A new Classification and prediction model with Two-stage Gene selection method using minimal subsets of Gene Expression data

Mallika R.* and Saravanan V.

*Department of Computer Science, Sri Ramakrishna College of Arts & Science for Women, Coimbatore, India.

Department of Computer Applications, Karunya University, Coimbatore, India

Abstract -Data mining models are extensively used in the field of disease diagnosis. Gene expression data are a main factor for the success of disease diagnosis. With thousands of gene expression data, gene selection is being a big challenge prior to classification. The proposed method incorporates two stages in gene selection. In the first stage pair wise gene selection was performed using a popular statistical technique. In the second stage the gene pairs that achieved 100% Cross Validation (CV) accuracy of those genes selected in first stage were used for classification. The testing results were compared with the single stage method and improvement on the computational burden was also proven to be the best in the proposed two-stage method. The paper also compares the performances of the three different classifiers Support Vector Machines (SVM), K Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA) and promising results have been achieved.

Keywords: Support Vector Machines, Linear Discriminant Analysis, K Nearest neighbour, Microarray data, Gene selection, Classification

INTRODUCTION

One of the principal features of microarray is the volume of quantitative data that they generate and the challenge is to interpret and use the data. Applying data mining algorithms and statistical techniques can ensure to the challenges [4]. Many classification methods and gene selection techniques are been computed for better use of classification algorithm in microarray gene expression data [7, 8]. Generally classification of microarray data in machine learning is to train the classifier to accurately recognise the genes from training sample and to classify the test samples with trained classifier [10]. Once such predictive model is built, it can be used to predict the class of objects. The issue of gene selection has become a central challenge in the field of microarray data analysis. With thousands of gene expression data many genes contain irrelevant information out of which only a small number of genes may be important. Therefore, there should be techniques capable of selecting the appropriate subset of genes from the entire set of microarray data [1]. Selection of important genes using statistical technique was carried out in various papers such as Fisher Criterion, Signal-to-Noise, traditional t-test, and Mann-Whitney rank sum statistic [17], chi-squared test, Euclidean distance [20] and the some of the classification algorithms used were SVMs, k-nn [18], Genetic algorithms (GA) [9] Naive bayes (NB)[2].In 2003, Tibshirani [16]

successfully classified the lymphoma data set with only 48 genes by using a statistical method called nearest shrunken centroids and used 43 genes for SRBCT data. Lipo Wang [11] in 2007 proposed an algorithm in finding out minimum number of gene up to 3 genes with best classification accuracy using C-SVM and Fuzzy Neural Networks (FNN). Tzu-Tsung [16] in 2008 proposed a classification method to classify the causality of a disease is of two stages. First stage the Gene Selection mechanism with individual or subset gene ranking and the second stage applying a classification tool with or without dimensionality reduction. Li-Yeh Chuang [12] in 2009 proposed a two-stage feature selection method using various cancer datasets. In the first stage all genes were ranked and in the second stage fixed number of gene subsets were ranked using particle swam optimisation and those gene subsets were classified. This paper proposes an efficient classification model using statistical model for individual gene ranking and data mining models for finding minimum number of gene rather than thousands of genes in two stages, which can be used to give good classification accuracy. In preprocessing stage, all genes are ranked and in the first stage of gene selection all possible top ranked gene pairs were ranked .In the second stage top ranked gene pairs were used to train the classifier. The gene pairs that achieved 100% accuracy were used to retrain the classifier and testing was performed. The results were well compared with the previous results and proved for best accuracies. Furthermore, the paper compares the SVM, KNN and LDA classifiers using three publicly available databases Lymphoma, Liver and Leukemia. Their performances are compared with their respective accuracies.

2. METHODOLOGY & DATASET USED

A. Gene Ranking -ANOVA p-values

Analysis of Variance (ANOVA) is a technique, which is frequently used in the analysis of microarray data, e.g. to assess the significance of treatment effects, and to select interesting genes based on P-values. [9]. The ANOVA test is known to be robust and assumes that all sample populations are normally distributed with equal variance and all observations are mutually independent.

The approach chosen in this paper is the one-way ANOVA that performs an analysis on comparing two or more groups (classes) for each gene and returns a single p-value that is significant if one or more groups are different from others. The most significantly varying genes have the smallest p-values. Of all the information presented in the ANOVA table, if the p value for the F- ratio is less than the critical value (α), then the effect is said to be significant. In this paper the α value is set at .05, any value less than this will result in significant effects, while any value greater than this value will result in non-significant effects. The very small p-value indicates that differences between the column means are highly significant. 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. Small p-values indicate a low probability of the between-sample variation being due to sampling of the within-sample distribution, small p-values indicate

interesting genes. The paper uses the p-values for individual gene ranking and pair wise gene ranking.

B. Support Vector Machines (SVM)

SVMs are the most modern method applied to classify gene expression data, which works by separating space into two regions by a straight line or hyper plane in higher dimensions. SVMs were formulated for binary classification (2 classes) but cannot naturally extend to more than two classes. SVMs are able to find the optimal hyper plane that minimizes the boundaries between patterns [13]. SVMs are power tools used widely to classify gene expression data [6][19]. How to effectively extend SVM for a multi-class classification is still an ongoing research issue [3]. This paper gives effective methodology to classify a multi-class problem .To extend SVM for multi-class classification, SVMs were designed with SVM *one-against-one*, SVM *one-against-all*. This paper efficiently uses SVM with heavy tailed RBF. The SVM-OAA constructs 'n' binary SVM classifier with the i^{th} class separating from all other classes. Each binary SVM classifier creates a decision boundary, which can separate the group it represents from the remaining groups For training data t = $(X_1, Y1), (X_2, Y2)\ldots(Xt, Yt)$; $Xn \in R^n$ and n=1…t, Y_n =1.. K is the class labels corresponding to X_n. The n_{th} SVM solves

$$\begin{aligned} &\text{Min} \quad \tfrac{1}{2}(w^i)^T w^i + C\sum_{i=1} t\,\xi_i w^i b_i \xi_i \\ &(w^i)^T\phi(x_i) + b^i \geq 1 - \xi_i, \text{ if } y_i = i, \\ &(w^i)^T\phi(x_i) + b^i \leq -1 + \xi_i, \text{ if } y_i \neq i, \\ &\text{and } \xi_i \geq 0, j = 1\ldots t, \end{aligned} \qquad (1)$$

Where each data x_i is mapped to the feature space by function ϕ and c, the penalty parameter. The Radial basis function (RBF) is the most popular choice of kernel functions used in Support Vector Machines, which can be represented by

$$K(X_i, X_j) \equiv e^{-\gamma |Xi - Xj|2} \qquad (2)$$

Where X_i is the support vector of the i^{th} class and X_j is the support vector for the new higher dimensional space and γ is the tuning parameter. RBF kernel can be able to give the same decision as that of RBF network. [3].

C. Linear Discriminant Analysis (LDA)

LDA otherwise known as FLDA (Fishers Linear Discriminant Analysis calculates a straight line or hyper plane that separates 2 known classes. Unlike SVM where the hyper plane is chosen to minimize the misclassification errors, LDA chooses hyper plane to minimize within class variance on either side of the line and minimize the between-class variance. Then the side of the line or hyper plane determines the class of the unknown sample. The disadvantage in using LDA is that it can perform classification well only for linearly separable data. Dudoit [5] used LDA to classify cancer gene expression data. In a training set T of size n, (t_i, c_i) is the representation of each tuple, where t_i is in the form $(t_i.X_1, t_i.X_2\ldots. t_i.X_m)$, a vector of expression values of m number of genes in tuple i and class C_i is the class label for the corresponding t_i. LDA tries to find the linear combination $M\alpha$ of M samples that has large ratio of between-class variance (S_e) to within-class variance (S_n) , α being the transformation matrix denoted by

$$M\alpha = \alpha' S_e \alpha \,/\, \alpha' S_n \alpha$$

The extreme values of M is obtained by the eigen values and eigenvectors of the matrix $S_n^{-1}S_e$ which has corresponding eigen

vectors $\nu_1,\nu_2,\ldots\nu_h$ .For any sample t, the discriminant variables are defined as $u_k = t\nu_l$, where l= 1,2…h and h=min (K-1,m) for K number of classes and ν_l maximizes $\alpha' S_e\alpha / \alpha' S_n\alpha$.

C .K- Nearest Neighbour (KNN)

K-Nearest neighbour is the simplest method for deciding the class to which a sample belongs and a popular nonparametric method. KNN classifies a new object based on attributes and training samples. To classify an unclassified vector X , the KNN algorithm ranks the neighbours of X amongst a given set of N data (Xi Ci),i=1,2…N and uses the class labels Cj (j=1,2..k) of the K most similar neighbours to predict the class of the new vector X. The classes of these neighbours are weighted using the similarity between X and each of its neighbours measured by the Euclidean distance metric. Then X is assigned the class label with the greatest number of votes among K nearest class labels.

D.Algorithm

Step1: Randomly divide the dataset for training and testing

Step2: For the training dataset, rank all genes Step3: (Gene selection Stage 1) Perform pair wise gene ranking from the top ranked genes

Step 4: (Gene selection Stage 2) Use the Gene pairs selected from stage 1 for training.

Step 5: Validate the classifier using 5 fold cross validation method

Step 6: Use the gene pairs that achieved 100% CV accuracy and retrain the classifier.

Step 7: Use the classifier to predict the samples in the testing database

III RESULTS AND DISCUSSION

The proposed methodology was applied to the publicly available cancer database namely Liver, Lymphoma and Leukaemia cancer databases. The description of the database is shown in the Table.1. This section reports the experimental results of all the datasets exhibiting the SVM, KNN and LDA classifiers. The dataset was with few missing data. The K-Nearest neighbour algorithm as used by [15] with k=3 was used and the missing data were filled. Half of the samples were picked randomly for training and all the samples for testing. For the training dataset with m x n dimensions ANOVA p-value was calculated for each gene and the top ranked genes were selected .All possible combinations of the top n genes were generated. For n number of top genes all possible combinations are n (n+1)/2. All these combinations were ranked ANOVA p-values. The gene pairs that were above the threshold value were used to train the classifier. The

performance of the classifier was validated using cross validation (CV) technique with 5-folds. For 5 folds, the samples in the training dataset was randomly divided into 5 equal parts, classification was performed for 5 runs, using the 4 parts as training and the other as testing. Each time the classifier was trained, a different test set was used, so that over 5 runs of the classifier, all the samples were used as test set. The average 5-fold accuracy for each run was calculated and average error rate in training were calculated. Fig.(**1**) shows the average Training accuracy for all possible gene pairs using 3 different classifiers SVM, KNN and LDA. In all the cases SVM one-against-all was superior to all other methods were well proven. Then the classifier was retrained with gene pairs that achieved 100% CV accuracy. Then the classifier was used to predict the samples in the testing dataset. In the earlier work of V. Saravanan et al [14] gene combination of 3 was used for training and all those combinations that achieved 100% CV accuracy were retrained with the SVM and LDA classifiers. The learning results showed a maximum of 100% testing accuracy for the lymphoma dataset and 99.39% for the liver dataset and 97.22% for the leukaemia dataset. Later in the work of V.Saravanan et al [15], the number of gene combinations was reduced to 2 aiming to achieve better results using few numbers of genes. The work well exposed all the SVM varieties such as one against all, one against one with 2 different kernel functions such as Gaussian and heavy tailed RBF. The work also proved the SVM one against all with RBF kernel was the best compared with other SVM varieties. The maximum testing accuracy results were 98.39% for lymphoma dataset and 97.44% for the Liver dataset with just 2 genes. The proposed work in this paper also used gene pairs to better predict the type of cancer but using a two-stage gene selection method, which reduced computational burden very much.

Fig.(1) Comparison of 5_fold Cross Validation accuracy results of the three datasets.[X axis denotes the top selected genes and Y axis denotes the number of gene pairs that achieved 100% CV accuracy 1(a) Lymphoma data,1(b) Liver data and 1(c) Leukaemia data],KNN-K Nearest neighbour, LDA-Linear Discriminant Analysis, SVM-Support Vector Machines. In the two-Stage method ANOVA was performed for all individual genes and all gene pairs generated from the top selected genes where again performed ANOVA and the gene pairs that achieved a p-value greater than 0.5 were chosen for training. Hence the method could able to train not all possible gene pairs but only the gene pairs that crossed the threshold, thereby reducing the computational burden in training all possible gene pairs. Let us suppose that one gene pair needs 0.16 seconds to train. Then, all possible gene pairs from the top selected 10 genes were 45 pairs * 0.16 = 7.2 seconds. Similarly for the top selected 100 genes the time taken was 792 seconds that is 13.2 minutes to train. But in the case of two-Stage method, all possible gene pairs were not trained but only important gene pairs selected using ANOVA p-values were used for training.Table.3 shows the number of gene pairs selected for training using the method without pair wise gene ranking [14, 15] and two-Stage gene selection method. All of the methods were carried out under MATLAB environment on Pentium Centrino, 2.0 GHz and 2GB

memory. Using the selected gene pairs, SVM classifier was trained and cross validated. The gene pairs that achieved 100% CV accuracy i.e., no errors in training were chosen for classification. From the Table.3, for the lymphoma dataset the gene pairs filtered were 6 out of 45 pairs using two-stage method. Then the training time would be just 0.96 seconds (6* 0.16). Class labels were assigned for the testing data comprising all samples. The best testing prediction accuracy using the two-stage was 100% using the lymphoma dataset. The testing results for all the three datasets are shown in Table.2.The table depicts the best testing accuracies that could be achieved by the gene pairs generated from top selected 10,20,30,50,100 genes for the method without pair wise gene ranking [14, 15] and the proposed two-stage method. For all the three datasets, accuracies in two-stage were promising. Furthermore, Table.4 shows a comparison of the proposed method with the previous results. For the Leukaemia dataset Alireza Osareh et al [1] used 25 to 1000 top ranked features to classify ALL and AML samples. In Comparison to Alireza Osareh et al's [1] work showing a classification accuracy of 95.2%, the proposed two-stage method produced a best testing accuracy of 97.22% with only two genes. Table.4 shows the average prediction accuracy for the gene pairs generated from top 10,20,30,50,100 genes for Lymphoma dataset. It is clear from the table that the average testing accuracy for two-Stage for most of the cases was slightly higher compared with the method without pair wise gene ranking, KNN and LDA classifiers. In few cases KNN performed well when compared to all methods, but the maximum prediction accuracy achieved by KNN was 98.39% for Lymphoma data. But in the case of the proposed method the maximum testing accuracy was up to 100%. For all the three dataset training and testing was performed using SVM-OAA choosing RBF kernel function. It should be noted that all the genes after ranking were given numbers in ascending order. Overall SVM outperformed KNN and LDA classifiers, since SVMs are less sensitive to high dimensionality and robust to outliers with respect to 5 fold CV accuracy, KNN performed well and are the second best classifier. Finally, LDA produced the worst classification accuracy. From the Plot in Fig (**2**) the gene pairs (13,23) for liver dataset a doctor can able to diagnose that a patient has HCC if and only if the expression level is less than 0.75 and greater than –0.3 otherwise the tissue is of a patient without tumour. Similarly for the lymphoma dataset Fig (**3**) shows a clear separation of all the three subtypes of lymphoma cancer for the gene pair (4,6). The figure plots the gene pairs that achieved 100% CV accuracy in training. The results in Table.5 and table.6 show the importance of pair wise gene selection. Table .5 shows the average testing accuracy of the three datasets for the top ranked genes without pair wise ranking. As indicated in Table.5 the average performance was 89.6% using the method without pair wise gene ranking and Table.6 showing an average performance of 93.53% for proposed two stage method thus improving the performance by 3.93%. Furthermore Table.7 illustrates the maximum accuracy achieved by previous methods and the proposed method. Although the proposed two-stage method could achieve a maximum accuracy of

100% like the method with 3 gene combination [14], the computational burden in training all gene pairs is reduced in the proposed method because most of the uninteresting gene pairs are filtered.

CONCLUSION

The paper focussed on finding the best gene pair, which can give better prediction accuracy with less computational burden for three publicly available dataset. In particular the paper also investigated the performance of the classifiers: SVM, KNN, LDA and proved SVM to be the best classifier for the proposed model. Comparative studies have been performed between the methods without pair wise gene ranking and two-stage gene selection methods and proved for an improved classification performance of 3.93%. Finally, amongst all methods the best prediction accuracy is achieved in the two-stage method using just two genes and effective results with the previous work were well proven.

REFERENCES

[1] Alireza Osareh, Bita Shadgar (2009) *Journal of Applied Sciences 9(3):459-468,ISSN 1812-5654.*

[2] Andrew D. Keller, MichH Schummer, Lee Hood, Walter L. Ruzzo (2000), *Technical Report UW-CSE-2000-08-01*

[3] Chih-wei Hsu and chih jen Lin (2002) *IEEE transactions on neural networks.*

[4] Dov Stekel (2003) Microarray Bioinformatics, Cambridge university press, ISBN 0 521-670500.

[5] Dudoit, S., Fridlyand, J., & Speed, T. (2002) *Journal of the American Statistical Association*, 97, 77–87.

[6] Elena Marchiori, Michele Sebag (2005) *Evo Workshops PP.74-83.*

[7]Hong Chai and Carlotta Domeniconi (2001) *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*.

[8]J. Jaeger, R. Sengupta, W.L. Ruzzo (2003) *Pacific Symposium on Biocomputing* 8:53- 64.

[9] Juan Liu, Hitoshi Iba (2001) *Genome Informatics* 12: 14–23.

[10] Li Y., Campbell C., Tipping M. (2002) *Bioinformatics* , **18:**1332-1339.

[11] Lipo Wang, Feng Chu, and Wei Xie (2007) *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. 4, no. 1, January-march.

[12] Li-Yeh Chuang,Chao-Hsuan Ke, Hsueh-Wei Chang,Cheng-hong Yang (2009) *OMICS A journal of Integrative Biology,Volume.13,number 2.*

[13] Mingjun Song, Sanguthevar Rajasekaran (2007) *21st International Conference on Advanced Information Networking and Applications workshop, IEEE.*

[14] Saravanan V., Mallika R. (2009) *International conference on e-CASE and e-Technology.*

[15] Saravanan V., Mallika R. (2009) *International conference on computer engineering and technology,IEEE.*

[16] Tibshirani R., Hastie T., Narasimhan B., Chu G. (2003) *Statistical Science*, vol. 18, pp. 104-117.

[17] Venu Satuluri (2007) March 15.

[18] Yeo lee chin, Safaai Ders, *Jurnal Teknologi (D) 43 (D). pp. 111-124. ISSN 0127-9696*

[19] Yoonkyung Lee, Cheo Koo Lee (2003) Vol. 19 no. 9, B*ioinformatics*

[20] Yvan Saeys, I naki Inza and Pedro Larra naga (2007) *Bioinformatics* Advance Access published August .

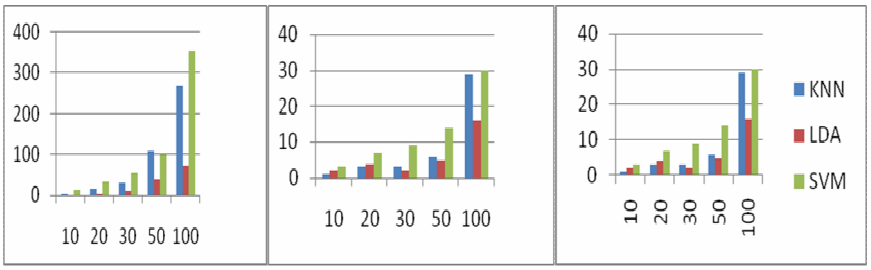

Fig 1(a) Fig 1(b) Fig 1(c)

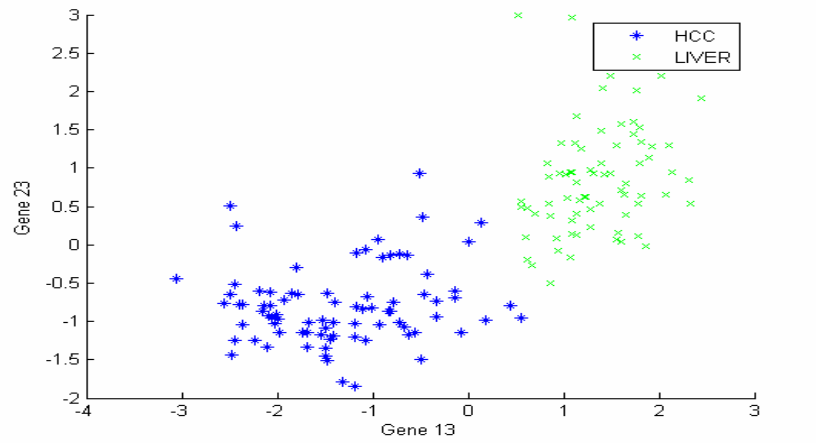

Fig.(2) Plot showing best separation for liver dataset that achieved the best testing accuracy for Liver data

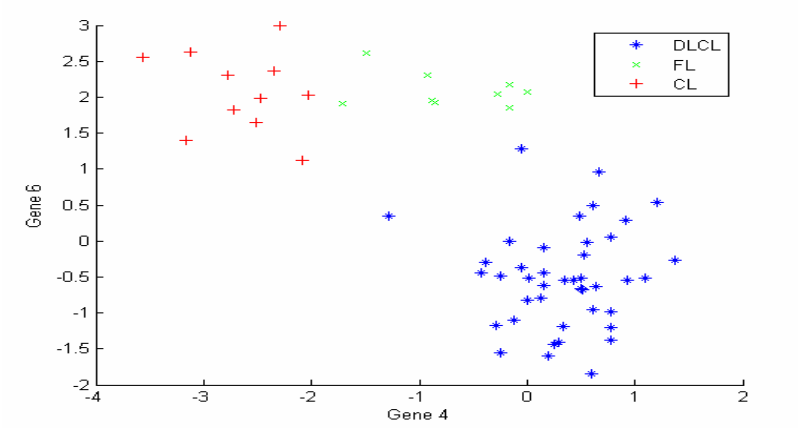

Fig.(3) Gene Expression level of gene pairs showing good separation of different classes for Lymphoma dataset

Table 1.Description of the database used

Datasets	No of features	No of samples	Tasks
Lymphoma Cancer	4026	62	Large B-cell (DLBCL),follicular Lymphoma (FL),chronic lymphocytic leukaemia (CLL)
Liver cancer	1648	156	Non-tumour liver and HCC
Leukaemia cancer	7129	72	Acute lymphoblastic (ALL) and acute myelogenous (AML).

Table 2. Comparison of the Best Testing Accuracies for the top selected genes for the three datasets using method without pair wise gene ranking and Two-Stage method

Top genes	10	20	30	50	100
LYMPHOMA DATA					
Method without pair wise gene ranking	98.39	100	100	100	100
Two-Stage	100	100	100	100	98.39
LIVER DATA					
Method without pair wise gene ranking	94.87	97.44	95.51	94.87	92.31
Two-Stage	97.44	98.08	95.51	96.69	96.79
LEUKAMIA DATA					
Method without pair wise gene ranking	94.44	93.06	94.44	94.44	94.44
Two-Stage	95.83	95.83	95.83	95.83	97.22

Table 3. Gene pair selection comparison between single gene selection method and Two-Stage method

# genes	Possible gene pairs	Gene pairs selected for training in single gene selection method	Gene pairs selected for training using Two- Stage method		
			Lymphoma	liver	Leukaemia
10	45	45	6	22	26
20	190	190	25	48	121
30	435	435	66	116	296
50	1225	1225	169	418	892
100	4950	4950	844	2075	3770

Table.4 Average testing accuracy results-Lymphoma data

#genes	Method without pair wise gene ranking	Proposed two stage method	KNN	LDA
10	95.75	100	96.24	96.77
20	94.96	97.21	95.70	87.10
30	94.96	96.01	92.92	88.55
50	92.16	93.55	91.61	76.27
100	78.74	91.34	90.94	75.48

Table.5: Average testing accuracies (%) without pair wise Gene ranking computed over all data sets

# gene	Lymphoma	Liver	Leukaemia	*Average*
10	100	97.12	95.83	***97.65***
20	97.21	97.76	85.69	***93.55***
30	96.01	94.62	91.45	***94.02***
50	93.55	87.01	88.61	***89.72***
100	91.34	89.66	97.22	***92.74***

Average performance: ***93.53***

Table.6. Average testing accuracies (%) using two stage gene selection method computed over all data sets

# gene	Lymphoma	Liver	Leukaemia	*Average*
10	95.75	91.54	68.98	***85.42***
20	94.96	93.35	87.7	***92.003***
30	94.96	93.12	93.98	***94.02***
50	92.16	88.14	93.85	***91.38***
100	78.74	87.45	82.79	***82.99***

Average Performance: ***89.16***

Table.7 Maximum Accuracy Comparison for the three datasets

Method	Lymphoma Accuracy	Liver Accuracy	Leukaemia Accuracy
3 genes [14]	100%	99.39 %	97.22%
2 genes [15]	98.39%	97.44%	-
2 genes-proposed two-stage method	100%	98.08%	97.22%