International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010
1793-8201

Cancer Classification of Bioinformatics data using ANOVA

A. Bharathi, Dr.A.M.Natarajan

Abstract—**The main aim of this paper is to find the smallest set of genes that can ensure highly accurate classification of cancer from micro array data by using supervised machine learning algorithms. The significance of finding the minimum gene subset is three fold:1) It greatly reduces the computational burden and noise arising from irrelevant genes.2) It simplifies gene expression tests to include only a very small number of genes rather than thousands of genes, which can bring down the cost for cancer testing significantly. 3) It calls for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment. Our simple yet very effective method involves two steps. In the first step, we choose some important genes using a 2 way Analysis of Variance (ANOVA) ranking scheme. In the second step, we test the classification capability of all simple combinations of those important genes using a good classifier such as Support Vector Machines. Our approach obtained very high accuracy with only two genes.**

Index Terms—**Gene expressions, Cancer classification, Neural networks, Support vector machines.**

I. Introduction

Compared with traditional tumor diagnostic methods based mainly on the morphological appearance of the tumor, the method using gene expression profiles is more objective, accurate, and reliable [2]. With the help of gene expression obtained from micro array technology, heterogeneous cancers can be classified into appropriate subtypes. Recently, different kinds of machine learning and statistical methods, such as artificial neural network [3], evolutionary algorithm [4], and nearest shrunken centroids [5], have been used to analyze gene expression data. Supervised machine learning can be used for cancer prediction as follows: First, a classifier is trained with a part of the samples in the cancer data set. Second, one uses the trained classifier to predict the samples in the rest of the data set to evaluate the effectiveness of the classifier. The challenge of this problem lies in the following two points:.

1) In a typical gene expression data set, there are only very few (usually from several to several tens) samples of each type of cancers. That is, the training data are scarce. .
2) A typical gene expression data set usually contains expression data of a large number of genes, say, several thousand. In other words, the data are high dimensional.

A. Bharathi, Dr.A.M.Natarajan, Bannari Amman Institute of Technology Sathyamangalam,Tamil Nadu(email:abkanika07@gmail.com, amn@bitsathy.ac.in)

In 2003, Tibshirani et al. successfully classified the lymphoma data set [6] with only 48 genes by using a statistical method called nearest shrunken centroids with an accuracy of 100 percent [7]. For the method of nearest shrunken centroids, it categorizes each sample to the class whose centriod is nearest to the sample. The difference between standard nearest centroids and nearest shrunken centroids is that the latter uses only some important genes rather than all the genes to calculate the centroids. In the same year, Lee and Lee also obtained 100 percent accuracy in this data set with an SVM classifier and the separability-based gene importance ranking [8], [9]. They used at least 20 genes to obtain this result. At the same time, they generated three principal components (PCs) from the 20 top genes. Their SVM also obtained 100 percent accuracy in the space defined by these three principal components. In fact, taking advantage of testing samples in any step of the classifier-building process,

In this paper, we propose a simple yet very effective method that leads to cancer classification using expressions of only a very few genes. Furthermore, we evaluated our methods in an honest way, which excluded the influence of the bias [11]. This paper is organized as follows: We first introduce our procedure to find the minimum gene combinations. Then, the numerical results of Lymphoma data sets demonstrate the effectiveness of our approach.

II. Method

Our proposed method is comprised of 2 steps. In step 1, we rank all genes in the training data set using a scoring scheme. Then we retain the genes with high scores. In step 2, we test the classification capability of all simple two gene combinations among the genes selected in step 2using a good classifier such as support vector machines.

2.1 Step 1: Gene Importance Ranking

In step 1, we compute the importance ranking of each gene using an Analysis of Variance (ANOVA) method. Analysis of variance (ANOVA) is a technique for analyzing experimental data in which one or more response variables are measured under various conditions identified by one or more classification variables. The combinations of levels for the classification variables form the cells of the experimental design for the data. In an analysis of variance, the variation in the response is separated into variation attributable to differences between the classification variables and variation attributable to random error. An analysis of variance constructs tests to determine the significance of the classification effects. A typical goal in an analysis of variance is to compare means of the response

variable for various combinations of the classification variables. An analysis of variance may be written as a linear model. The two-way analysis of variance is an extension to the one-way analysis of variance. There are two independent variables. Two-way ANOVA determines how a response is affected by two factors. The two independent variables in a two-way ANOVA are called factors. The idea is that there are two variables, factors, which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels. In the 2 way ANOVA interactions between row and column. These are differences between rows that are not the same at each column, equivalent to variation between columns that is not the same at each row. For each component in the 2 way ANOVA table consists of sum-of-squares, degrees of freedom, mean square, and the F ratio. Each F ratio is the ratio of the mean-square value for that source of variation to the residual mean square (with repeated-measures ANOVA, the denominator of one F ratio is the mean square for matching rather than residual mean square). [26]

2.2 Step 2: Finding the minimum gene subset

After selecting some top genes in the important ranking list, we attempt to classify the data set with one gene. We input each selected gene into our classifiers. If no good accuracy is obtained we go on classifying the data set with all possible 2 gene combinations within the selected genes. If still no good accuracy is obtained, we repeat this procedure with all of the 3-gene combinations and so on until we obtain a good accuracy. In this paper, we used the following classifier to test 2-gene combinations.

2.2.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) [21] were originally designed for binary classification. Recently, SVM [22] have become a popular tool for learning methods since they translate the input data into a larger feature space where the instances are linear separable, thus increasing efficiency. In the SVM methods a kernel which can be considered a similarity measure is used to recode the input data. The kernel is used accompanied by a map function .Even if the mathematics behind the SVM is straight forward, finding the best choices for the kernel function and parameters can be challenging, when applied to real data sets. We will use the Libsvm developed by Chang [23]. Usually, the recommended kernel function [24] for nonlinear problems is the Gaussian radial basis function, because it resembles the sigmoid kernel for certain parameters and it requires less parameters than a polynomial kernel. The kernel function parameter γ and the parameter C, which controls the complexity of the decision function versus the training error minimization, can be determined by running a 2 dimensional grid search, which means that the values for pairs of parameters (C, γ) are generated in a predefined interval with a fixed step. The performance of each combination is computed and used to determine the best pair of parameters.

The non-sparse property of the solution leads to a really slow evaluation process. Thus, for the microarray datasets a data reduction [25] can be done in terms of genes or features of the dataset considered. Redundant or highly correlated features can be replaced with a smaller uncorrelated number of features capturing the entire information. This is done by applying a method called Principal Component Analysis (PCA) before using the SVM algorithm. The method is performed by solving an eigenvector problem or by using iterative algorithms and the result is a set of orthogonal vectors called principal components. The mapping of the larger set into the new smaller set is done by projecting the initial instances on the principal components. The first principal component is defined as the direction given by a linear regression fit through the input data. This direction will hold the maximum variance in the input data. The second component is orthogonal on the first vector, uncorrelated and it is defined to maximize the remaining variance. This procedure is repeated until the last vector is obtained.

The envisioned research will follow the main steps of knowledge discovery processes:-

Gene selection - the irrelevant attributes (genes) are removed and the selected data is represented as a two-dimensional table.

Preprocessing - if the selected table contains missing values or empty cell entries, the table must be preprocessed in order to remove some of the incompleteness. Statistics should be run to obtain more information about the data.

Training and validation sample - the initial table is divided into at least two tables by using a cross validation procedure. One will be used in the training step, the other in the validation or testing step.

Interpretation and evaluation - the validation or test data set is then used to test the classificatory performance of the methods in terms of efficiency and accuracy.

2.2.2 Algorithm Description

We used five fold cross validation in the experiments because formal training and test datasets are not available for this data set. More specifically, we randomly divide data in each class into five groups. In each fold, data points in four groups are used as a training set, the data points in the remaining group is used as a test set. Hence, we have five folds of the data. The training and test sets in each fold are independent. Moreover, the experiment using data in each fold is done independently. Hence, cross validation is used here for separating the data set into several groups of training and testing sets, not for avoiding over fitting [1]. Fig.1 shows the procedure for cross validation.

III. RESULTS

In the lymphoma data set [13] there are 42 samples derived from Diffuse Large B-cell Lymphoma (DLBCL), nine samples from Follicular Lymphoma (FL), and 11 samples from Chronic Lymphocytic Leukemia (CLL). The entire data set includes the expression data of 4026 genes. In this data set, a small part of the data is missing. A k-nearest neighbor algorithm was applied to fill those missing values [10]. In the first step, we randomly divided the 62 samples into 2 parts: 31 samples for testing, 31 samples for training. We ranked the entire set of 4,026 genes according to their

ANOVA in the training set. Then we picked out the 20 genes with 2 gene combinations with 190 iterations (see table 1) and picked the highest ANOVA. (See the table 2).

TABLE1.

Gene	Correct rate	Error rate
1,4	1	0
1,8	1	0
1,9	1	0
1,14	1	0
1,15	1	0
1,16	1	0
1,18	1	0
2,4	1	0
2,8	1	0
2,9	1	0
2,11	1	0
2,14	1	0
2,15	1	0
2,16	1	0
2,18	1	0
4,7	1	0
4,12	1	0
4,17	1	0
7,8	1	0
7,9	1	0
7,18	1	0
8,17	1	0
9,12	1	0
9,17	1	0
11,17	1	0
12,14	1	0
12,18	1	0
14,17	1	0
17,18	1	0
18,20	1	0

In the complete combinations of the 190 iterations the average error rate is 7.35 percent

TABLE 2. MAXIMUM ACCURACY ACHIEVED BY THE FOLLOWING COMBINATIONS

1 ,4	1,8	1, 9	1,14	1,15	1,16	1,18	2,4
2,8	2, 9	2,11	2,14	2,15	2, 16	2,18	4 ,7
4,12	4,17	7, 8	7, 9	7,14	7,18	8,17	9,12
9,17	11,17	12,14	12,18	14,17	17,18	18,20	

We applied our SVM to classify the lymphoma micro array data set. At first, we added the selected 20 genes one by one to the network according to their ANOVA ranks. That is, we first used only a two gene that is ranked 1 as the input to the network. We trained the network with the training data set and subsequently, tested the network with the test data set.

The excellent performance of our SVM motivated us to search for the smallest gene subsets that can ensure highly accurate classification for the entire data set. We first attempted to classify the data set using two gene tested for all possible combinations within the 20 genes. Fig.1 shows the CV procedure used here. 2007, Lipo et al successfully classified the lymphoma data set [1] using T-Test method; the average accuracy was 93.85 percent. To our pleasant surprise, among all possible two gene combinations the best five fold CV accuracy for the training data reached 97.15 percent for the SVM. The corresponding testing accuracies varied from 96.77 to 100 percent. We are comparing all possible combination of tests. The results are shown in table 2. Comparing existing method, our approach obtained very good accuracy.

TABLE 2.

Knnimpute	No. fold	No.of Genes	No.of Comb.	CV Acc	Acc
(Data,3)	5	20	2	91.7	96.77
(Data,3)	5	20	3	93.97	97.6
(Data,3)	5	10	2	92.11	96.77
(Data,3)	5	10	3	93.31	100
(Data,3)	10	20	3	93.42	97.3
(Data,3)	10	20	2	91.26	96.77
(Data,3)	10	10	2	91.25	96.77
(Data,3)	10	10	3	92.47	100
(Data,5)	5	20	2	93.11	98.39
(Data,5)	5	20	3	96.4	98.4
(Data,5)	5	10	2	94.62	98.38
(Data,5)	5	10	3	97.15	100
(Data,5)	10	20	2	93.41	98.38

With the application of linear SVM raking, we have the overall mean misclassification error to be equal to 18.5%. However, the most important conclusion is drawn from the so-called confusion matrix. A confusion matrix contains information about the actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Table 3 presents such matrix corresponding to the best gene combination among 20 genes.

TABLE 3. (9, 12)

42	**0**	**0**
0	**8**	**1**
0	**0**	**11**

From the table 3. The accuracy rate is (42+8+11)61/62 = 0.9838 and the error rate is 8/62=0.129

Using Back Propagation Networks, we obtain the accuracies are shown in table 4.

TABLE 4.

Knnimpute	No.fold	No.of Genes	No.of Comb.	CV Accuracy	Accuracy
(Data,3)	5	20	2	89.85	96.77
(Data,3)	5	10	3	90.29	97.77
(Data,3)	5	10	2	89.16	96.77
(Data,5)	5	10	2	89.66	98.39
(Data,5)	5	10	3	90.08	96.77
(Data,5)	5	20	2	88.75	98.36
(Data,5)	5	20	3	90.05	96.87
(Data,3)	10	20	2	88.66	96.77

Knnimpute	No.fold	No.of Genes	No.of Comb.	CV Accuracy	Accuracy
(Data,3)	5	20	2	89.85	96.77
(Data,3)	5	10	3	90.29	97.77
(Data,3)	5	10	2	89.16	96.77
(Data,5)	5	10	2	89.66	98.39
(Data,5)	5	10	3	90.08	96.77
(Data,5)	5	20	2	88.75	98.36
(Data,5)	5	20	3	90.05	96.87
(Data,3)	10	20	2	88.66	96.77

Using T-test, we obtain the 93.85 percent accuracy Comparing all the three classifiers shown in table 5.

TABLE 5

Classifiers	Accuracy
T-test	93.85
SVM	97.91
BPN	97.43

Comparing all the three classifiers, our SVMs classifier obtained very good accuracy.

IV. Conclusion

For our purpose of finding the smallest gene subsets for accurate cancer classification, both ANOVA and CV are highly effective ranking schemes, whereas SVM is sufficiently good classifiers. As we have known from the results in the lymphoma dataset, the gene combination that gives good separation may not be unique. In the lymphoma data set, we clustered the 20 selected genes using K-means method. Mat lab 7.0 is used to implement this procedure. Finally we obtained very good accuracy compared to T-Test method.

References

[1] Lipo Wang, Feng Chu, and Wei Xie, Accurate Cancer Classification using expressions of very few genes, IEEE/ ACM transactions on computational Biology and Bioinformatics, 4, 40-52,2007

[2] Gloub et al., Molecular Classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 286,531-537

[3] Perou et al., Molecular portraits of human breast tumors. Nature, 406,747-752

[4] Cho J. H, Lee J.H, and Lee I.B, New gene selection method for classification of cancer subtypes considering within –class variation. FEBS Letters, 551, 3-7

[5] Kim H, and Park H, Multi class gene selection for classification of cancer subtypes based on generalized LDA

[6] Shipp M. A et al,.Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, Nat. Med., 8,68-74.

[7] Van't Veer L.J et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, 530-536.

[8] Vapnik V.,The nature of Statistical Learning theory, Springer-Verlag, New York.

[9] Alter O., Brown P.O. , and Botstein D., Singular value decomposition for genome-wide expression data processing and modeling, Proceedings of Natural academic Science, USA, 97(18), 10101-10106.

[10] Alter O., Brown P.O. , and Botstein D., Generalized Singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms, Proceedings of Natural academy of Science, USA, 100(6), 3351-3356.

[11] Troyanskanya I. et al., Missing value estimation methods of DNA micro array, Bioinformatics, 17(6), 520-525.

[12] Oba S. et al., A Bayesian missing value estimation method for gene exoresion profile data, Bioinformatics, 19(16), 2088-2096.

[13] Friedland S., Niknejad A., and Chihara L.,A Simultaneous reconstruction of missing data in DNA microarrays, Institute of Mathematics and its Applications preprint series, No.1948.

[14] A. A. Alizadeh et al., Distinct types of diffuse Large b-cell lymphoma identified by gene expression profiling, Nature, 403,503-511.

[15] Y. Lee and C. K. Lee, Clasication of multiple cancer types by Multicategory Support Vector Machines using gene expression data, Bioinformatics, 19, 1132-1139.

[16] M. P. Brown et al., Knowledge- based analysis of micro array gene expression data by using support vector machines, Proceedings of Natural academy of Science, USA, 97, 262-267.

[17] Roseberg S. A., Classification of lymphoid neoplasm's, Blood 84, 1359-1360.

[18] Schena M. Shalon D, Davis R. W, and Brown P.O., Quantitative monitoring of gene expression pattern with a complementary DNA micro array, Science 270, 467-470.

[19] J.M. Khan et al., Classification and diagnostic prediction of cancers using gene expression profiling and Artificial Neural Networks, Nature Medicine,7,673-679.

[20] C. Ambroise and G. J. MaKachlan, Selection Bias in Gene Extraction on the Basis of Micro array Gene-Expression data, Proceedings of National Academy of Sciences USA, 99, 6562-6566.

[21] C. Cortes and V. Vapnik, "Support-vector network," Machine Learning, vol. 20, pp. 273–297, 1995.

[22] T. Joachims, "Making large-scale SVM learning practical.", In B. Scholkopf, C. J. C. Burges and A. j.Smola, editors, Advances in Kernel Methods – Support Vector Learning, pp. 169-184, MIT Press, Cambrige, MA, 1999.

[23] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[24] N. Cristianini and J. Shawe -Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, England, 2000.

[25] Y.-J. Lee and O.L. Mangasarian, "RSVM: Reduced Support Vector Machines", Proc. of the First SIAM International Conference on Data Mining, Chicago, April 5-7, 2001.

[26] H Zhang, N Ye, J He, A Roontiva and J Aguay,"Two-way ANOVA to identify impacts of multiple interactive behavioral factors on the neuronal population dependency during the reaching motion", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008

Mrs.A.Bharathi received her Bachelor of Engineering Degree from Kongu Engineering College in 1998, Perunduai, Master of Engineering Degree from Bannari Amman Institute of Technology, Sathyamangalam, in 2007 and she is doing Doctor of Philosophy in Computer Science and Engineering from Anna University, Coimbatore. She is currently the Assistant Professor, Department of IT, Bannari Amman Institute of Technology, Sathyamangalam. Her Professional activities include…Guided Ten UG projects and guiding Seven UG and Three PG projects. Published and presented 10 papers in International and National Conferences and also published 2 international and 3 national journals.

Dr. A. M. Natarajan received his Bachelor of Engineering Degree from PSG College of Technology in 1968, Coimbatore, Master of Engineering Degree from PSG College of Technology in 1970, Coimbatore and Doctor of Philosophy in Systems Engineering from Bharathiar University, Coimbatore in 1984. He was the Principal in Kongu Engineering College at the time of relieving. He is currently the Chief Executive and Professor in Bannari Amman Institute of Technology, Sathyamangalam. His Professional activities include…Guided 15 Ph.Ds and guiding 21 Ph.Ds in the field of CSE, EEE and ECE. Published and presented more than 150 Papers in International and National Journals and also in Conferences

Fig. 1 Procedure for Cross Validation (CV)