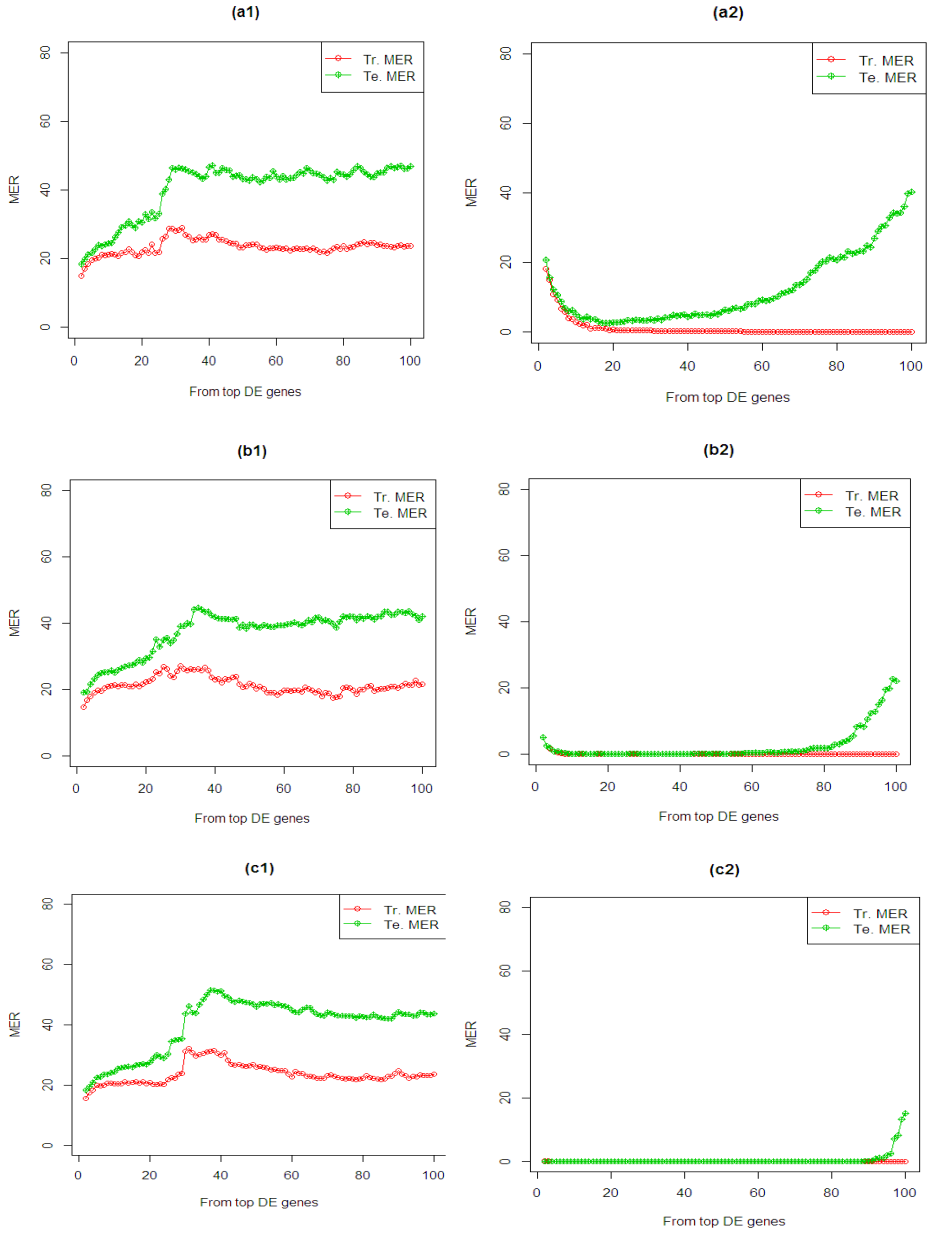

FIGURE 2: Plots of average MER against the number of top DE genes in presence of 5%-10% contaminated patient samples with 30% genes in each of training and test 200 datasets. Datasets are generated using the data generating model described in figure 1. (a1-a2) Average MER for classical (without outlier modification) and proposed (with outlier modification) methods respectively, with μ=0.5 and σ^2=1. (b1-b2) Average MER for classical and proposed methods respectively, with μ=1.0 and σ^2=1. (c1-c2) Average MER for classical and proposed methods respectively, with μ=2.0 and σ^2=1.

Again we generated 3 types of artificial gene expression datasets using μ = 0.5, 1.0 and 2.0 and common variance σ^2 = 1 as before in the data generating model. Each dataset contains p =1000 genes of which p_1=50 DE genes of pattern 1, p_2=50 DE genes of pattern 2 and p_3 = 900 EE genes of pattern 3. Each gene is generated with N *=12, 20, 40, 60 and 80* sample expressions of which N_1=6, 10, 20, 30, 40 expressions are generated from normal patients and N_2=6, 10, 20, 30, 40 expressions are generated from cancer patients respectively for each cases of μ =0*.5,* 1.0 and 2.0. Then we construct training and test datasets from each dataset by choosing $n_1 = N_1/2$=3, 5, 10, 15, 20 random samples from N_1 normal patients and $n_2 = N_2/2$=3, 5, 10, 15, 20 random samples from N_2 cancer patients for the test dataset, respectively. The rest of the patients for each case belong to the respective training dataset. Then we contaminated 5%-10% patients with 30% genes in both training and test datasets by outliers for each case. Then we computed both training and test MER for both the classical and proposed methods for each case. We repeated this procedure 200 times and calculate the average of training and test MER for each case. Table 1 represent the average values of training and test MER with sample sizes (n1, n2): (3, 3), (5, 5), (10,10), (15, 15) and (20, 20) for each value of μ *= 0.5, 1.0 and 2.0,* respectively. It is obviously seen that both training and test MER becomes smaller with $k < \max(n_1, n_2)$ top DE genes for the proposed method than the classical method using the Gaussian linear Bayes classifier.

Sample Size (n_1, n_2)	**Number of Top DE genes (k)**	**μ=0.5 and σ^2 =1**			
		Classical Method		**Proposed Method**	
		MER.Tr	**MER.Test**	**MER.Tr**	**MER.Test**
(3, 3)	2	8.3333	18.3333	0.83333	5.0000
(5, 5)	4	4.0000	18.5000	0.0000	4.5000
(10, 10)	9	0.7500	7.7500	0.0000	4.2500
(15, 15)	14	0.5000	5.3333	0.0000	3.6666
(20, 20)	19	1.6250	10.1250	0.0000	2.0000
Sample Size (n_1, n_2)	**Number of Top DE genes (k)**	**μ=1.0 and σ^2 =1**			
		Classical Method		**Proposed Method**	
		MER.Tr	**MER.Test**	**MER.Tr**	**MER.Test**
(3, 3)	2	4.6666	12.0000	0.0000	1.8333
(5, 5)	4	2.0000	8.4000	0.0000	1.5000
(10, 10)	9	0.8000	2.8000	0.0000	0.2500
(15, 15)	14	0.2666	3.6000	0.0000	0.1666
(20, 20)	19	2.0000	6.5000	0.0000	0.0000
Sample Size (n_1, n_2)	**Number of Top DE genes (k)**	**μ=2.0 and σ^2 =1**			
		Classical Method		**Proposed Method**	
		MER.Tr	**MER.Test**	**MER.Tr**	**MER.Test**
(3, 3)	2	5.0000	10.8333	0.0000	0.2500
(5, 5)	4	2.5000	9.0000	0.0000	0.0500
(10, 10)	9	0.0000	3.0000	0.0000	0.0000
(15, 15)	14	0.1666	3.1666	0.0000	0.0000
(20, 20)	19	1.6250	7.1250	0.0000	0.0000

TABLE 1: Training and test average MER for k= top DE genes such that $k < \max(n_1, n_2)$ with each case of μ =0.5, 1.0 and 2.0, respectively for both classical (without outlier modification) and proposed (with outlier modification) methods.

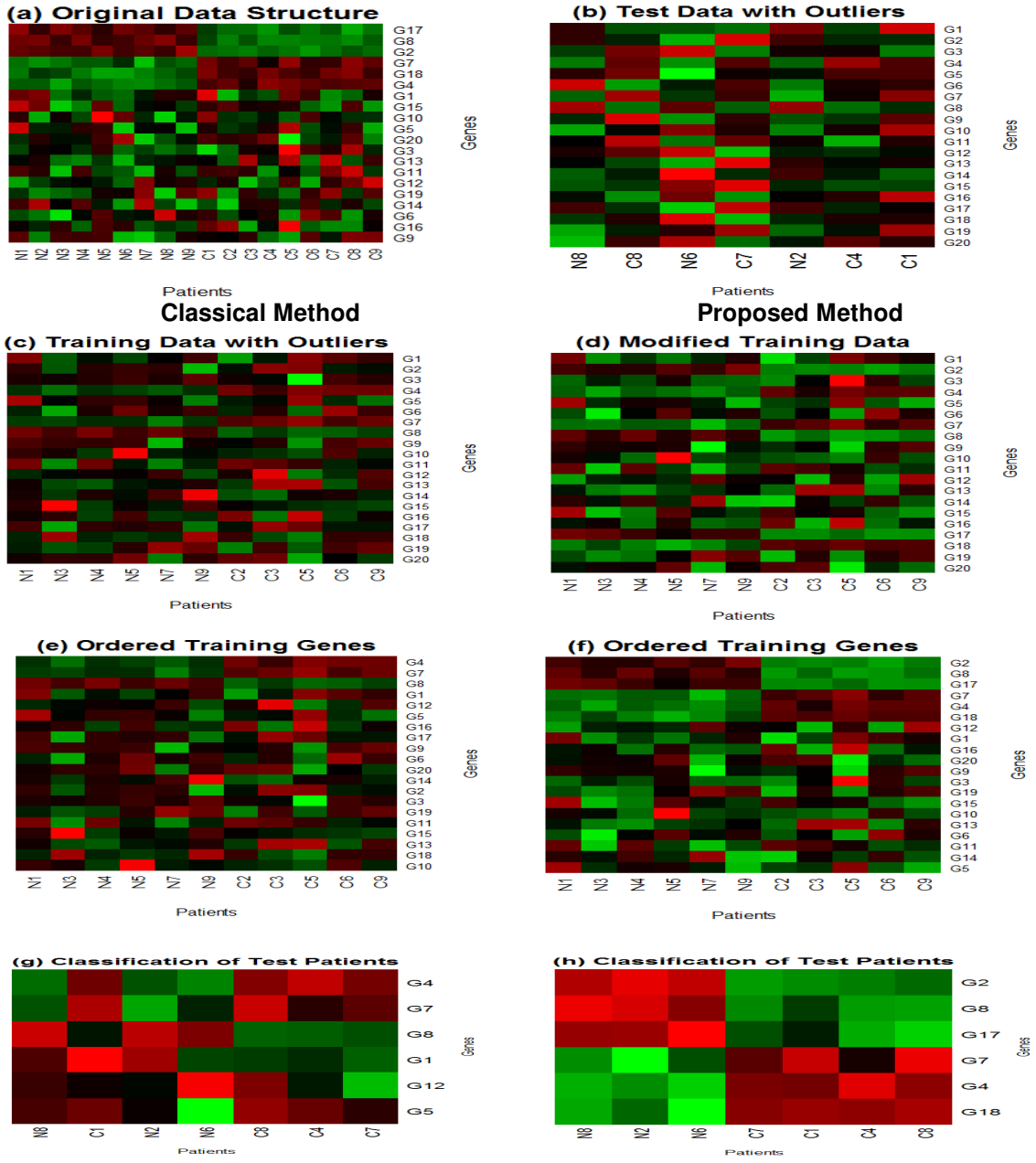

FIGURE 3: Simulated gene expression data analysis for a comparison between classical and proposed methods. Dataset is generated using the data generating model described in figure 1 with parameters μ=0.5 and σ^2=1. (a) Unobservable original structure of gene expressions with normal and cancer patients, where normal and cancer patients are labeled with N_i and C_j respectively for $i=1, 2,,\ldots, n_1$ and $j=1,2,\ldots,n_2$. (b) Test data obtained from (a) with outliers. (c) Training data obtained from (a) with outliers. (d) Modified training data obtained from (c) by replacing outlying observation for each gene by their respective group (normal/cancer) median values. (e) Ordered training DE genes obtained by ranking the *p*-values of *t*-test for each gene in the training dataset. (f) Ordered training DE genes obtained by ranking the *p*-values of *t*-test for each gene in the modified training dataset. (g) Classification of test patients using 6 top DE genes obtained from (e) with outlier (classical method). (h) Classification of test patients using 6 top DE genes obtained from (f) by outlier modification (proposed method).

For motivation of the proposed approach, we have generated a simple artificial gene expression dataset with μ = 2.0 and variance σ^2 = 1 as before. This dataset contains p =20 genes of which p_1=3 DE genes of pattern 1, p_2=3 DE genes of pattern 2 and p_3 = 14 EE genes of pattern 3. Each gene is generated with N =18 sample expressions of which N_1=9 expressions are generated from normal patients and N_2=9 expressions are generated from cancer patients. This dataset is visualized using Figure 3(a). Then we constructed training and test datasets from that dataset by choosing n_1=3 normal patients randomly from N_1=9 normal patients and n_2 = 3 cancer patients randomly from N_2=9 cancer patients for the test dataset. The rest of the patients belong to the training dataset. Then we contaminated 10% patients with 30% genes in both training and test datasets by outliers. Both test and training datasets are visualized using Figures 3(b) and 3(c) respectively. Then we detect outlying components for each gene in each group using IQR rule in the training dataset and replace outlying observations for each gene by their respective group (normal/cancer) median values. Then we call this dataset as the modified training dataset (Figure 3(b)). Then, we apply t-test on both the training datasets to obtain the ordered top DE genes using p-values. Figures 3(e-f) visualize the ordered genes for both the training datasets, respectively. Then we select $k = 6 < \max(n_1, n_2)$ top DE genes from each of the ordered training dataset to construct the Gaussian Bayes classifier. Then we select those k=6 top DE genes from the test dataset also for the classification of test patients. Figures 3 (g-h) show the patient classification results by the classical and proposed method respectively. We observe that the patient C1 is not correctly classified based on the 6 top DE genes selected by the classical method, while all test patients are correctly classified based on the 6 top DE genes selected by the proposed method. So, we may conclude that the Bayes classifier is controlled by only the top DE genes. From this point of view, we can overcome the inverse problem of variance-covariance matrix for the Gaussian Bayes classifier by using the small number $k = 5 < \max(n_1, n_2)$ of top DE genes.

3.2 Example of Real Gene Expression Data Analysis

To investigate the performance of the proposed method with the real gene expression datasets, we consider two publicly available microarray gene expression datasets (i) head and neck cancer dataset which is previously analyzed in [26, 27] and (ii) Colon cancer dataset which is previously analyzed in [28, 29]. These datasets can be downloaded from the web links http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6631 and in the R-package 'plsgenomics', respectively. The head and neck cancer dataset consists of p=12625 genes with N=44 samples having N_1=22 normal and N_2=22 cancer individuals/patients. We constructed 200 bootstrap training and test datasets from this dataset by choosing n_1=10 random samples from N_1 =22 normal patients and n_2 = 10 random samples from N_2 =22 cancer patients with replacement for each training dataset and the rest of the patients belong to the respective each test dataset. Then we computed both training and test average MER using both classical and proposed methods with respect to top 2, 4, 6, 9 genes respectively. It is seen that both method produces smaller test average MER with top 6 genes, where the proposed method produces smallest test MER (see Table 2). Then we analyzed the colon cancer dataset. This dataset contains p=2000 genes with N=62 samples having N_1=22 normal and N_2=40 cancer individuals/patients. We constructed 200 bootstrap training and test datasets as before from this colon cancer dataset by choosing n_1=10 random samples from N_1 =22 normal patients and n_2 = 10 random samples from N_2 =40 cancer patients with replacement for each of the training dataset. The rest of the patients belong to the respective each test dataset. Then we computed both training and test average MER using both classical and proposed methods with respect to top 2, 4, 6, 9 genes respectively as before. It is seen that both method produces smaller test MER with top 4 genes, where the proposed method produces smallest test MER (see Table 2) in this case also.

Sample Size (n_1, n_2)	Number of Top DE Gene (k)	Head and Neck Cancer Gene Expression Dataset		Colon Cancer Gene Expression Dataset	
		MER.Training	MER.Test	MER.Training	MER.Test
(10, 10)	2	3.0000 (3.1000)	6.8333 (6.0563)	12.3333 (12.7863)	13.9583 (12.3422)
(10, 10)	4	1.0000 (1.0045)	6.6500 (6.1040)	8.4444 (7.8988)	**8.8333 (6.0666)**
(10, 10)	6	0.0000 (0.0065)	**2.0833 (1.9034)**	7.9333 (7.1223)	12.0833 (10.0234)
(10, 10)	9	0.0000 (0.0000)	7.1111 (6.0111)	2.1333 (3.7677)	14.9206 (12.0123)

TABLE 2: Bootstrap training and test average MER for the classical (without outlier modification) and the proposed (with outlier modification) methods with respect to top k=2, 4, 6 and 9 DE genes for the real (i) head and neck cancer (ii) colon cancer gene expression datasets. The MER results with first bracket (.) indicate the results of the proposed method.

4. CONCLUSION

Cancer classification using gene expression data is one of the major research areas in the medical field. Accurate cancer classification has great value to cancer treatment and drug discovery. There exist some computational algorithms for cancer classification. However, most of them are sensitive to outlying gene expression data. To overcome this problem, we proposed outlier modification based linear Bayes classifier which is one of the most popular approaches for classification. Our proposed method has shown that it is a simple efficient yet accurate approach for binary cancer classification problems. The precision of proposed approach is comparable to others such as SVM, however, the time consumption of this approach is much less than other approaches. Support vector machines (Vapnik, 1998) cannot be easily extended to multiclass cancer classification problem because the elegant theory behind the use of large margin hyperplanes, whereas, our proposed approach can be easily used for multiclass classification that we will show in our next paper. However, it has another problem for cancer classification using the modified gene expression data due to the inverse problem of its covariance matrix in presence of large number of gene variables with small number of cancer patients/samples in the training dataset. To overcome these problems, we propose few top differentially expressed (DE) genes from both upregulated and downregulated groups for binary cancer classification using linear Bayes classifier. Top DE genes are selected by ranking the *p*-values of *t*-test procedure. The performance of the proposed method is investigated using both simulated and real gene expression datasets. It is observed that the proposed method improves the performance with outlier modifications for binary cancer classification. However, there are several future research directions with this work. In our next project, we would like to extend the proposed method for multiclass cancer/disease classification and compare with other existing methods.

5. REFERENCES

[1] A. Sharma, and K.K. Paliwal. “Cancer classification by gradient LDA technique using microarray gene expression data.” *Data Knowl. Eng.*, vol. 66, pp. 338-347, 2008.

[2] S. Dudoit, J f Fridlyand, T. P Speed. “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.” *Journal of the American Statistical Association*, vol. 97, No. 457, pp. 77-87, Mar. 2002.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, M. Gaasenbeek C. Huard, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.” *Science*, pages 531–537, Oct 1999.

[4] V. Van't, , L.J. Dai, H. Van de, M.J. Vijver and Y.D. He et al. "Gene expression profiling predicts clinical outcome of breast cancer." *Lett. Nature. Nature*, vol. 415, pp. 530-536, 2002.

[5] A. Berns. "Cancer: Gene expression in diagnosis." *Nature*, pages 491–492, Feb 2000.

[6] A. Azuaje. "Interpretation of genome expression patterns: computational challenges and pportu-nities." *IEEE Engineering in Medicine and Biology*, 2000.

[7] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans and J.E. Blumenstock et al. "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma." *Cancer Res.*, vol. 62, pp. 4963-4967, 2002. .

[8] L. Ziaei, A. R. Mehri, M. Salehi. " Application of Artificial Neural Networks in Cancer Classification and Diagnosis Prediction of a Subtype of Lymphoma Based on Gene Expression Profile." *Journal of Research in Medical Sciences*, vol. 11, No. 1, Jan. & Feb. 2006.

[9] S. Lakhani and A. Ashworth. "Microarray and histopathological analysis of tumours: the future the past?" *Nature Reviews Cancer*, pages 151–157, Nov 2001.

[10] D. Nguyen and D. Rocke. "Classification of Acute Leukemia based on DNA Microarray Gene Expressions using Partial Least Squares." *Kluwer Academic*, 2002.

[11] I. Guyon, J. Weston, S. Barnhill, M. D., and V. Vapnik. "Gene selection for cancer classification using support vector machines." *Machine Learning*, 2000.

[12] A.C. Tan and D. Gilbert. "Ensemble machine learning on gene expression data for cancer classification." *Applied Bioinform.*, vol. 2, pp. S75-83, 2003.

[13] G. Cong, K.L. Tan, A.K.H. Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski and M. Schummer et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." *Bioinformatics*, vol. 16, pp. 906-914, 2005.

[14] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank and A. Facius et al. "Gene selection from microarray data for cancer classification - a machine learning approach." *Comput. Biol. Chem.*, vol. 29, pp. 37-46, 2005.

[15] A. Statnikov, L. Wang and C. F. Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." *Journal BMC bioinformatics*, 2008.

[16] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. "Tissue classication with gene expression profiles." In Proc. of the Fourth Annual Int. Conf. on Computational Molecular Biology, 2000.

[17] D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander. "Class prediction and discovery using gene expression data." *In Proc. 4th Int. Conf. on Computational Molecular Biology(RECOMB)*, pages 263–272, 2000.

[18] Liang-Tsung Huang. "An integrated method for cancer classification and rule extraction from microarray data." *Journal of Biomedical Science*, 2009.

[19] Kun-Huang Chen**,** Kung-Jeng Wang**,** Min-Lung Tsai**,** Kung-Min Wang**,** Angelia Melani Adrian**,** Wei-Chung Cheng**,** Tzu-Sen Yang**,** Nai-Chia Teng**,** Kuo-Pin Tan and Ku-Shang Chang. "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithms." *BMC Bioinformatics,* 15:49*,* 2014

[20] Desheng Huang, Yu Quan, Miao He and Baosen Zhou. “Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data.” *Journal of Experimental & Clinical Cancer Research, 28:149, 2009*

[21] Ubharup Guha, Yuan Ji and Veerabhadran Baladandayuthapani. “Bayesian Disease Classification Using Copy Number Data.” *Cancer Informatics*, vol. 13 (S2), pp. 83–91, 2014.

[22] Sharma, A., C.H. Koh, S. Imoto and S. Miyano. “Strategy of finding optimal number of features on gene expression data.” *Elect. Lett.*, vol. 47, pp. 480-482, 2011a.

[23] H.A.L. Thi, V.V. Nguyen and S. Ouchani. “Gene selection for cancer classification using DCA.” *Adv. Data Min. Appli.,* vol. 5139, pp. 62-72, 2008.

[24] H. Rattikorn, K. Phongphun, “Tumor classification ranking from microarray data.” *BMC genomics journal*, vol. 9, pp. s21, September 2008.

[25] T.W. Anderson. An Introduction to Multivariate Statistical Analysis, *Wiley Interscience*, 2003.

[26] Liu,H., et al. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform., 13, 51–60.

[27] Wu,B., et al. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19, 1636–1643.

[28] Jafari,P. and Azuaje,F. (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, Vol. **6**.

[29] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel. “Robust Statistics: The Approach Based on Influence Functions.” *John Wiley and Sons*: New York, 1986.

[30] P. J. Huber. Robust Statistics. *John Wiley and Sons*: New York, 2004

[31] P. J. Rousseeuw and A. M. Leroy. Robust Regression and Outlier Detection. Wiley: New York, 1987.

[32] A. Bharathi, A. M. Natarajan. “Cancer Classification of Bioinformatics data using ANOVA” *International Journal of Computer Theory and Engineering*, Vol. 2, No. 3, June, 2010.

[33] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data” *Journal of the American Statistical Association,* Vol. 97, No. 457, Applications and Case Studies, March 2002.