A Review on Gene Selection for Cancer Classification from Microarray Data

Mrunali Vaidya
Lecture in Computer Science & Engg.Dept.
Ballarpur Institute of Technology, Ballarpur

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

***Abstract*— As Cancer is the genetic disease arising from the progressive accumulation of many genetic alterations, identification of differences in the expression profile of tumor cells in comparison to their normal counterpart would provide a better platform for understanding the process of tumor formation and development. 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 reviews different technologies offered by different authors This paper aims to develop a classification algorithm by employing a hybrid method for gene selection. This is a new feature selection method which uses ANNOVA statistical test, principal component analysis, KNN classification & recursive cluster elimination. At each step redundant & irrelevant features are get eliminated.**

Keywords— *ANNOVA, 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 leukemia and some other classification algorithms .
In the literature, different schemes are reviewed to achieve more accurate & correct dataset for cancer classification.

In 2002, Pabitra Mitra , C.A. Murthy proposed a unsupervised feature selection algorithm suitable for dataset , large in both dimension & size. This method is based on measuring similarity between features whereby redundancy is removed. The algorithm has low computational complexity with respect to both number of features & number of samples of the original data. With respect to the dimension, the method has the complexity O(D) though each evaluation is time consuming.
In 2004, Sach Mukherjee *and* Stephen J. Roberts proposed a theoretical analysis of gene selection, in which the probability of successfully selecting relevant genes, using a given gene ranking function, is explicitly calculated in terms of population parameters.

In 2007, Xian Xu and Aidong Zhang *fxianxu|azhang* presents a novel general framework BFSS: Boost Feature Subset Selection to improve the performance of single-gene based discriminative scores using bootstrapping techniques. Features are selected from dynamically adjusted bootstraps of the training dataset. A nice feature of this approach is that most if not all single-gene based discriminative scores can be plugged into our system and the resulted BFSS feature selectors are expected to perform better than the original scores according

to experiments. Our approach is also independent of the classifier used.

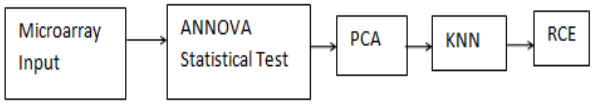
In 2005, Yuchun Tang presents the novel Granular Support Vector Machines - Recursive Feature Elimination (GSVM-RFE) algorithm for the gene selection task. GSVM-RFE can separately eliminate irrelevant, redundant or noisy genes in different granules at different stages and can select positively related genes and negatively related genes in balance. Firstly, GSVM-RFE explicitly groups genes with similar expression patterns into clusters. Therefore, the lower-ranked genes in each cluster can be safely removed as redundant genes because the more significant genes with similar functions will survive. Secondly, GSVM-RFE deals with complex correlation between genes by assigning a gene into several clusters with different membership values so that a really informative gene is more possible to survive. GSVM-RFE is more reliable to predict unseen testing samples on the prostate cancer dataset. More importantly, GSVM-RFE extracts a compact cancer related gene subset on which a SVM with 100% accuracy can be modeled.

In 2005**,** Kai-Bo Duan, Jagath C. Rajapakse proposes a new feature selection method that uses a backward elimination procedure similar to that implemented in support vector machine recursive feature elimination(SVM-RFE). Unlike the SVM-RFE method, at each step, the proposed approach computes the feature ranking score from a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data. As in SVM-RFE, we can eliminate several feature variables per step for computation efficiency at the risk of possible performance degradation. Also note that in order to use SVM-RFE and MSVM-RFE, it is important to normalize the values of each feature variable across the samples. The proposed MSVM-RFE is computationally more expensive than SVM-RFE. However, as feature selection is a pre step for building a good classifier, it is worthwhile to go through a computationally more expensive way if a better feature subset can be selected.

In 2006**,** Jiang Li proposed an efficient feature selection algorithm for the general regression problem, which utilizes a piecewise linear orthonormal least squares (OLS) procedure. The algorithm 1) determines an appropriate piecewise linear network (PLN) model for the given data set, 2) applies the OLS procedure to the PLN model, and 3) searches for useful feature subsets using a *floating search* algorithm. Initially, the feature space of the training data set is partitioned into a large number of clusters using a self-organizing-map (SOM) [48]. For each cluster, a linear regression model is then designed, and the total unexplained variance (training error) is calculated. The trained PLN model is then applied to the validation data set to get its validation error. Our goal is to find the PLN structure such that its validation error reaches the minimum. A cluster is pruned if its elimination leads to the smallest increase of the training error, and the remaining local linear models are redesigned if necessary. The pruning procedure continues till only one cluster remains. Finally, curves of the training and validation errors versus the number of clusters are produced. We find the minimum value on the validation error curve, and the number of clusters corresponding to the minimum is chosen for the PLN model. In the feature selection procedure, once the number of clusters is determined, the algorithm uses the SOM to partition the feature space and accumulates the autocorrelation and cross-correlation matrices for each cluster. These matrices remain unchanged during the whole feature search procedure. This implies that the algorithm uses the initial partition, which involves all features, for any feature subspace. The advantage of doing this is that we can significantly reduce the computational load of the algorithm. One could repartition the sub feature space and recalculate the autocorrelation and cross-correlation matrices for each feature subset, which may produce a more accurate estimate of the unexplained variance for the selected features, but this is not feasible for data with a large number of features. However, our approach may produce an optimistic estimate of the unexplained variance, for a small feature subset.

II. PROPOSED SCHEME

In this paper a new hybrid method is illustrated for selecting datasets for cancer classification. This is a new feature selection method which uses ANNOVA statistical test, recursive cluster elimination, principal component analysis, & finally KNN classifier. At each step redundant & irrelevant features are get eliminated. The microarray technology is used to carry gene expression data. As microarray contains large amount of gene expression data and few samples it is required to find out the relevant samples from it. ANNOVA statistical test is used to create clusters from microarray data. Recursive cluster elimination (RCE) is then applied to iteratively remove those clusters of genes that contribute the least to the classification performance (depending on some weight or rank).Then PCA is applied. The goal of PCA is the reduction of data matrix dimensionality by finding r new variables. Finally the KNN classifier is used for classification of genes.



A. ANNOVA 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. If the between variation is much larger than the within variation, the means of different samples will not be equal. If the between and within variations are approximately the same

size, then there will be no significant difference between sample means.
Assumptions of ANOVA:
(i) All populations involved follow a normal distribution.
(ii) All populations have the same variance (or standard deviation).
(iii) The samples are randomly selected and independent of one another.
Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.

B. Principal Component Analysis

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.

C. KNN Classification

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.

D. 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. One of the merits of the SVM-RCE is its ability to group the genes using different metrics. In this way, the outcome would be a set of significant genes that share biological networks or functions. The success of SVM-RCE suggests that estimates based on the Support Vector Machines (SVMs), a supervised machine learning classification method, to identify and score (rank) those gene clusters for accuracy of classification. K-means is used initially to group genes into clusters. After scoring by SVM the lowest scoring clusters are removed. The remaining clusters are merged, and the process is repeated.

III. CONCLUSION

Thus this algorithm will read the gene expression data from microarray extract only relevant features & classify them into datasets. Firstly the clusters of genes will be created using ANNOVA. AS these clusters may contain irrelevant & redundant features. Using recursive cluster elimination method, only those clusters having relevant features will be extracted. These clusters will be having large dimensions. Only relevant dimensions will be retained for final classification using Dimension reduction method. Thus finally only relevant & useful information carrying genes will be classified resulting this method to be more accurate than existing hybrid novel method.

REFERENCES

[1] Sach Mukherjee and Stephen J. Roberts, " A Theoretical Analysis of Gene Selection", *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*
[2] 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)*
[3] Xian Xu and Aidong Zhang,Boost, "Feature Subset Selection: A New Gene Selection Algorithm for Microarray Dataset", *State University of New York at Buffalo, Buffalo, NY 14260, USA*
[4] Leandro N. de Castro, "Learning and Optimization Using the Clonal Selection" , *IEEE transactions on evolutionary computation, vol. 6, no. 3, June 2002*
[5] 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)*
[6] 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 Cité Internationale, Lyon, France August 23-26, 2007.*
[7] E. K. Tang, P. N. Suganthan and X. Yao, "Feature Selection for Microarray Data Using Least Squares SVM and Particle Swarm Optimization" , *2005 IEEE*
[8] 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*
[9] Roberto Ruiz, José C. Riquelme, "Incremental wrapper-based gene selection from microarray data for cancer classification", *Pattern Recognition Society. Published by Elsevier Ltd., 2005*.
[10] Pradipta Maji 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*
[11] Jirapech, U.T., & Aitken, S., (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6, 148.
[12] 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.
[13] Zhang, J.G., & Deng, H.W., (2007). Gene selection for classification of microarray data based on the Bayes error, BMC Bioinformatics, 8, 370.

[14] 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.
[15] VenuSatuluri, V., (2007). A survey of parallel algorithms for classification.
[16] Saeys, Y., Inza, I., &Larrañaga, P., (2007). A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2507-2517.
[17] 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.
[18] 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.
[19] Wong, T.T., & Hsu, C.H., (2008). Two-stage classification methods for microarray data, Expert Systems with Applications, 34(1), 375-383.
[20] R. Markus, P. Carsten, (2003) Microarray-based cancer diagnosis with artificial neural networks, BioTechniques Journal, 30-35.