Classification of Tissue Samples and Gene Selection a Study

M. Sofia and Dr.D. Ramyachitra

Abstract--- *The gene selection was supported roughest to pick discriminative genes. The fundamental power of microarrays lies within the ability to conduct parallel surveys of gene expression. This paper compares many feature ranking techniques and Support Vector Machine (SVM) techniques, to use these genes to classify tissue samples of microarray information. This paper provides a comparative study of gene selection strategies for multi-class classification that can be used to reach high prediction accuracies with a tiny low number of selected genes.*

Keyword--- *Microarray, Gene selection, Classification, Rough Sets*

I. Introduction

MICROARRAY technology permits coincident activity of the expression levels of thousands of genes inside a biological tissue sample. Gene expression is to classify samples according to their gene expression profiles. Gene selection ways are classified into three types: Filter technique, Wrapper technique and embedded ways. Filter technique valuate a set of genes by viewing the intrinsic characteristics of knowledge. Wrapper technique valuate the goodness of a sequence set by the accuracy of its learning or classification. Gene choice is embedded within the construction of the classifier. Microarray expression experiments permits the recording of expression levels of thousands of sequence at the same time. These experiments primarily consist of either observing every sequence multiple times below several conditions or alternately evaluating every sequence in an exceedingly single atmosphere however in numerous genes attributable to common expression patterns. Whereas the later experiments have shown promise in classifying tissues sorts and within the identification of genes whose expression are good diagnostic indicators. Clustering analysis groups genes that have interconnected patterns. It provides gene to gene interactions and gene function. The k-nearest neighbors and genetic technique is employed for choosing a set of predictive genes from a large data. Different theoretical measures like t-test, entropy and mutual information's are wide used.

M. Sofia, Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore

Dr.D. Ramyachitra, Assistant Professor, Department of Computer Science, Bharathiar University,Coimbatore

II. Gene selection Methods

Many methods are used for gene selection and tissue sample classification using microarray.

A. K-nearest Neighbors

K- nearest neighbor is a non parametric classification method ,that predicts the sample of a test case[7].To apply K-nearest neighbor each sample was represented by a pattern of expression that consists of D genes.Each sample was then classified according to the class memberships of its k nearest neighbors, as determined by the Euclidean distance in the d-dimensional space. Dudoit S.Fridly says that the number of neighbors used is choosen by cross validation[14].By using the prediction top features are extracted and the method is used to classify unknown samples. When unclassified is accepted as a possible output, one needs to consider the various outcomes in analyzing the value of a classification[8].

B. Genetic Algorithm

A genetic algorithm (GA) is a global optimization procedure that uses the genetic evolution of biological organisms. It generates a new population from the current population using cross over and mutation methods [13]. Genetic algorithm is an intelligent technique used to find a useful subset. Since genetic algorithm has been shown to be effective in searching complex high-dimensional space.As Holland and Goldberg adapted Genetic algorithm as search tool[7].Each'chromosome' consists of d distinct genes that are initially randomly selected from all genes. A set of chromosomes is constructed to from a 'population' or a 'niche'. The genes to be selected is correspond to the features attributes.[2],[3].

C. Support Vector Machines

The ability of support vector machine is to deal with high dimensional data. The four different kernels are used for testing the genes. SVM try to find an optimal gene separating hyper plane between the classes. When the classes are linearly separable, the hyper plane is located so that it has maximal margin which should lead to better performance on data not yet seen by the SVM. When the data are not separable, there is no separating hyper plane; in this case it tries to maximize the margin but allow some classification errors subject to the constraint that the total error is less than a constant. There are several possible approaches; In this method "one against- one" approach, as implemented in "libsvgm"[12]Chan CC. 200 genes as predictors tended to perform as well as, or better than, smaller numbers. Guyon used the support vector machine as a tool for discovering informative patterns [4].

ISBN 978-93-83459-15-5 © 2014 Bonfring

III. Performance Metrics

A. Feature Ranking with Correlation Coefficients

For gene selection testing is not possible to achieve an errorless separation with a single gene. These methods include correlation methods and quantitative relation methods. Moreover, complementary genes that severally don't separate well the information are incomprehensible. The coefficient used is defined as:

$$wi=(\mu i(+)-\mu i(-))/(\sigma i(+)+\sigma i(-))(2) \qquad (1)$$

where μi and σi are the mean and standard deviation of the gene expression values of gene i for all the patients of class (+) or class (−), $i = 1, \dots n$.

$$(\mu i\ (+)-\mu i\ (-))2/(\sigma i\ (+)2 + \mu i\ (-)2) \qquad (2)$$

B. Ranking Criterion and Classification

One possible use of feature ranking is the design of a class predictor based on a pre-selected subset of features. Each feature that is correlated with the separation of interest is by itself such a class predictor, an imperfect one. This suggests a simple method of classification based on weighted voting: the features vote proportionally to their correlation coefficient, the method being used in Golub (1999). The weighted voting scheme yields a particular linear discriminate classifier:

$D(x) = w \cdot (x - \mu)$ (3) (1)
where w is defined in
$\mu = (\mu(+) + \mu(-))/2.$ (2)

It is interesting to relate this classifier to Fisher's linear discriminant. Such a classifier is

also of the form of Eq. (3), with
$\mathbf{w} = S-1(\mu(+) - \mu(-))$ (3)

And where μ is the mean vector over all training patterns. Coefficients are denoted by X(+) and X(−) the training sets of class (+) and (−). This particular form of Fisher's linear discriminant implies that S is invertible. It retains some validity if the features are uncorrelated, that is if the expected value of the product of two different feature is zero, after removing the class mean. Approximating S by its diagonal elements is one way of regularizing it.

C. Feature Ranking by Sensitivity Analysis

For classification problems, the ideal objective function is the expected value of the error. The OBD algorithm approximates DJ(i) by expanding J in Taylor series to second order. At the optimum of J , the first order term can be neglected, yielding:

$DJ(i) = (1/2)\partial 2\ J/\partial w2i(Dwi)2$ (1)

The change in weight Dwi =wi corresponds to removing feature i . The authors of the OBD algorithm advocate using DJ(i) instead of the magnitude of the weights as a weight pruning criterion. For linear discriminant functions whose cost function J is a quadratic function of wi these two criteria are equivalent. This is the case for example of the mean-squared-error

classifier (Duda, 1973) with cost function

$J =(1/2)||w||2$ (2)

D. Recursive Feature Elimination

A good feature ranking criterion is not a good feature subset ranking criterion. The criteria DJ(i) or (wi)(wi) estimate the effect of removing one feature at a time on the objective function. It will become very sub-optimal when it comes to removing several features at a time, which is necessary to obtain a small feature subset. This problem can be overcome by using the following iterative procedure that we call Recursive Feature Elimination .Optimize the weights wi with respect to J.

(DJ(i) or (wi)(wi).

This iterative procedure is an instance of backward feature elimination. In such a case, the method produces a feature subset ranking, as opposed to a feature ranking.

Feature subsets are nested $F1 \subset F2 \subset \cdots \subset F$.

E. Ranking with Correlation Coefficients

The classification of genes with the best separation between means for the two classes, was by "G-S correlation" metric are chosen:

$GS - correlation(g) = (\mu g1- \mu g2)/(\sigma g1 + \sigma g2)$, (1)

where μg1, σg1 and μg 2, σg2 are the mean and standard deviation for values of gene g among training samples of class 1 and 2, respectively. Genes with the most positive and most negative G-S correlation values are selected in parallel and grouped together in equal number in the final classifier. This method tends to not select genes for which class values have large standard deviations with respect to the training data, though some of those are most relevant and biologically informative.

IV. Conclusion

A study on the method used to perform prediction of genes such as support vector machine, k Nearest neighbor and genetic algorithms given. It is informed from the reviewthat the number of gene selection has to be reduced and classification accuracy rate has to be increased. The performance measures such as feature ranking with correlation coefficients, ranking criterion and classification, feature ranking by sensitivity analysis, recursive feature elimination and ranking with correlation coefficients are also studied. In future, the techniques given in this paper can be modified to give better performance.

References

[1] Roberto Ruiza,, Jose C. Riquelmea, Jesus S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification", Pattern Recognition 39 (2006) 2383 – 2392.

[2] JinHyukHong,SungBaeCho,"Gene boosting for cancer classification based on gene expression profiles", Pattern Recognition 42 (2009) 1761 – 1767.

[3] Goldberg.D.E,"Genetic Algorithm in search optimization and machine learning",Addison –Wesley,Reading ,MA.

[4] Guyon, "SVM Application Survey":http://www.clopinet.com/SVM.applications.html.

[5] Terrence S.furey,Nello cristianini,Nigel Duffy,"Support vector machine classification and validation of cancer tissue samples using microarray expression data", vol.16no.102000.

[6] Leping Li,Clarice R.Weinberg, Thomas A.Darden ,"Gene Slection a study of sensitivity to choice of parameters of the GA/KNN Method,vol.17no.122001.

[7] Fan Li and Yiming Yang,"Gene expression Analysis of recursive gene selection approaches from

[8] microarray data", Vol. 21 no. 19 2005.

[9] Xin Zhou and K. Z. Mao1, "Gene expression LS Bound based gene selection for DNA microarray data", Vol. 21 no. 8 2005.

[10] Christophe Ambroise and Geoffrey J. McLachlan," Selection bias in gene extraction on the basis of microarray gene-expression data".

[11] 1Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. 1999," Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotidearrays," Proc. Nat. Acad. Sci. USA 96, 6745–6750.

Cortes, C. & Vapnik, V. (1995). Support vector networks. Machine Learning, 20:3, 273–297.