International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 6 (July 2014) http://ijirae.com

Innovative Technique for Gene Selection in Microarray Based on Recursive Cluster Elimination and Dimension Reduction for Cancer Classification

MrunaliVaidya
Lecture in Computer Science &Engg.Dept.
Ballarpur Institute of Technology,Ballarpur
Chandrapur

Prof. P.S. Kulkarni
Professor in Information Technology Dept.
Rajiv Gandhi College of Engg., Research & Technology,

Abstract: Gene selection is usually the crucial step in microarray data analysis. A great deal of recent research has focused on the challenging task of selecting differentially expressed genes from microarray data ('gene selection'). Numerous gene selection algorithms have been proposed in the literature, but it is often unclear exactly how these algorithms respond to conditions like small sample-sizes or differing variances. Choosing an appropriate algorithm can therefore be difficult in many cases. This paper presents combination of Analysis of Variance (ANOVA), Principle Component Analysis (PCA), Recursive Cluster Elimination (RCE) a classification algorithm by employing a innovative method for gene selection. It reduces the gene expression data into minimal number of gene subset. This is a new feature selection method which uses ANOVA statistical test, principal component analysis, KNN classification &RCE (recursive cluster elimination). At each step redundant & irrelevant features are get eliminated. Classification accuracy reaches up to 99.10% and lesser time for classification when compared to other convectional techniques.

Keyword-- ANOVA, Recursive Cluster Elimination, microarray, PCA (Principle component analysis).KNN classifier.

I. INTRODUCTION

DNA microarrays offer the ability to look at the expression of thousands of genes in a single experiment one of the important applications of microarray technology is cancer classification. With microarray technology, researchers will be able to classify different diseases according to different expression levels in normal and tumor cells, to discover the relationship between genes, to identify the critical genes in the development of disease. A main task of microarray classification is to build a classifier from historical microarray gene expression data, and then it uses the classifier to classify future coming data. Due to the rapid development of DNA microarray technology, gene selection methods and classification techniques are being computed for better use of classification algorithm in microarray gene expression data .The analysis of large gene expression data sets is becoming a challenge in cancer classification. So gene selection is one of the critical aspects. Efficient gene selection can drastically ease computational burden of the subsequent classification task, and can yield a much smaller and more compact gene set without the loss of classification. In classifying microarray data, the main objective of gene selection is to search for the genes, which keep the maximum amount of information about the class and minimize the classification error. Data mining methods typically fall in to either supervised or unsupervised classes.

In unsupervised analysis, the data are organized without the benefit of external classification information. Hierarchical clustering, K-means clustering, or self-organizing maps are examples of unsupervised clustering approaches that have been widely used in microarray analysis. In Supervised analysis, the entire data set is divided into training set and a testing set and it also involves construction of classifiers, which assign predefined classes to expression profiles. Once the classifier has been trained on the training set and tested on the testing set, it can then be applied to data with unknown classification used a k-nearest neighbor strategy to classify the expression profiles of leukemia samples into two classes: acute myeloid leukemia and acute lymphocytic Classification method to classify the gene expression data into gene subset for cancer classification including ANNOVA and Principal Component Analysis (PCA) was proposed by P.Rajkumar et al.[1].This method also uses KNN for creating the gene clusters. He called the method as hybrid method. The survey of application of machine learning algorithms for classification and diagnosis of cancer is provided by Markus et al.[22], Dudoit et al.[21] compared the performance of different discrimination methods for the classification of tumors based on gene expression data.

In the present paper, we concentrate on the feature selection ANNOVA method for finding the number of gene subsets followed by PCA for further reduction of gene subsets. For the classification accuracy of gene subsets KNN classifier and Recursive cluster elimination (RCE) is used.

II. DATASETS USED IN THE EXPERIMENTATION

We have applied the algorithm on Leukemia cancer dataset by Armstrong et al.(2001).It contains DNA microarray gene expression data.Three types of leukemia samples are provided namely acute lymphoblastic leukemia(ALL), acute myeloid leukemia(AML) and mixed lineage leukemia(MLA) samples.

III. PROPOSED METHOD

Cancer classification proposed in this paper comprises of three steps. In the first step, all genes in the training dataset are tested for variations. This paper uses Analysis of Variance(ANNOVA) method for finding out variations in gene samples. In the second step Principal Component analysis (PCA) is applied. PCA is applied for further reduction in gene subsets. As samples may contain duplicate or irrelevant dimensions which can be removed after applying PCA. The last step is the combination of KNN Classification & Recursive Cluster Elimination (RCE) .It is applied for achieving the classification accuracy.

Fig1. Block Diagram of Proposed System

1. ANOVA STATISTICAL TEST

Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken. ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that cause the mean in one group to differ from the mean in another.

Most of the time ANOVA is used to compare the equality of three or more means, however when the means from two samples are compared using ANOVA it is equivalent to using a t-test to compare the means of independent samples.

ANOVA is based on comparing the variance (or variation) *between* the data samples to variation *within* each particular sample. To calculate F distribution ANOVA needs to calculate following terms

1. Total sum of squares(SST) : the total variation in the data. It is the sum of the between and within variation.

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{c} (Xij - \overline{\overline{X}})2 \quad \text{--- (1)}$$

Where r is the number of rows & c is number of columns, is the grand mean and x^{ij} is the i^{th} observation in the j^{th} column.

2. Between Sum of Squares(SSTR) (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

$$SSTR = \sum rj (\overline{Xj} - \overline{\overline{X}})2 \quad \text{--- (2)}$$

Where r^j is the number of rows in the j^{th} treatment $\overline{Xj}$

3. Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSE = \sum \sum (Xij - \overline{Xj})2 \quad \text{--- (3)}$$

Hence, you only need to compute any two of three sources of variation to conduct an ANOVA.The next step in an ANOVA is to compute the "average" sources of variation in the data usingSST, SSTR, and SSE.

4. Average total variation in the data i.e. Total Mean Square

$$MST = \frac{SST}{N-1} \quad \text{--- (4)}$$

Where N is the total number of observations.

5. Mean Square treatment (MSTR) = SSTR/c-1 is average between variation & c is the number of columns in the data.
   Mean Square Error (MSE) =SSE/(N-c) is the average within variation

6. Calculation of F Value: The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal to the ratio of MSTR and MSE. This is the ratio of the "average between variation" to the "average within variation." this ratio is known to follow an F distribution. Hence,

F =MSTR/MSE --- (5)

7. Obtain the Critical Value($\propto$)
   To find the critical value from an F distribution you must know the numerator (MSTR) and denominator (MSE) degrees of freedom, along with the significance level.
   FCV has df1 and df2 degrees of freedom, where df1 is the numerator degrees of freedom equal to c-1 and df2 is the denominator degrees of freedom equal to N-c.

In this paper the value of $\propto$ is set at 0.5, any value less than this will result in significant effects , while any value greater than this will result in non-significant effects. The probability of the F –value arising from two identical distributions gives us a measure of the significance of the between sample variation as compared to the within sample variation.

2. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is appropriate when you have obtained measures on a number of observed values and wish to develop a smaller number of artificial variables called principal components that will account for most of the variance in the observed variables.

It is a statistical pattern analysis technique for determining the key variables in a multidimensional data set that explain the differences in the observations and is very useful for analysis visualization and simplification of high dimensional data sets. Given m observations (samples or arrays) on n variables (genes) which form mxn data matrix, the goal of PCA is the reduction of data matrix dimensionality by finding r new variables, where r
is less than n. These r new variables are termed as principal components and together they account for as much of the variance in the original n variables as possible while remaining mutually uncorrelated. We start with a matrix of expression data, A, where each row corresponds to a different gene and each column corresponds to one of several different conditions to which the cells were exposed. The ait entry of the matrix contains the ith gene's relative expression ratio with respect to a control population under condition t.
To compute the principal component , the n eigenvalues and their corresponding eigenvectors are calculated. Each eigenvector defines the principal component . A component can be viewed weighted sum of the conditions. The general formula for calculating the score or weight of the first component extracted is

C1 = b 11(X1) + b12(X 2) + ... b1p(Xp)

--- (6)

Where
C1 = the subject's score on principal component 1 (the first component extracted)
b1p = the regression coefficient (or weight) for observed variable p, as used in creating principal component 1
Xp = the subject's score on observed variable p.

3. KNN CLASSIFICATION

KNN Classification is applied in combination with Recursive Cluster Elimination. The key idea behind classification is that similar observations belong to similar classes. Thus, one simply has to look for the class designators of a certain number of the nearest neighbors and weight their class numbers to assign a class number to the unknown. The weighing scheme of the class numbers is often a majority rule, but other schemes are conceivable. The number of nearest neighbors, K, should be odd to avoid ties, and it should be kept small, since a large K tends to create misclassifications unless the individual classes are well separated. One of the major drawbacks of KNN classifiers is that the classifier needs all available data. This may lead to considerable overhead, if the training dataset is large. Given an input vector, KNN extracts K closest vectors in the reference set based on similarity measures, and makes decision for the label of input vector using the labels of the K nearest neighbors. In this paper, we are combining the KNN classification with Recursive cluster Elimination to increase the accuracy.In this paper Euclidean distance, Pearson'scorrelation are used as a similarity measure. If we have input X and reference dataset D={d1,d2,…..dn}, the probability that X belongs to class cj, P(X,cj) is :

$$P(X, cj) = \sum_{di \in KNN} Sim(X, di) P(dj, cj) - bj$$

Where sim(X,di) is the similarity between X and di and bj is the bias term.

4. RECURSIVE CLUSTER ELIMINATION

The relationship between the genes of a single cluster and their functional annotation is still not clear. The clustered genes do to not have correlated functions as might have been expected. We need to remove those clusters which are contributing least to the classification.

We assume that given dataset D with S genes. The data is partitioned into two parts, one for training and other for testing. Let X denotes a two-class training dataset that consist of t samples and s genes. We define a score measurement for any list of genes as the

ability to differentiate the two classes of samples. To calculate the score we carry out the random partition the training set X of samples into f non overlapping subsets and remaining subset is used to calculate the performance. The clusters having lowest score are removed. If the number of remaining clusters is not equal to the desired number of clusters. The samples are again merged and clusters are created till we get the desired number of clusters. This procedure is repeated r times to take into account different possible partitioning. The Flowchart for RCE is as shown below.

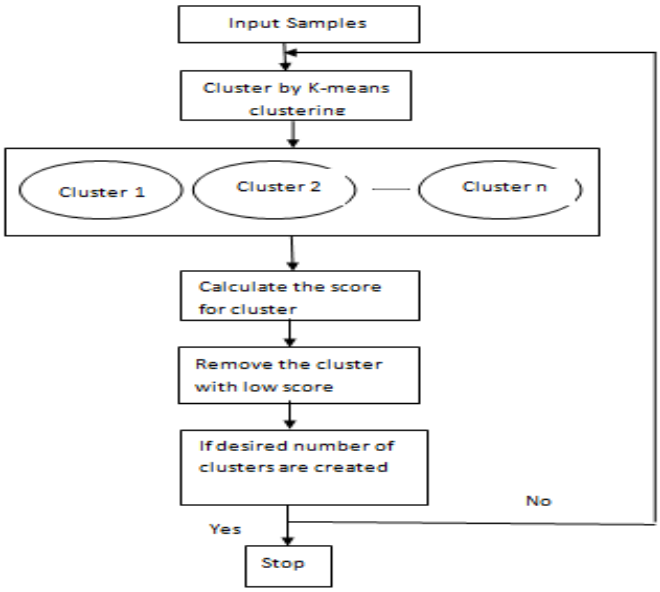

Fig.2 KNN With Recursive Cluster Elimination

5. EXPERIMENTAL RESULT AND COMPARISON WITH HYBRID (ANOVA+PCA) METHOD

In hybrid method ,on applying the ANOVA statistical test with the critical value of 0.5,to Leukemia dataset, the number of differentially expressed genes is 8144 followed by the application of PCA results into 3 clusters with 98.60% classification accuracy. When Leukemia dataset is used in our method, NOVA gives 8144 samples, after applying PCA 19 irrelevant samples were removed, followed by application of KNN with RCE results into 2 clusters.The experimental result and comparison for Leukemia dataset containing 12582 genes and 72 samples, is shown in following table.

Sr. No	Method	No. of Classes	% Accuracy
1.	Hybrid (ANOVA+PCA)	3	98.60%
2.	Our Method (ANOVA+PCA+RCE)	2	99.10%

Table 1. Classification accuracy with hybrid & ANOVA+PCA+RCE) method.

The Classification accuracy of ANNOVA Statistical Selection is 97.20 %, Hybrid method (ANOVA+PCA) is 98.60% and in our method (ANOVA+PCA+RCE) is 99.10%.

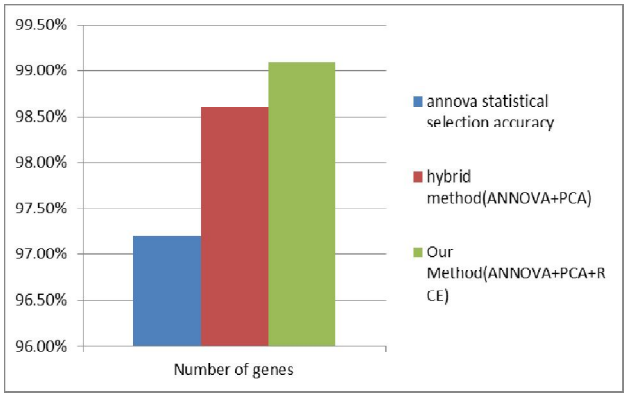

Fig. 3 Comparison of Hybrid method with new method

IV. CONCLUSION

In this paper, we have investigated the method which uses ANOVA model for analyzing the variations in the gene samples followed by PCA to reduce the number of differentially expressed genes. These steps are followed by the combination of KNN and RCE to increase the classification accuracy. As RCE is used to remove those clusters which are least contributing to the cancer classification. This method increases the accuracy up to 99.10% by reducing the number of irrelevant samples thus resulting in the less time consumption in processing the gene expression data.

V. REFERENCES

[1]P.Rajkumar,Dr.IlaVennila,K.Nirmalkumari, "A New Hybrid Method for Gene Selection in Microarray Based cancer Classification"International Journal of Engineering Science & Technology, Vol.5, No.5, November 2013.

[2]Dina A Salem,Rania Ahmed A. A. Abdul Seoud,andHeshamA.Ali "A New Gene Selection TechniqueBased on Hybrid Methods For cancer Classification Using Microarrays" International Journal of Bioscience,Biochemistry and Bioinormatics Vol.1 No.4 November 2011

[3] Sach Mukherjeeand Stephen J. Roberts, "A Theoretical Analysis of Gene Selection",*Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*

[4]Lijun Sun, Duoqian Miao &Hongyun Zhang,"Gene Selection with Rough Sets for Cancer Classification",*Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*

[5] Xian Xu and AidongZhang,Boost, "Feature Subset Selection: A New Gene SelectionAlgorithm for Microarray Dataset", *State University of New York at Buffalo, Buffalo, NY 14260, USA*

[6] Leandro N. de Castro, "Learning and Optimization Using the Clonal Selection" ,*IEEE transactions on evolutionary computation, vol. 6, no. 3, June 2002*

[7]Yuchun Tang, Yan-Qing Zhang & Zhen Huang,Granular, "SVM-RFE Gene Selection Algorithm for Reliable Prostate Cancer Classification on Microarray Expression Data",*Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*

[8] Daniele Apiletti, Elena Baralis, Giulia Bruno, Alessandro Fiori, "The Painter's Feature Selection for Gene Expression Data", *Proceedings of the 29th Annual InternationalConference of the IEEE EMBS CitéInternationale, Lyon, France August 23-26, 2007.*

[9] E. K. Tang, P. N. Suganthan and X. Yao, "Feature Selection for Microarray Data Using Least Squares SVM and Particle Swarm Optimization" , *2005 IEEE*

[10] Kai-Bo Duan, Jagath C. Rajapakse, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data", *IEEE transactions on nanobioscience, vol. 4, no. 3, september 2005*

[11] Roberto Ruiz, José C. Riquelme, "Incremental wrapper-based gene selection from microarray data for cancer classification", *Pattern Recognition Society. Published by Elsevier Ltd., 2005.*

[12] PradiptaMaji and Sankar K. Pal,"Fuzzy–Rough Sets for Information Measures and Selection of Relevant Genes From Microarray Data", *IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 40, no. 3, june 2010*

[13] Jirapech, U.T., & Aitken, S., (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6, 148.

[14] Blanco, R., Larranaga, P., Inza, I., & Sierra, B., (2004). Gene selection for cancer classification using wrapper approaches, International Journal of Pattern Recognition and Artificial Intelligence , 18(8), 1373-1390.

[15] Zhang, J.G., & Deng, H.W., (2007). Gene selection for classification of microarray data based on the Bayes error, BMC Bioinformatics, 8, 370.

[16] Wang, L., Chu, F., &Xie, W., (2007). Accurate Cancer Classification Using Expressions of Very Few Genes, IEEE/ACM Transactions on computational biology and bioinformatics, 4(1), 40-53.

[17] VenuSatuluri, V., (2007). A survey of parallel algorithms for classification.

[18] Saeys, Y., Inza, I., &Larrañaga, P., (2007). A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2507-2517.

[19] Yoonkyung Lee, Cheo Koo Lee , (2003). Classification of multiple cancer types by Multicategory support vector machines using gene

expression data, Bioinformatics, 19(9), Liu, J., &Iba, H., Selecting Informative Genes with Parallel Genetic algorithms in Tissue Classification, Genome Informatics, 12, (2001) 14-23.

[20] Keller, A. D., Schummer, M., Hood, L., &Ruzzo, W. L., (2000). Bayesian Classification of DNA Array Expression Data (Tech. Rep.No. UW-CSE-2000-08-01), Seattle: University of Washington, Department of Computer Science & Engineering.

[21] Wong, T.T., & Hsu, C.H., (2008). Two-stage classification methods for microarray data, Expert Systems with Applications, 34(1), 375-383.

[22] R. Markus, P. Carsten, (2003) Microarray-based cancer diagnosis with artificial neural networks, BioTechniques Journal, 30-35.

[23]D. Coomans; D.L. Massart (1982). "Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearestneighbour classification by using alternative voting rules". *AnalyticaChimicaActa***136**: 15–27. doi:10.1016/S0003-2670(01)95359-0.

