Class Prediction Methods Applied to Microarray Data for Classification

Fatima .S. Shukir
The Department of Statistic, Iraqi Commission for Planning and Follow up Directorate
Computers and Informatics (ICCI), Ministry of Higher Education and Scientific Research, Baghdad, Iraq.

Abstract

The use of microarray data for the analysis of gene expression has been seen to be an important tool in biological research over the last decade. The important role of this tool is indicated by providing patients a great benefit of predicted treatment. There is an important question about a classification problem. The question is which genes play an important role in the prediction of class membership? There are many classification methods applied to microarray data to solve the classification problem. In bioinformatics, Statistical method is addressed by using microarray data. For example breast tissue samples could be classified as either cancerous or normal.
Microarray expression profiling has provided an exciting new technology to identify classifiers for selection treatments to patients. Sometime in special cases, prognostic prediction is included in class prediction. In order to predict which patient will respond to a specified treatment we can think about two classes, including responders and no responders. The objective may be to predict whether a new patient is likely to respond based on the Microarray expression profile of her or his tissue sample. That it is mean accurate prediction is of obvious value in treatment selection. To achieve the above objectives I used many methods for class prediction using gene expression profiles from microarray experiments. This research aims to explain what these methods are, how these methods are applied to the microarray dataset, analyzes the results and how feature selection is used for classification. Furthermore, comparison of these methods and cross validation will be used to evaluate the predictive accuracy.

Key words: Cancer, microarray data, classification, cross validation, predictive accuracy, gene expression profiles, feature selection.

تطبيق طرق التوقع لبيانات المايكروأري (Microarray Data) لأجل التصنيف

فاطمة صادق شكر

قسم الإحصاء / مديرية التخطيط والمتابعة، الهيئة العراقية للحاسبات والمعلوماتية، وزارة التعليم العالي والبحث العلمي، بغداد، العراق.

الخلاصة

استخدام بيانات المايكروأري (microarray data) لتحليل (gene expression) عدت كأداة هامة في البحوث البايلوجية على مدى العشر سنوات الماضية . ويدل ذلك على تطلع اهمية دور هذه الأداة (microarray) عبر توفير فائدة كبيرة للمرضى أي (توقع علاج ملائم للمرضى). هناك سؤال مهم حول مشكلة التصنيف والسؤال هو أي من الجينات تلعب دورا هاماً في توقع صنف من مجموعة من الأصناف ؟ هنالك عدد من طرق التصنيف تطبق في (microarray) لحل مشكلة التصنيف.

في الإحصاء الحيوي (Bioinformatics) هنالك طرق إحصائية تعالج باستخدام بيانات (microarray) , مثلا مرض سرطان الثدي أو القولون, إذ يمكن تصنيف عينة من هذا المرض إما تكون سرطانية او طبيعية.
إن (microarray expression profiling) قد جهز تقنية جديدة ومثيرة للتعرف على المصنفات لأجل اختيار علاج للمرضى, في بعض الأحيان وفي حالات خاصة تشخيص التوقع يتضمن من خلال توقع الصنف(class prediction). من أجل التوقع (predict) اي من المرضى سوف يستجيب لعلاج محدد نستطيع ان نفكر في صنفين مستجيبين وغير مستجيبين. الهدف هو قد يكون توقع أي من المريض الجديد محتمل للاستجابة بناءا على بيانات (microarray expression profiling) المأخوذة من عينة من الأنسجة.ذلك يعني أن دقة التوقع تكون قيمة واضحة في إختيار العلاج.
لتحقيق الأهداف السابقة استخدمت عدد من طرق توقع الصنف (class prediction) باستخدام (gene expression profiles) من خلال تجارب المايكروأري (microarray experiments). يهدف هذا البحث إلى توضيح ما هذه الطرق, وكيف يتم تطبيق هذه الطرق لمجموعة من بيانات المايكروأري (microarray data) وتحليل النتائج وكيف يتم اختيار (feature selection) المستخدم لأجل التصنيف وإضافة الى ذلك مقارنة بين هذه الطرق وسيتم استخدام مصطلح (cross validation) لتقييم دقة التوقع.
الكلمات المفاتيح : السرطان , بيانات المايكروأري , التصنيف , التحقق (cross validation) , دقة التوقع, التعبير الجيني (gene expression profiles) , اختيار خاصية (feature selection).

1. Introduction

Oncologists need improved tools for selecting treatments for individual patients. Microarray data is one of the tools which have been used in biology for the purposes of testing to develop and evaluate relevant classifiers. The important role of this tool is that it gives patients the benefit of predicted treatment. The goal for class prediction is to develop a multivariate class predictor for accurately predicting class (membership) of a new individual sample. Furthermore, it is necessary for supplemental class information to be ready for use for each individual in the data set from which the predictor will be created. For example, breast tissue samples could be classified as either cancerous or normal. The main aim of this study is to define many class prediction methods and apply all of them to microarray data. In this study I focus on feature selection because it is a very important stage in classification, particularly with microarray datasets that have thousands of feature

A human tissue has become an important part of biological research over the last several years. Therefore it used a dataset of colorectal cancer that is classified by the Dukes stages system into four stages A, B, C and D. Sub staging of the cancer is not just important for the prognosis but for the treatment, at the same time these stages tell the doctor how far the cancer could have spread. In this study it will attempted to clarify the materials and methods needed to develop predictors for classifying samples using two classes, for example, classifying the colorectal cancer patients. Moreover, compared the methods and their application and discus the extension of class prediction methods to develop a gene predictor.

2. Datasets and Feature Selection

2.1 Data Description

The data gse14333 is used for this study. It is comprised of 290 samples taken from primary colorectal cancer patients. The Affymetrix Human Genome U133 plus 2.0 Array design has been used for this experiment with 50 gene probe sets. This data included the four stages Dukes A, B, C and D. Data from the A and D samples used microarray data as a classifier, and the data B and C samples were preprocessing with reference to a training set prior to application of our prognosis classifier [1]. In this study, I was interested in the classification of sub staging that is based on DNA microarray technology. I used three stages A, B and C. Thus the data set contains 229 patients with 50 genes found by Croner et al [2]. For this study, I used a new class to work with this data, which takes the letter T if the patient is in stage C and F if the patient is in stage A or B.

2.2 Feature Selection

In general, one of the properties of microarray data is that the number of features is very large in the tens of thousands. Sometimes, these features are removed because they are likely to be unrelated for given classification purposes. In terms of feature selection, this property has two advantages: the first is that a large number of (unrelated) feature adds more effect for the inject noise during the classification task which leads to bad classifier. The second one from the explanation the feature selection is to be most related to genes in the data. For feature selection, we have general method that could be used: the method is the filter method in which features are gained individually by using statistical methods earlier to use of the classifier [3].

3. Filter Methods

The choice of filter method could depend on a prior assumption. A prior assumption is about to know the road in which the significance of individual features are ranked. Therefore, this method can be viewed as having two groupings. Firstly, those measures more affected by the consistency of the difference between classes. Secondly, those more affected by the amount of differences. At the same time, for classification algorithms there are many methods, which can be used. In this study, there are two commonly used methods. I will describe each one. The first one is the Fisher and Golub Scores. The second one is the t-test. Furthermore, an evaluation of these scores on a new dataset for cancer research enables comparison of their performance [3].

3.1 The Fisher and Golub Scores

The first kind of feature scoring is Fisher and Golub Scores. From the expression values between classes derive the means and standards deviation for the samples in each class. Therefore, the large separation between the means and small standard deviations that is lead to a good discriminating feature. The Fisher (F) and Golub (G) scores are defined as follows:

$$F = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \qquad G = \frac{|\mu_1-\mu_2|}{\sigma_1+\sigma_2}$$

For the two classes, we can define the means of t he samples as μ_1 and μ_2 also the standard deviations as σ_1 and σ_2 .

3.2 Scoring Using the t-Test

The second kind of score is the t-test, for the difference between the means of two populations. For the two classes we can define and calculate the means and variances

$$(\mu_1,\ \mu_2,\ \sigma_1^2,\sigma_2^2)$$

, and after that a weighted average of the two variances as follows:

$$s^2 = \frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}$$

and using a test statistic t

$$t = \frac{\mu_1-\mu_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

and a probability measure can be obtained from a Student t-Test distribution.

4. Cross Validation of Prediction Accuracy

At the end, the target is to construct a classifier that can accurately predict class labels for new samples. There are methods to estimate the error rate of the predictor that is "the probability of incorrectly classifying a randomly selected future case" with less bias from the reconstitution estimate. One common method is cross validation. Cross validation is a procedure which leads to greater efficiency of the data. Most of the samples are used to build the predictor, but a small number of samples are held back. Furthermore, the predictor is used to predict class membership for the held back samples.

This procedure is repeated, at each step each time leaving out a new set of samples, until all samples have been classified. For instance, each sample is left of the training set one at a time and after that classified, and the predictor will build from the all the other samples. Leave one out cross validation (LOOCV) [4] procedure gives an unbiased estimate of the true error rate of the classification.

In general, there are three parts of class prediction methods have three parts . The first part is the selection of the informative gene. The second part is computing the weights for the selected informative genes, and the last part is the creation of a prediction rule. All the three parts are very necessary for the cross validation procedure,(see Figure4.1). Feature selection is usually the most important components for developing this model. When the cross

validation is failing this means all parts of class prediction could lead to a large bias in the error rate estimated [5]. Furthermore, feature selection is an aspect of building the predictor.
Thus, when using cross validation to estimate error, feature selection should be done not on the entire training set, but separately for each cross validation sample, which is used to build the classifier. Leaving out feature selection from cross validation based on prediction methods leads to the results of error rates being overly high.

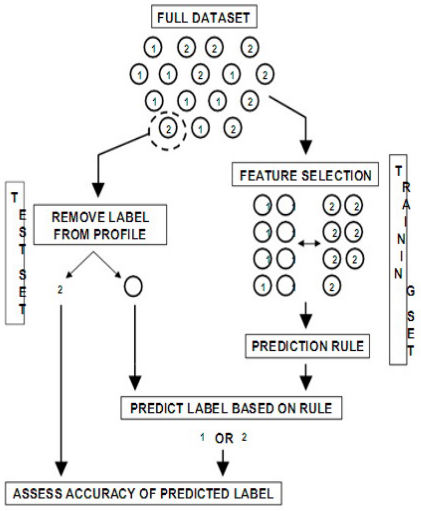

Figure 4.1- A Single Step Of The Leave-One-Out Method Of Cross-Validation [5] (Fig. 8.5).

In the depicted figure, there are two main steps for the leave-one-out method of cross validation. The first step is a process where a single sample is removed from the full dataset. The left-out sample is the test set and remaining samples comprise the training set. In other words, the comparison is done with the remaining samples in the training set. Consequently, in the training set feature selection is performed in a supervised fashion that leads to a comparison between class 1 and class 2. By the second step the prediction rule has been built from the selected features. Thus, the prediction approach (prediction rule) is applied to the gene expression profile from this process (left _out sample), which is stripped of its class label. From this step the correctness is highly observed for the prediction rule. Therefore, it is clear from this procedure why the process of leave-one-out step is performed for every sample in the full dataset [5].

4.1 Validation Dataset

There are many gene selection algorithms for building a model that can be applied inside the leave one out (LOO) training set to estimate the misclassification rate by using the bootstrap or leave-one-out cross validation (LOOCV). Before analysis of the data, the step of validation of the test set and training set is the best step for separating data that are required to estimate the error rate of the predictor. The validation set is not used until the single predictive model can be developed in the training set. Then, in the validation set the model that single predictive has been applied to during the observation is used to make a prediction. At the end, the rate of misclassification has been computed.

5. Materials and Methods

The goal of this study is to determine the accuracy level in classifying an unknown gene sample based on microarray data using prediction methods for classification of a given microarray data. It also focuses on analyzing different methods for class prediction and determining the prediction and cross validation. Therefore, there is a large amount of literature on methods for developing multivariate predictors of class membership. These methods involve linear discriminant analysis, support vector machine, and classification trees.

5.1 Methodology

There are many steps in this study. The first step is selection of the dataset. One microarray dataset, colorectal cancer (gse14333 dataset) which is used in this project. The size of the dataset is shown in Table 6.1. The second step is classification using LDA, SVM and CT. The third step is predicting class labels for testing data by using cross validation. The process of determining the validation set was achieved using an automated process in the R program and. The fourth step is obtaining and analyzing the results. At the end of these steps is a comparison of the different methods.

5.2 Dataset Selection

One microarray dataset is used in the study of colorectal cancer that is classified by the Dukes stages system into four stages A, B, C and D.
The following table provides detailed information on Gse14333 Dataset

Table 5.1- The Size Of The Gse14333 Dataset For Two Stages Of Cancer. The First Stage Has 138 Classes Of (A And B) And 91 Classes Of C. The Second Stage Has 44 Classes Of A And 91 Classes Of C, In Addition The Data Consists Of 50 Genes For Each Class.

	Gse14333 Dataset				
stages	A	B	C	total	total number of genes
(A , B) versus C	44	94	91	229	50
A versus C	44		91	135	50

5.3 Main Functions LDA

In general, there are many objectives when the data has been classified into known classes. Firstly, data are made more effective by using more information in the display. Secondly, informal assessment is used to examine the nature of difference between the classes. Thirdly, dimensionality minimization is to be achieved between the classes. Finally, from the predicted classes, the future observations are classified.

In this project, the data consist of observations that have been classified into three classes A, B and C. Data from the different classes will of course, be different in some way. We are very interested in finding the least error rate between the three stages of cancer, particularly A and B versus C stage, because stage C describes invasive cancers that have spread outside the colorectal to other parts of the body. For that reason, I used a data set containing 229 observations (patients) with 50 expression genes utilized by Croner et all [2], also the second stage contains 135 patients with 50 genes. Furthermore, the stage of the patient is classified by the Dukes system.

Hence, we presented the discriminant analysis for the (gse14333) dataset by the trained linear discriminant analysis (LDA) method, as can be seen from the Figures 5.1 and5.2 which depict the first and second stage respectively. The plot of the linear classification method is used for the expression levels of individual genes. LD1 and LD2 represent the linear discriminate analysis. The panel depicts the linear discriminate analysis, and every point on the plot denotes a gene used for classification by the LDA method.

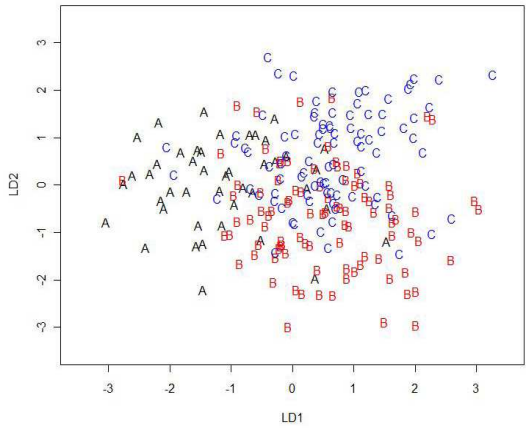

Figure 5.1- This Plot Shows The Linear Discriminant Analysis Of The Gse14333 Dataset.

LDA is used to find a linear combination of variables for separating the classes. There are three different classes A, B and C. In addition, to find the linear discriminant function the class variance is partitioned in decreasing order. Furthermore, the vertical line displays the LD1 and the horizontal line displays the LD2. We can see from the data, LDA is discriminant, but the data are very close to each other in this plot. Linear discriminant analysis seeks to maximize the separation among the different classes. This means that we should have some measure of the separation between the classes. The dataset consists of 44 A classes, 94 B classes and 91 C classes.

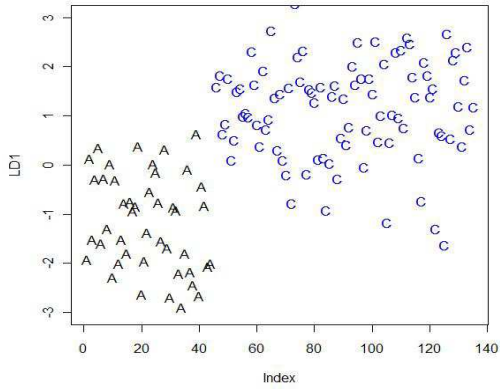

Figure 5.2-The Plot Displays The Linear Discriminant Analysis (LDA) Of The Gse14333 Dataset That Are Similar To The First Stage, But There Are Two Classes A And C For The Discriminant.

In this panel, we can see the classes are much closer to each other but, they are made discriminant and the dataset consists of 44 A classes and 91 C classes. Linear discriminant analysis (LDA) seeks to find the maximum separation between the classes, which require minimizing the variation among sample points within the same class.

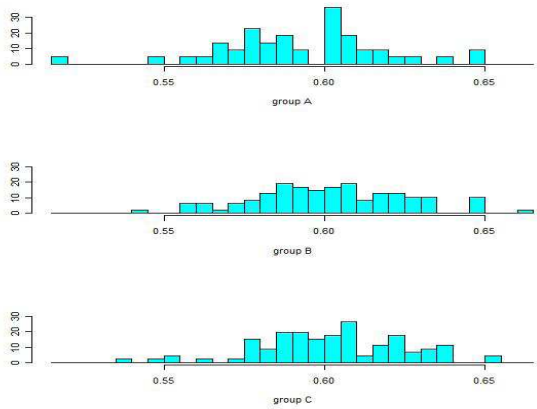

Figure 5.3- Histogram Of The Genes Represented By One Expression Gene (X336_At) Which Used The Three Groups A, B And C.

There is information about how the data should look. It is necessary to identify each gene with the three groups. For this reason I used the data to produce graphs of all the genes. As mentioned previously, every point on the plot denotes a gene used for classification by the LDA method. In addition, each gene involved the three groups A, B and C which are used for discriminant analysis. Therefore, Figure 5.3 is presented by grouping variables.

Consequently, predictions of this model from the three groups of classes is described as follows:

Class by using cross validation LDA is often called **leave-one-out testing.**

```
 [1] B B C B B C C A C A C A A A B C A C B B A A C C C C C C C B A B A B A A C A
[38] B C A B B B B B B C B C B B B A B C C B B C C B B A A B B C C C C B B B B B C C
[75] B C C A C A C A B C B B B C A A B A B C C A B C C C B A B A C B B B B B A B C B
[112] A B C A B B C B C B B B B B C A B B C C B C B B B B C B C B B B C C B B C C B
[149] C C B C C C B B B B C C B B B B C B C B C C A B C B A C B A A B C B A C B
[186] C C C C C C A C C C C B B B C A B A C B B A C C C C C C A C C B C A B C A C B B
[223] C C B C B B B
Levels: A B C

 [1] A C A C C A C A C A C A A C A C A A C A A A C C C C A C A A C C A A A C A
[38] A C A C C A A C C C C C C C A A C C C C C C C C C A C C C C C C A C C A C A C C
[75] C C A C C C A C A A A C C A C C C C C C C C C A C C C C C C C A A A C C C C C
[112] C C C C C A A C C C C C A C A A C A C C C C A C C A C
Levels: A C
```

stage 1 (A, B) versus C	right prediction	wrong prediction
A	14	30
B	48	46
C	43	48
stage 2 A versus B		
A	24	20
C	67	24

Table 5.2- Results Of Right And Wrong Prediction Using Predicted Class For Classification.

As can be seen, the table shows 30, 46 and 48 wrong predictions for class A, B and C respectively, this is used for stage 1. Similarity, for stage two, there are 20 and 24 wrong predictions for class A and C respectively. This means that the bold A represents a right prediction in class A that is used for stage 1 while B and C represent a wrong prediction. Therefore, many members of class A are misallocated to class B and C in the first stage as follows: (1, 2, 3, 4, 5, 6, 7, 9, . . .). As a result, it is clearly that finding the discriminant is not perfect. In the second stage the members of class A that are misallocated to class C is as follows: (2, 4, 5, 7,9,11 ...). In comparison with the first case, the discriminant is almost good but not perfect.

As a result, after the model of the single predictive gene is developed in the training set, the confusion matrix in these models helps the analyst to see where misclassifications are actually happening. The numbers along the diagonals of the matrix express correct classification, while off diagonal numbers express misclassification. The confusion matrices for these models are as follows:

Predicted Class

Actual class

	A	B	C
A	14	15	15
B	15	48	31
C	12	36	43

Therefore, it is necessary to accurately predict class labels for new samples. Cross validation is a very important method to make data more efficient.

The following two tables provide detailed information on Misclassification:

Table 5.3- Misclassification For Cancer Stage 1.

classes	Prior probabilities	N	Misclassification
A	0.1921397	44	30
B	0.4104803	94	46
C	0.3973799	91	48
	0.999999	229	124

Table5.4- Misclassification For Cancer Stage 2.

classes	Prior probabilities	N	Misclassification
A	0.3259259	44	20
C	0.6740741	91	24
	1	135	44

The two tables show the misclassification by class and summarize the number of stages in each class in the data like the prior probability of classes, N (number of class), for the first stage the total misclassification rate is 124/229=0.541 and for the second stage it is 0.325. Hence, at the end of this step is cross validation is achieved. The accuracy of these models is shown in Table 5.5.

Table5.5 - Cross Validation Classification Table Displaying The Total Cases Correctly Classified For The First Stage Is 45% And 67% For The Second Stage.

stages	Correctly Classified	Incorrectly Classified
(A , B) versus C	45%	55%
A versus C	67%	33%

The performance of the linear discriminant analysis was tested using cross validation. The number of cases misclassified is an estimate or 'best' prediction of cases if the LDA is applied to a new datasets.

Predicted Class

Actual class

	A	C
A	24	20
C	24	67

5.4 Support Vector Machines

A support vector machine is a classification algorithm. The idea of this method is to find a linear combination which gives the best separation of the samples in the two groups from the class labels. When finding the perfect separation is difficult, we can obtain the best linear combination by minimizing the number of misclassifications. This method is applied to microarray data by using log expression value.

The simple way to build a binary classifier is by using a hyperplane separating the class members with a positive instance from the nonmembers with a negative instance in the space. Each vector $\vec{x}$ in the gene expression matrix could be represented as a point in an m-dimensional expression space. Sometimes, the data may include nonseparable members. If this is the case, no hyperplane exists which can successfully separate the positive from the negative instances. This is a problem. One solution is to define a separating hyperplane and to map the data onto a higher dimensional space. The higher dimensional space is known as the feature space. Thus, with an appropriate feature space of a sufficient dimension, the training set has been made separable.

On the other hand, separating the data in this way in the learning system runs the risk of discovering a solution that could be the result of overfitting the data. It is necessary to avoid this phenomenon by choosing the hyperplane which has the maximum separation margin which means it is possible to separate the positive from the negative instances in the higher dimensional (feature space).

It is possible to apply the data in R. For example, the real dataset of the colorectal cancer observations is used in the R language, which is based on the S3 class mechanisms. In general, the R language is provides a training function, prediction method, support vectors and plot method, and moreover, decision boundaries. For the classification task, the radial basis function kernel with a fixed hyper parameter is

applied to the data of colorectal cancer, see Figure 5.4 stage 1 of colorectal cancer and Figure 5.5 stage 2 of colorectal cancer. Thus, after the dataset is applied using the SVM classifier, we can visualize a 2-dimensional projection of the (gse14333) with highlighted classes and support vectors. Of course, the highlighted class will be different depending on which stage of cancer is used.

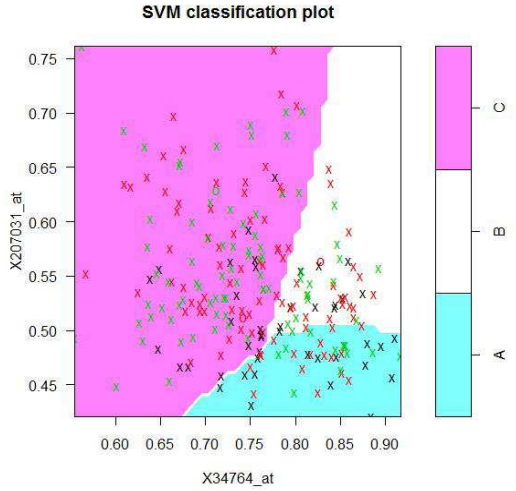

Figure 5.4- The Plot Shows SVM Visualizing The Gse14333 Dataset For Three Classes(A,B)Versus C.

The parameter for the support vector machine is determined by the function svm from the e1071 package. Each 'X' is represents a support vector, and the predicted class regions can be determined using the coloured backgrounds. In addition, the true classes are highlighted by symbol colour. For hard margin SVM, support vectors are 'X' which are "on the margin." In the figure above, SVM is closer to a hard margin, and you can see the 'X'es that are touching the margin (the margin is about 0 in that figure, so it is necessary to use hyperplane for separating the data).

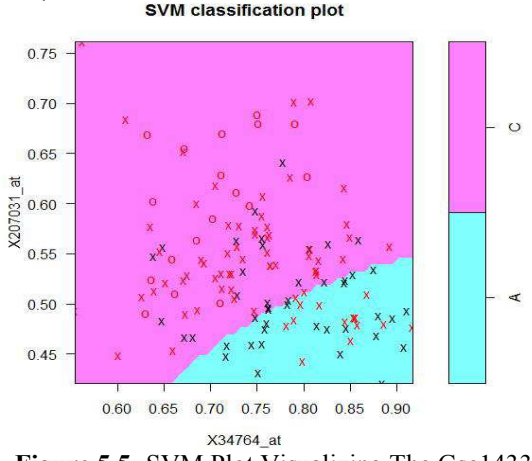

Figure 5.5- SVM Plot Visualizing The Gse14333 Dataset For Two Classes A Versus C.

As can be seen 'X'es represent the support vectors and predicted class regions are visualized using coloured backgrounds. In addition, the true classes are highlighted by the symbol colour. We can see 'X'es that are touching the margin (the margin is about 0 in this figure) . It is necessary to find an SVM that does not automatically select genes, and which is designed for continuous gene prediction. [6][7].

5.4.1 Estimate Accuracy Using 10-Fold Cross Validation

Table 5.6-: Table Shows The Estimated Accuracy Using 10-Fold Cross Validation On Training Data. The Result Of The Total Accuracy For Stage 1 Is 42.35808, And For Stage 2 It Is 71.85185. The Following Two Tables Provide Detailed Performance Results By Using 10-Fold Cross Validationon The Training Data For Two-Stage Cancer.

10-fold cross-validation on training data (A, B)	
(A, B) versus C (first stage)	A versus C (second stage)
Single Accuracies	Single Accuracies
27.27273	69.23077
39.13043	64.28571
56.52174	76.92308
47.82609	71.42857
52.17391	69.23077
30.43478	57.14286
43.47826	69.23077
47.82609	85.71429
47.82609	76.92308
30.43478	78.57143

5.4.2 Computing The Predictive Accuracy On The Test Set

Consequently, we can train our final model. This involves computing the predictive accuracy on the test set. Thus, the confusion matrices for these models are as follows:

Predicted Class

Actual class	A	B	C
A	14	15	15
B	15	48	31
C	12	36	43

Predicted Class

Actual class	A	C
A	24	20
C	24	67

The following table shows misclassification by stage and summarizes the number of classes in each stage in the data like the number of classes, Misclassification (M), the number of support vectors, correctly classified and incorrectly classified. For the first stage the total misclassification rate is 42/229=0.18334 and for the second stage it is 0.1703.

Table5.7- Cross Validation Classification Table Showing That The Total Number Of Support Vectors Correctly Classified In The First Stage Is 82% And For The Second Stage It Is 83%.

stage	N of classes	N of SV	M	Correctly Classified	Incorrectly Classified
(A , B) versus C	3	196	42	82%	18%
A versus C	2	95	23	83%	17%

5.5 Classification Trees

Another method for classification using microarray gene expression profiles are classification trees that use classification for prediction. The construction of a classification tree is started by splitting the gene from the two nodes or by building a subset based on the expression level of one of the genes. Thus, one node includes the remaining samples, and the other node includes samples with an expression level that is selected for the gene from the selected threshold value. Through this process, the selected split produces two nodes. The first node will consist of specimens from class 1 and the second one of specimens from class 2. Furthermore, the important step is the optimizing function because the split depends on it and is to be selected based on this function. Usually, for an ideal split to be obtained from this process, there is no threshold value or gene to produce in this split.

In addition, after finishing, the process of optimal split that is uses the gene and threshold value in the training set for the two nodes, the process then is repeated for each of the two nodes. Finally, the process is used to determine the best split of the samples. At the end of this procedure, it is clear that every node is split based on the threshold expression level and the gene and also every node is expressed as a set of samples.

In general, at the top of the tree, the root node consists of all the samples. Consequently, in the tree, the terminal nodes have been assigned to a class. Then the class which we assigned in the terminal node could be simply the class with the most widespread samples relating to that node. Furthermore, there are many halting procedures that could be used in the tree after it has grown by using the hierarchical method. For instance, the splitting of a node may stop if there is less than a specified number of samples contained in the node, or if the samples that appear in the node are sufficiently homogeneous with regard to class labels.

Consequently, this rule could be modified for class members to take the prior of the probability if it is different or to take into account the different costs of misclassification. Usually the classification stage of a new sample is useful because it can be used to determine in which terminal node the new sample would reside. Then, from this stage, the class of the new sample that is unknown has been predicted [5].

5.5.1 Simple Example of a Classification Tree

For our example, we have a dataset of the colorectal cancer patients with 50 genes, the number of the patients with the first stage of cancer (A, B) versus C is 229, and the number of terminal nodes (NOTN) is 29. In addition, the number of genes actually used in tree construction is 22 genes. The misclassification error rate (MER) is 0.1441 and the Residual Mean Deviance (RMD) is 0.6443. The first step of the structure tree is Node 1, and the second step is to split gene (X207031) into two profiles. The first one is at Node 2 with log ratio less than

(0.565616) and the other profile is at Node 3. Consequently, a third step for Node 2 and Node 3 is needed to further divide the heterogeneous samples in this subset. From gene (X209353_ s_ at), we can see two profiles. The first one is at Node 4 with log ratio less than 0.507443 and the other one profile is at Node 5. Furthermore, from the gene in Node 4 we can see the prediction of class B and a second one for the profile at Node 5. As a result, the fourth step is where the gene (X34764_s_at) is split into two Nodes. The first one is Node 6 needs another split for prediction and Node 7 also needs another split for prediction see Figure 5.6.

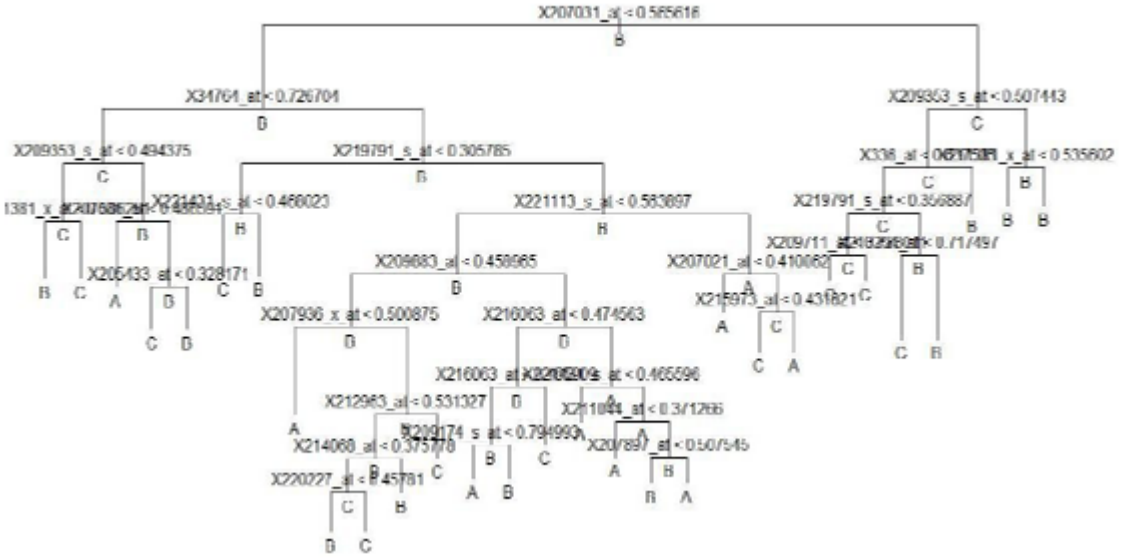

Figure5.6-Maximum Classification Tree For Prediction Including 2 Classes (A, B) Versus C.

On the other hand, the process of creating the tree at the second stage is similar, but the number of terminal nodes and genes actually used in tree construction of course is different (see Figure5.7). Thus, the following table provides details of the information with two stages used for constructing the classification tree.

Table 5. 8- Results For Construction Tree By Using Classification Tree.

stages	NOTN	Genes actually used in tree construction	RMD	MER
(A,B) versus C	29	22	0.6443	0.1441
A versus C	14	7	0.2815	0.05185

The table shows for each stage of cancer the number of terminal nodes(NOTN), the number of expression genes actually used in tree construction, the Misclassification Error Rate(MER) and the Residual Mean Deviance(RMD). In the first stage, as can be observed RMD = 128.9 / 200 = 0.6443 and MER = 33 / 229 = 0.1441. Furthermore, for the second-stage RMD = 0.2815= 34.06 / 121 and MER= 0.05185 = 7 / 135. The process of creating the tree from the dataset was achieved using an automated process in the R program.

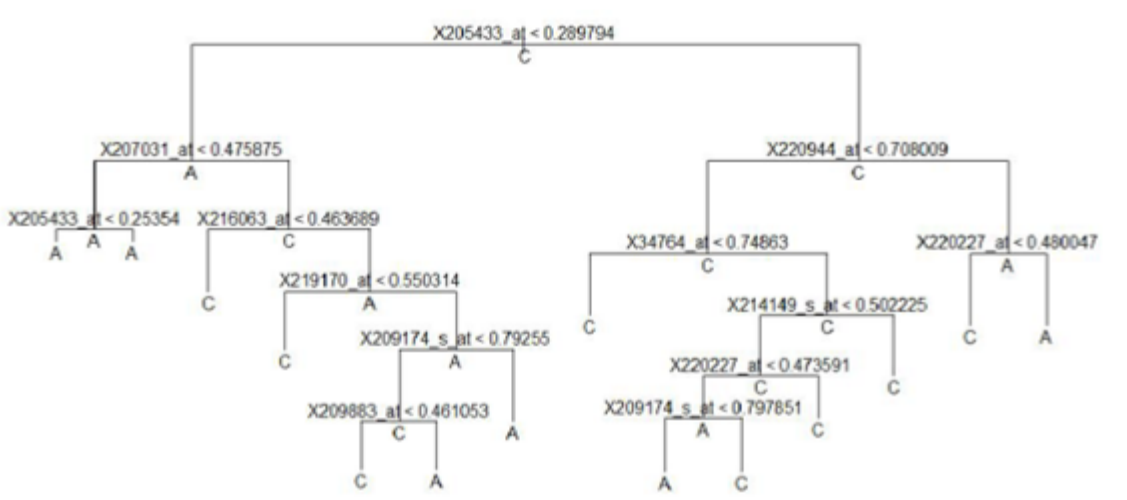

Figure 5.7- Maximum Classification Tree For Prediction Including Two Classes A Versus C.

In the two figures depicted above, we can see the name of the gene for splitting as well as the numbers beside the arrows denoting the threshold value that is used for splitting. It is clear that the best step in the structure is the splitting point of the nodes. It is necessary to split all the terminal nodes that consist of just specimens from one class. Nevertheless, that is more likely to lead to overfitting of the data in the classification tree. Therefore, overfitting is more likely to appear in the model, that is the model applied to the data does not generate new data well (leading to wrong prediction). The reason for this is related to variation and random noise, so one of the major problems is high variance in the classification tree (CT). This problem has been addressed by a method of pruning which stops the generation of new split nodes [8]. Hence, the best tree is selected by using the cross validation method. Thus, the pruning of the tree is determined by this method (see Figure 5.8 for the first stage and Figure 5.9 for the second stage).

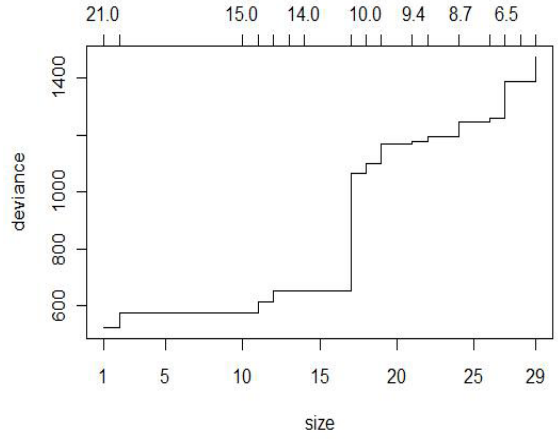

Figure 5.8- Cross Validation For Prediction Accuracy Including Two Classes (A, B) Versus C

To find the optimal tree size the only way is to use the cross validation procedure. Hence, the maximum size of a tree could have very high complexity as well as a large number of levels. In this case, if we had to perform classification of new data, it would have to be optimized beforehand to use classification. Therefore, tree optimization is choosing the right size of the tree by use cut off significant subtrees, though in this case, we can use pruning algorithms (see figure 5.10). Hence, in this figure we can see the best size of the tree is above 2.

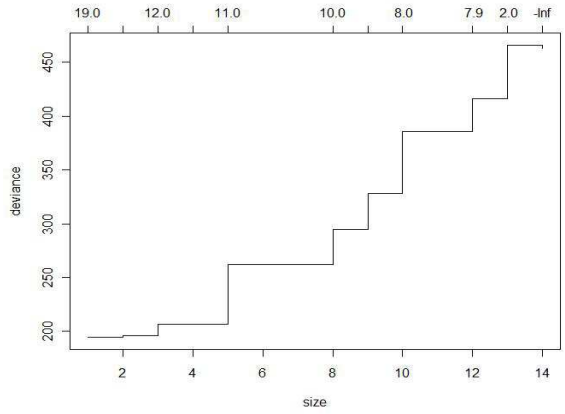

Figure 5.9- Cross Validation For Prediction Accuracy Including Two Classes A Versus C

The figure displays the size of the tree by using the cross validation procedure, and the best size of the tree is above 3. It is clear that the maximum size of a tree would be very complex. Moreover, it would contain a large number of levels. As mentioned previously, for tree optimization it is necessary to choose the right size of the tree by use cut off significant subtrees and we can also use pruning algorithms (see figure 5.11).

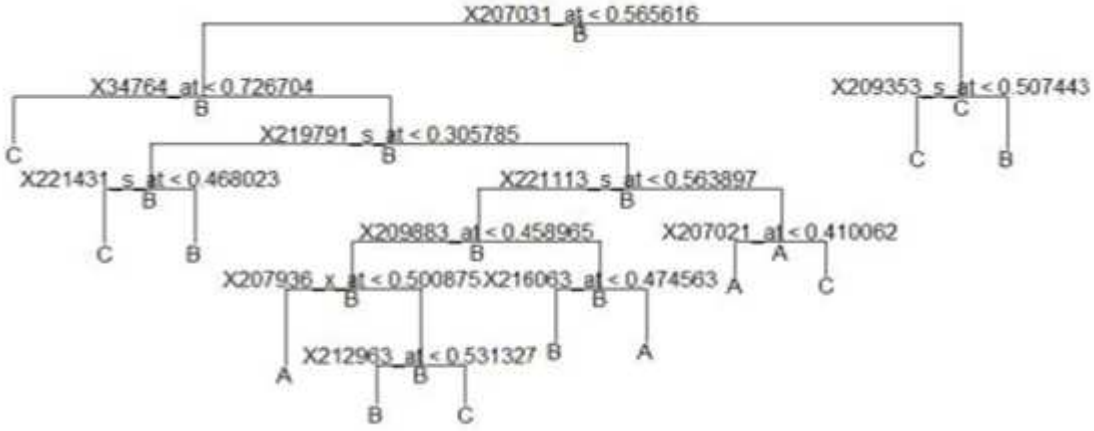

Figure5.10- Prune Misclassifications After Cross Validation Selection For The First Stage.

It can be seen that the tree prune misclass ideally from the simple tree rather than the first one obtained. After the algorithm of the tree building has stopped. Therefore, the prune misclass has constructed with 11 actual genes, and the number of terminal nodes is 12. Furthermore, the Misclassification Error Rate (MER) is about 0.3144.

Figure 5.11- Prune Misclassifications After Cross Validation Selection For The Second Stage.

The figure displays the tree prune misclass ideally from a simple tree. In comparison with the first one that is obtained the pruned misclassification have just 10 actual genes and the number of terminal nodes is 13. Moreover, the Misclassification Error Rate (MER) is about 0.05185.

6. Comparison of Methods

In this study, the class prediction methods applied to microarray data by using easy classification problems such as, between two classes A and B versus C (e.g., non invasive versus invasive) are not made in order to compare prediction methods. Furthermore, the real problem is not finding a standard prediction method for comparison purposes as probably there is method that would be ideal in all cases. "The relative performance of methods is likely to depend on the biological classification under investigation, the genetic disparity among classes, within-class heterogeneity, and size of the training set " Simon. et al [5].

This study suggests neglecting gene correlation and interaction leads to the best performing method. For the two stages, the linear discriminant analysis (LDA) had the best overall performance, with few missing data. Another method for class prediction is Classification Tree (CT), the method of (CT) gave the worst performance with a high number of missing data. This means that the (CT) model that it is used for class prediction could not be fitted. The last method for classification the Support Vector Machine (SVM), gave an intermediate performance. The following table provides the Misclassification Error Rate (MER) for tested classifiers.

Table6.1- Results Of Misclassification Error Rate (MER) For Classification Methods LDA, SVM And CT.

	LDA	SVM	CT
stages	MER	MER	MER
stage 1 (A, B) versus C	0.3955	0.3974	0.4162
stage 2 A versus C	0.2970	0.3324	0.4081

For microarray data, if I compare the LDA and linear SVM classification methods, the only difference is in the approach of determining the weights of the linear combination. Furthermore, for the nonlinear support vector machine, we can use many kinds of approach [9]. However, in comparison with linear versions of gene expression data it is clearly more effective. Sometimes, it is clear that for a comparison of methods, the interaction and correlation between genes in the system of the biology is not important.

In general, it is necessary to find from this study that the class prediction methods used that include gene interaction and correlation probably have more of an advantage. The reason for this is that datasets of gene expression are very large and available.

7. Conclusions

This research dealt with class prediction methods for classification. The results of this study exhibit interesting directions for further research. The class prediction methods are used in microarray datasets such as LDA, SVM and Classification and Regression Trees (CART).

From my analysis of the data it is clear that feature selection is important for classification. For prediction methods, it is necessary to use a classification tree to perform automatic feature selection. This means that features are selected at each step based on the number of features. Consequently, after using cross validation to prune the tree, the number of features is already determined. Therefore, feature selection in a classification tree is a necessary part to build the tree. Also pruning deals with overfitting. On the other hand, linear discriminant analysis does not perform feature selection because all gene expression values are used in building the classifier. Furthermore, the support vector machine does perform feature selection. As it was mentioned previously, the second method of feature selection is Recursive Elimination of Features. Thus, feature selection is implemented directly within the algorithm.

According to the results obtained, the class prediction methods proved to have an effective

level of accuracy in classifying an unknown gene sample based on microarray data. It is necessary to find the lowest error rate that distinguish for the easy classification problem. Therefore, in this work, the result of LDA for two stages had the best performance with fewer missing data. Even in the best performance prediction method observed there is still dependence on the interaction and correlation genes. But, the classification tree (CT) had the worst performance. Thus, the results depend on the data of the clinical study and the given sample size. This means there is no optimal method in all situations.

References:

[1] NCBI, **2011,** 'GEO' Series gse14333 .http:www.ncbi.nlm.nih.gov (26-06-2011).

[2] Croner, R., Förtsch T., Brückl, W., Rödel, F., Rödel, C., Papadopoulos, T., Brabletz, T., Kirchner, T., Sachs, M., Behrens, J., Klein-Hitpass, L., Stürzl, M., Hohenberger, W., & Lausen, B, **2008,** ' Molecular Signature for Lymphatic Metastasis in Colorectal Carcinomas'.Annals of Surgery , 247 (5) , pp. 803-810.

[3] Simon, R., Richard, D., Williams,&Colin, C,.**2005** "Class Predication with Microarray Datasets" In: Udo, S., Lakhmi, C., & Patric, S. Bioinformatic Using Computational Intellegience Paradigms. Gemany: Springer-Verlag Berlin Heidelberg. Pp.119-141.

[4] Lee, J., Lee, J., Park, M., & Song, S. **2005**'An extensive comparison of recent classification tools applied to microarray data'Computational Statistics & Data Analysis . 48 (4) pp.869-885.

[5] Simon, R., korn, E., Mcshane, L., Radmacher, M., Wright, G.,& Zhao, Y.**2003** Design and Analysis of DNA Microarray Investigations. 2th ed. USA, Springer-Verlag New York Berlin Heidelberg.

[6] Krijnen, W. **2009** Applied Statistics for Bioinformatics using R. The Netherlands.

[7] Karatzoglou, A., Meyer, D. & Hornik, K,.**2006**'Support Vector Machines in R' Journal of Statistical Software. 15(9) pp. 1-28.

[8] Breiman, L., Friedman, J., Stone, C., & Olshen, R., **1984** Classi_cation and Regression Trees. Belmont, CA: Wadsworth.

[9] Vapnik V., **1998** Statistical Learning Theory . New York: Wiley.