Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Research Article
Volume 1 Issue 2 - April 2017
DOI: 10.19080/BBOAJ.2017.01.555557

Biostat Biometrics Open Acc J

Feature Selection for Cancer Classification Using Microarray Gene Expression Data

Wenyan Zhong, Xuewen Lu and Jingjing Wu*

Department of Mathematics and Statistics, University of Calgary, Canada

Submission: November 12, 2016; **Published:** April 06, 2017

***Corresponding author:** Jingjing Wu, Department of Mathematics and Statistics, University of Calgary, Canada, Tel: 1-403220-6303; Fax: 1-403282-515; Email: jinwu@ucalgary.ca

Abstract

The DNA microarray technology enables us to measure the expression levels of thousands of genes simultaneously, providing great chance for cancer diagnosis and prognosis. The number of genes often exceeds tens of thousands, whereas the number of subjects available is often no more than a hundred. Therefore, it is necessary and important to perform gene selection for classification purpose. A good subset of discriminative genes can improve prediction accuracy of classifiers and save computational cost with reduced dimension of data. In this paper, we use data on microarray gene expression level to determine marker genes that are relevant to a type of cancer. We investigate a distance-based feature selection method for two-group classification problem. In order to select marker genes, the Bhattacharyya distance is implemented to measure the dissimilarity in gene expression levels between groups. We use the support vector machine to make classification with use of the selected marker genes. The performance of marker gene selection and classification are illustrated in both simulation studies and two real data analysis.

Keywords: Feature selection; Microarray gene expression data; Classification; Support vector machine

Introduction

The rapid development of microarray technology has given rise to a wealth of statistical studies that aimed at detecting significantly differentially expressed genes. DNA microarray data has been extensively studied. Currently, there already exist a variety of analysis methods for DNA microarray data to deal with different research interests. In general, the data gathered from microarrays broadly prompt two types of questions:

A. Those about variables, such as which genes or subsets of genes are associated with a specific

Phenotype, biological mechanism or outcome

B. Those regarding biological samples, such as what prediction can be made about a specific tissue [1,2].

Most statistical methods, including clustering, easily address the first question. Pattern classifiers, such as support vector machines (SVM) and other machine learning systems, however, are much better for solving the second question. In this paper, we mainly focus on marker gene selection for cancer classification. Here marker gene selection, or more broadly feature selection, belongs to the first type of question, while classification falls into the second type.

In statistics, feature selection, also known as variable selection, is the process of selecting a subset of relevant variables for constructing statistical models. Feature selection techniques are used under the central assumption that the data contain many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are often used in domains where there are many features and comparatively few samples, such as DNA microarray data. It is also useful as part of the data analysis process since it shows which features are important for prediction and how these features are related.

Due to the high-dimension nature of microarray data and their small sample sizes, microarray data impose a great challenge to computational techniques. In order to tackle the difficulties in analyzing microarray data, the apparent need of feature selection methods was realized by researchers; see, e.g. Alon et al. [3], Golub et al. [4], Ross et al. [5] & Ben-Dor et al. [6] among many others. This has led to a recent surge of dimension reduction approaches presented in both bioinformatics and statistics.

According to the literature, it is widely believed that in most microarray gene expression data, only some relevant genes play an important role in classification and the rest are irrelevant to

classification. Thus, feature selection techniques are needed in order to and out those most important ones among all the genes that have been measured. With microarray data, the selection of important genes for classification of different phenotypes, such as cancer types, aims at providing a better understanding of the underlying biological system and improving the prediction performance of classifiers Ramaswamy et al. [7], Tibshirani et al. [8], Guyon et al. [10], Liu et al. [11]. Feature selection techniques could be classified into three types, i.e., filter method, wrapper method and embedded method, depending on how they combine the feature selection with the construction of classification model.

In this paper the classification problem under our consideration is limited to two classes. Without loss of generality, we use (+) and (-) to denote the two classes. The input, as a "pattern", is a vector of p components, called variables or features, measured on each of n subjects. The output is the observed class labels for all n subjects. We use F to denote the p-dimensional feature space. In our situation, the features are expression levels of thousands of different genes. Thus, a given dataset consists of n vectors of features $\{x_1,...,x_k,...,x_n\}$ with corresponding known class labels $\{y_1,...,y_k,...,y_n\}$. Here with indicating kth subject belongs to class(-) and for class(+). The entry for the kth subject is then $\{x_k,y_k\}$. We randomly select a certain proportion of all subjects to form the training dataset. The training data are used to construct a classifier, more specifically, to build a scalar discriminant or decision function D(x) of an input pattern x. New patterns are classified according to the sign of the decision function as

$$D(x) > 0 \Rightarrow x \in class(+)$$

$$D(x) < 0 \Rightarrow x \in class(-)$$

$D(x) = 0$ decision boundary.

Decision functions that are simply weighted sum of the training patterns plus a bias are called linear discriminant functions Guyon et al. [10], In this case, the decision function can be written as

$$D(x) = w.x + b \quad (1)$$

Where w is a weight vector and b is a bias value. A dataset is called linearly separable if a linear discriminant function can separate it without error. Classification methods can be generally summarized into two branches: supervised learning and unsupervised learning. In this paper, we will use the supervised learning where the class labels are known beforehand. With supervised learning methods for gene expression data, various classifiers with promising performance have been constructed. These classifiers include k-nearest neighbor rule (kNN) [4], Fisher Linear Discriminant Analysis (LDA), weighted gene voting [7,8], naive Bayes classifier (NBC) [11], artificial neural networks classification and regression trees (CART) [12] and random forest [13], Many comparative reviews and studies indicate that SVM-based classifiers outperform other methods on most benchmark microarray datasets [13,14], though in general no one classifier uniformly outperforms others.

In this paper, we propose a new method for marker gene selection and cancer classification based on SVM. The newly proposed method is based on Bhattacharyya distance which measures the dissimilarities between categories. We first calculate the Bhattacharyya distance between two classes for each gene and then rank the genes according to their corresponding Bhattacharyya distances. Then we evaluate certain subsets, starting with the top-ranked gene with largest Bhattacharyya distance and progressively adding the next one on the list until all genes are included. At last, we choose, by forward selection method, the final gene subset based on those subsets' individual classification capability with use of SVM. In order to improve the performance of SVM, parameter optimization of SVM is conducted with the final gene subset. Then the SVM is trained with use of the optimized parameters and the final optimal gene subset, which is finally applied to make prediction for testing/validation data.

This paper is organized as follows. In Section 2, we present our newly proposed method for marker gene selection. We first introduce the Bhattacharyya distance to measure dissimilarities between categories and then describe the proposed B/SVM algorithm for marker gene selection. Section 3 contains the numerical results for our proposed method. The simulation studies are given in Section 3.1 while two real data analysis are demonstrated in Section 3.2. The final concluding remarks and future work discussions are presented in Section 4.

Bhattacharyya distance with SVM classifier (B/SVM)

The proposed method in this paper follows a similar framework to that of filter method in feature selection for two-category classification problems. Particularly, it selects informative genes according to their discriminative power without considering any knowledge of the classifier adopted. With use of DNA microarray data, we propose a new approach for marker gene selection in cancer classification with improved classification accuracy.

Bhattacharyya distance

In statistics, the Bhattacharyya distance is often used to determine the similarity of two probability distributions, discrete or continuous. In classification, it measures the separability of classes and is considered to be more robust and reliable than the Mahalanobis distance, as the latter is a special case of the former when the standard deviations of the two classes are the same. In cases when two classes have similar means but different standard deviations, the Mahalanobis distance would be close to zero whereas the Bhattacharyya distance would grow depending on the difference between the two standard deviations. The Bhattacharyya distance is closely related to the Bhattacharyya

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.

coefficient which is a measure of the amount of overlap between two statistical samples or populations Bhattacharyya [15]. This coefficient can be used to measure the relative closeness of two samples under consideration. For discrete probability distributions p and q over the same domain X, the Bhattacharyya distance is defined as

$$D_B(p,q) = -\ln(BC(p,q)), \quad (2)$$

Where $BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$ is the Bhattacharya coefficient. When p and q are continuous density functions, the Bhattacharya coefficients defined as

$$BC(p,q) = \int \sqrt{p(x)q(x)}dx. \quad (3)$$

In either case, $0 \le BC \le 1$ and $0 \le D_B \le \infty$. Particularly when the two populations p and q are normal, the Bhattacharyya distance can be calculated by extracting the means and variances of the two distributions or classes; specifically,

$$D_B(p,q) = \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)\right] + \frac{1}{4}\left[\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2}\right], \quad (4)$$

Where μ_p and σ_p^2 are the mean and variance of p and μ_p and σ_p^2 are those of q.

B/SVM algorithm for marker gene selection

The B/SVM algorithm that we propose for marker genes selection is based on the Bhattacharyya distance along with SVM. In this algorithm, the Bhattacharyya distance is used to obtain the ranking list of genes for classification where genes with larger Bhattacharyya distances are suggested to be more discriminative on the differentiation of two categories than those with smaller ones. Then SVM is used to evaluate each subset of genes in terms of classification performance. Eventually, we identify a subset of discriminative genes that provides the lowest classification error rate as marker genes that will be used for the validation data. Denote pi and qi the distribution functions of the ith gene for the two classes respectively, say class(+) and class(-), i= 1,..., p. Then in general, the proposed B/SVM algorithm includes the following four steps:

A. Compute the Bhattacharyya distance given in (4) for the ith gene, i= 1,..., p

B. Order genes according to the magnitude of their Bhattacharyya distances

C. Evaluate subsets of important genes using SVM in terms of classification power, start-ing with the gene(s) with the largest Bhattacharyya distance:

i. Randomly divide the given data into training data and testing data

ii. Use the training data to train the SVM with optimal hyper parameters selected by 10-fold cross-validation criterion or any subset of important genes

iii. Obtain the classification error rate of the trained SVM applied to the testing data

D. Repeat Step 3 progressively by adding the next important gene until all genes are added once. We identify as marker genes the subset of important genes that provides the lowest classification error rate for testing data.

In the first step, we assume for simplicity that for any fixed gene, the expression levels of subjects in the two classes are from two (possibly the same) normal distributions, i.e. we assume that both pi and qi are normal distributions, i= 1,...,p. Based on the data, we can calculate the sample mean and variance for both pi and qi. Then we plug in these sample means and variances into (4) to get the estimated Bhattacharyya distance for each gene. In order to evaluate and compare the e effectiveness of this feature selection approach, the given dataset is divided into two, training set and testing set. The training set is used to train the model while the testing set is used as an independent validation set to evaluate the performance of the model through prediction results.

Numerical Results

We illustrate how well B/SVM performs in marker gene selection through both simulation studies and real data applications. Specifically, we present the results of simulation studies for the proposed B/SVM, supervised weighted kernel clustering/SVM (SWKC/SVM) and the well known SVM with recursive feature elimination (SVM-RFE). The SVM-RFE, a widely used and well developed method proposed by Guyon et al. [10], is proved to be an effective embedded feature selection method. The idea of SWKC/SVM was introduced by Shim et al. [16]. As alter feature selection method, it sets the ranking rule by a weight vector obtained from clustering approach with a distance measurement defined by using Gaussian kernel function. Furthermore, we apply and compare the three methods for the two benchmark datasets, i.e., the colon cancer dataset and leukemia dataset. Both simulation studies and real data applications suggest that the proposed B/SVM competitively in the sense that it produces relatively lower misclassification rate compared with SWKC/SVM and SVM-RFE.

Simulation Studies

To evaluate how well the proposed B/SVM method performs in feature selection for classification, a simulation study is carried out. As in Broberg [17] & Koo et al. [18], the data are simulated from normal distributions assumed as the populations of gene expression levels after log transformation. The means and standard deviations of those normal distributions are given in Table 1. In this table, the first three rows, designated as the null cases, represent the distributions of irrelevant genes. The last three rows in the table, designated as the significant cases, represent the distributions of marker genes that differentiate

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.

the two groups of subjects. For a particular gene, its population distribution is randomly chosen from either the first three rows or the last three rows, depending on whether it's an irrelevant gene or a marker gene. The numbers of subjects, in another word the sample sizes, in the normal group and the case group are chosen to be 47 and 25, respectively, as in Broberg [17] & Koo et al. [18]. These particularly chosen sample sizes are intended to match those of the real leukemia data.

Table 1: Normal distributions used to generate simulated data.

Genes Expression Levels	Mean 1	Standard Deviation 1	Mean 2	Standard Deviation 2
Null cases	-8	0.2	-8	0.2
	-10	0.4	-10	0.4
	-12	1	-12	1
Significant cases	-6	0.1	-6	1
	-8	0.2	-8.5	0.2
	-10	0.4	-11	0.7

In the simulation, we consider 1000 genes in total among which 1%, i.e., 10 genes, are assumed differentially expressed. As a result, the simulated data for a normal subject consists of 1000 genes randomly selected from the null cases, while the simulated data for a case subject contains 10 differentially expressed genes randomly selected from the significant cases and the rest 990 genes are all from the null cases. Note that the data are generated gene by gene across all 72 subjects, rather than subject by subject with all 1000 genes [19].

The simulated data are randomly split into a training set of 67% (two-third) and an independent testing set of 33% (one-third) of the whole dataset. We will use this same ratio for splitting data into training and testing sets throughout this paper. Based on the training set, the Bhattacharyya distance is calculated for each gene according to which all the genes are ranked in decreasing order. Then the SVM classifier is built based on the gene rankings and classification error rate is calculated for the testing set. Here, the SVM is used as a method of marker gene selection. The simulation procedure is repeated 50 times and large number of repetitions gives very similar results. We compare our proposed B/SVM method with the SWKC/SVM and SVM-RFE. For comparison purpose, we use the following indexes: average number of genes selected, average number of true marker genes selected, average recovery rate, and average misclassification rate of classification models based on the constructed classifier.

In Table 2, "Average number of genes selected" denotes the averaged number of genes selected out of 1000 over 50 repetitions; "Average number of true marker genes selected" denotes the averaged number of genes, among those selected, that are designated as significant; "Average recovery rate" denotes the averaged ratio of the number of true marker genes selected to the total number of genes selected; "Average misclassification rate" denotes the averaged proportion of misclassified subjects calculated as the ratio of the number of prediction errors to the size of testing set.

Table 2: Comparison of the proposed B/SVM with SWKC and SVM-RFE using simulated data.

Indexes	B/SVM	SWKC/SVM	SVM-RFE
Average number of genes selected	6.9	8.6	3.5
Average number of true marker genes selected	6	2.8	3
Average recovery rate (%)	95.7	83.9	94.1
Average misclassification rate (%)	1.1	7.3	2

Compared with SVM-RFE and SWKC/SVM, the proposed B/SVM returns a relatively higher average recovery rate of 95.7%. This indicates that B/SVM has higher power of detecting marker genes than the other two methods. B/SVM also gives the lowest misclassification rate among the three methods. SVM-RFE selects the smallest size, 3.5 on average, of gene subset among the three, at the price of expensive computation time. B/SVM as alter method takes much less computing time than both SVM-RFE and SWKC/SVM. Note that the smallest subset size doesn't mean the best since we know this simulation is designed with 10 truly differentially expressed genes. SVM-RFE can only identify 3 of the 10 while B/SVM identifies 6 on average. In this sense, SWKC/SVM performs the first, selecting the largest size of gene subset but identifying the fewest true marker genes. These observations suggest that the proposed B/SVM method outperforms both SWKC/SVM and SVM-RFE.

The numerical analysis given in this study is conducted with the use of R programming. The R package, e1071, has provisions for performing C-classification which corresponds to a soft-margin classifier using a Gaussian kernel or linear kernel. In this simulation study, 10-fold cross-validation on each training set was used to estimate the cost parameter C in SVM classifier.

Real Data Applications

In this section, we analyze two real microarray datasets

A. Colon cancer dataset

B. Leukemia dataset

to demonstrate the application of our newly proposed B/SVM method. These two datasets are publicly available and all the original observations in the two datasets were transformed to the base 10 log scale. Table 3 gives the description of the two benchmark microarray datasets.

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.

Table 3: Description of the two benchmark microarray datasets.

Dataset	Sample	Gene	Class	Publication
Colon cancer	62	2000	2	Alon et al. [2]
Leukemia	72	3571	2	Golub et al. [3]

To check the performance of the newly proposed B/SVM method, each dataset is randomly split into a training set and a testing set 50 times and the results are the average over the 50 repetitions. Since the true marker genes of the real data are unknown, average number of genes selected, average number of misclassified samples in testing set and average misclassification rate for testing set are used for comparing the proposed B/SVM with SWKC/SVM and SVM-RFE

Colon cancer data

The colon data contains 62 samples with 2000 genes Alon et al. [1]. Among the 62 samples, 40 are tumor tissues and the rest 22 are normal ones. In each of the 50 splitting of data, we randomly select 20 samples (33%) of the total 62 to form a testing set and the rest 42 samples (67%) form a training set. The training set is used to train a classifier and the testing set is used to evaluate the classification performance of trained classifier. Table 4 gives the analysis results of all the three methods for this colon cancer data. It presents average number of the genes selected, average number of misclassifications in testing set and average misclassification rate. Note that the misclassification rate is simply the number of misclassifications in testing set divided by the size of testing set 20. From Table 4 we can see that the proposed B/SVM performs best among the three in terms of either index. Comparatively, SWKC/SVM performs first among the three in the sense that it uses the largest gene subset for classification that nevertheless results in the highest misclassification rate. This is consistent with our observation on the simulated data. For this data we can infer that, on average, the trained classifier by B/SVM will achieve 90.5% classification accuracy with use of only the top 6 ranked genes out of the total 2000

Table 4: Analysis result of the colon cancer microarray data.

Indexes	B/SVM	SWKC/SVM	SVM-RFE
Average number of genes selected	6.36	15.42	7.62
Average number of misclassifications in testing set	1.9	2.96	2.3
Average misclassification rate (%)	9.5	14.8	11.5

Leukemia data

The leukemia data consists of 72 samples with 3571 genes Golub et al. [8] It is a gene expression data with a binary response indicating two types of leukemia: ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). Among the 72 patients, 47 are ALL patients and 25 are AML patients. The classifier built for this data aims at differentiating these two types of leukemia patient For each data splitting, we randomly select 24 samples (33%) of the total 72 to form a testing set and the rest 48 (67%) to form a training set. Table 5 presents the analysis results of the three methods for this leukemia data. From Table 5 we can see that the three methods are very competitive for this leukemia data in terms of misclassification rate or equivalently the number of misclassifications in testing set, with SVM-RFE slightly better than B/SVM followed by SWKC/SVM. For this leukemia data, B/SVM selects the smallest gene subset. SWKC/SVM performs the first among the three methods in the sense that it uses the largest gene subset for classification that nevertheless results in the highest misclassification rate. For this data we can infer that, on average, trained classifier by B/SVM will achieve 96.9% classification accuracy with use of the top 9 or 10 ranked genes out of the total 3571; in another word, there is only about one patient out of the total 24 in testing set that will be misclassified.

Table 5: Analysis result of the leukemia microarray data.

Indexes	B/SVM	SWKC/SVM	SVM-RFE
Average number of genes selected	9.54	13.48	9.68
Average number of misclassifications in testing set	0.74	0.78	0.62
Average misclassification rate (%)	3.08	3.25	2.58

As a summary for both datasets, Table 4 & 5 shows that B/SVM always selects a relatively smaller gene subset for classification than SVM-RFE and SWKC/SVM. Especially, the performance of B/SVM for the colon cancer dataset is outstanding. With the number of genes increases, the advantage of computational efficiency of B/SVM is impressive.

Conclusion and Discussion

Many DNA microarray data in practice are represented in form of extremely high dimensional vectors or matrices which brings great challenges in both data mining and further processing. High dimensionality not only increases the learning cost but also deteriorates the learning performance, known as the “curse of dimensionality”. Therefore, dimension reduction has attracted great attentions in pattern recognition, machine learning and their applications such as microarray data analysis. Under this framework, this paper focuses on dimension reduction, more specifically, feature selection in DNA microarray data analysis. We propose a new gene selection method for classification based on SVMs. In the proposed method, we first rank all the genes ac-cording to the magnitude of their Bhattacharyya distances between the two specified classes. Then the optimal gene subset is selected as the one which achieves the lowest misclassification rate in the constructed SVMs following a forward selection algorithm. Afterwards, the 10-fold cross-validation is applied to

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.

find the optimal parameters for SVM with the final optimal gene subset. As a result, the classification model is trained and built. Finally, the classification model is evaluated by its prediction performance for testing set.

We compare the performance of our newly proposed B/SVM method with that of SVM-RFE and SWKC/SVM. The simulation studies suggest that the proposed B/SVM method outperforms the other two in terms of small average misclassification rate (1.1%) and high average recovery rate (95.7%). This means that B/SVM appears to be more effective and has more power in finding marker genes than SVM-RFE and SWKC/SVM. B/SVM results in relatively high recovery rate with use of only a very small gene subset selected. The much smaller computational burden is another outstanding advantage of the proposed B/SVM. Based on the simulation results for B/SVM, about 7 genes on average are selected out of the total 1000 as marker genes. The dimension of the original data is dramatically reduced. The analysis results of the colon cancer data and the leukemia data also indicate that the B/SVM algorithm can e effectively reduce the dimension of data and select a small informative gene subset for classification with low misclassification rates.

For feature selection, alter method refers to selecting informative genes according to their discriminative power without considering any knowledge of the classifier adopted, while wrapper method selects the discriminative features dependently on the classifier used. The alter method possesses the advantages of fast computability and capability of dealing with large datasets, but lacks the ability of finding optimal feature subset. Wrapper method can be expected to have good performance, but it is difficult to be scaled to large datasets because of the expensive computation cost. Embedded method can be treated as a special case of wrapper method when feature selection space is exactly the same as the hypothesis space of a classifier. The proposed B/SVM, classified as alter method, has the drawbacks shared by all alter methods. In another word, B/SVM treats features independently, with the dependency and interaction among features ignored, and fails to take into account the interaction with classifier.

Our study shows that the proposed B/SVM algorithm is very promising in marker gene selection and cancer classification. For the proposed B/SVM, improvements could be obtained in our future work. Firstly, we can extend the B/SVM algorithm to the more general multi class classification. Secondly, other classifiers than SVM can be used to obtain potentially more reliable and more precise models for cancer classification. We use SVM to select the optimal feature subset simply for the sake of a reasonable comparison with the other two SVM-based methods SVM-RFE and SWKC/SVM. Lastly, we can relax the normal assumption imposed on the distributions of gene expression levels with use of nonparametric kernel density estimation.

Acknowledgements

The authors acknowledge with gratitude the support for this research via Discovery Grants from Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Olshen AB, Jain AN (2002) Deriving quantitative conclusions from microarray ex-pression data. Bioinformatics 18(7): 961-970.
2. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, et al. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer 8(1): 37-49.
3. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tu-mor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12): 6745-6750.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537.
5. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. National Genetics 24(3): 227-235.
6. Ben Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, et al. (2000) Tissue classification with gene expression profiles. J Comput Biol 7(3-4): 559-583.
7. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98(26): 15149-15154.
8. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple can-cer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99(10): 6567-6572.
9. Guyon I, Weston J, Barnhill S (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46(1): 389-422.
10. Liu H, Li J, Wong L (2002) A comparative study on feature selection and classification methods using gene expression pro les and proteomic patterns. Genome Inform 13: 51-60.
11. Breiman L, Friedman J, Stone J, Olshen A (1984) Classification and regression trees. CRC press, Boca Raton, Florida.
12. Breiman L (2001) Random forests. Machine Learning 45: 5-32.
13. Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multi-class classification methods for tissue classification based on gene expression. Bioinformatics 20(15): 2429-2437.
14. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehen-sive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21(5): 631-643.
15. Bhattacharyya A (1943) On a measure of divergence between two statistical populations de ned by their probability distributions. Bulletin of the Calcutta Mathematical Society 35: 99-109.
16. Shim J, Sohn I, Kim S, Lee SW, Green PE, et al. (2009) Selection marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. Computational Statistics and Data Analysis 53(5): 1736-1742.
17. Broberg P (2002) Ranking genes with respect to di erential expression. Genome Biology 3: 1-23.

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.

18. Koo JY, Sohn I, Kim S, Lee JW (2006) Structured polychotomous machine diagnosis of multiple cancer. types using gene expression. Bioinformatics 22(8): 950-990.

19. Rao PBLS (1983) Non-Parametric Functional Estimation, Academic Press, Or-lando, Florida, USA, p. 538.

How to cite this article: Wenyan Z, Xuewen L , Jingjing W. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostat Biometrics Open Acc J. 2017;1(2): 555557. DOI: 10.19080/BBOAJ.2017.01.555557.