Jurnal Teknologi, 43(D) Dis. 2005: 111–124

A STUDY ON GENE SELECTION AND CLASSIFICATION ALGORITHMS FOR CLASSIFICATION OF MICROARRAY GENE EXPRESSION DATA

YEO LEE CHIN[1]* & SAFAAI DERIS[2]

Abstract. The development of microarray technology allows researchers to monitor the expression of genes on a genomic scale. One of the main applications of microarray technology is the classification of tissue samples into tumor or normal tissue. Gene selection plays an important role prior to tissue classification. In this paper, a study on numerous combinations of gene selection techniques and classification algorithms for classification of microarray gene expression data is presented. The gene selection techniques include Fisher Criterion, Golub Signal-to-Noise, traditional t-test and Mann-Whitney rank sum statistic. The classification algorithms include support vector machines (SVMs) with several kernels and *k*-nearest neighbor (*k*-nn). The performance of the combined techniques is validated by using leave-one-out cross validation (LOOCV) technique and receiver operating characteristic (ROC) is used to analyze the results. The study demonstrated that selecting genes prior to tissue classification plays an important role for a better classification performance. The best combination is obtained by using Mann-Whitney Rank Sum Statistic and SVMs. The best ROC score achieved for this combination is at 0.91. This should be of significant value for diagnostic purposes as well as for guiding further exploration of the underlying biology.

Keywords: Microarray gene expression data, gene selection, statistical methods, classification algorithms, support vector machines, *k*-nearest neighbor

Abstrak. Pembangunan teknologi *microarray* membenarkan penyelidik untuk meneliti tahap ekspresi gen dalam sel. Salah satu aplikasi teknologi *microarray* adalah pengkelasan sampel tisu kepada tisu kanser atau tisu biasa. Pemilihan gen memainkan peranan yang penting sebelum pengkelasan. Dalam makalah ini, beberapa kombinasi teknik pemilihan gen dan teknik pengkelasan yang berlainan untuk pengkelasan data expresi gen *microarray* telah dikaji. Teknik pemilihan gen terdiri dari *Fisher Criterion*, *Golub Signal-to-Noise*, *traditional t-test* dan *Mann-Whitney rank sum statistic*. Teknik pengkelasan terdiri dari *support vector machines* (*SVMs*) dengan pelbagai *kernel* dan *k-nearest neighbor* (*k-nn*). Prestasi kombinasi teknik-teknik yang dikaji disahkan dengan menggunakan teknik *leave-one-out cross validation* (*LOOCV*) dan *receiver operating characteristic* (*ROC*) digunakan untuk menganalisa prestasi kombinasi teknik-teknik yang dikaji. Kajian yang telah dijalankan dalam eksperimen ini menunjukkan bahawa pemilihan gen sebelum pengkelasan adalah penting untuk memperolehi prestasi pengkelasan yang lebih baik. Kombinasi yang menghasilkan prestasi tertinggi adalah dengan menggunakan *Mann-Whitney rank sum statistic* dan *SVMs*. Nilai *ROC* tertinggi yang dicapai oleh kombinasi ini adalah 0.91. Ini adalah penting bagi tujuan rawatan dan kajian biologi seterusnya.

Kata kunci: Data expresi gen *microarray*, pemilihan gen, kaedah statistik, algoritma pengkelasan, *Support Vector Machines*, *k-nearest neighbor*

[1&2]Artificial Intelligence and Bioinformatics Laboratory, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
* Corresponding author: Email: yeoleechin@hotmail.com

1.0 INTRODUCTION

The development of microarray technology allows researchers to monitor thousands of gene expression levels in a simple microarray experiment [1-4]. Classification of tissue samples into tumor or normal tissue is one of the applications of microarray technology. There is a variety of different existing classification algorithms that can be used for tissue classification such as the Fisher linear discriminant analysis [5], *k*-nearest neighbor (*k*-nn) [6], and support vector machines (SVMs) [7]. However, there are some challenges due to the characteristics of the data that make tissue classification a non-trivial task, i.e. high dimensionality of the data, where the data usually contains thousands of genes and the available tissue samples is very small, some have sizes below 100. Most genes are irrelevant to tissue distinction and the data set might contain noise. Therefore, gene selection plays an important role prior to tissue classification. Performing gene selection helps to reduce data size thus improving the classification running time. More importantly, gene selection removes a large number of irrelevant genes which improves the classification performance.

There are many different gene selection techniques exist. Initially, selection of important genes was simply carried out by comparing the ratio of expression levels between the two tissues, a method known as fold change approach. They set an arbitrary cutoff level for the value of the fold change ratio (eg. 2) and declare any gene with a higher observed ratio of expression in the two tissues to be important genes [8]. However, this approach is known to be unreliable [9] because statistical variability was not taken into account. Since then, many more sophisticated statistical methods have been proposed. One approach is by defining a null hypothesis of equal tumor and normal mean expression levels for each gene in the data set. For each gene, some statistical methods are used such as the traditional t-test or Welch approximation to calculate the statistical score [10-12]. Large score suggests that the corresponding gene has different expression levels in the tumor and normal tissue and thus is an important gene and will be selected for further analysis. Besides that, some researchers used a variation of correlation coefficient to select genes, for example, Fisher Criterion [13] and Golub Signal-to-Noise [14]. Non-parametric tests like threshold number of misclassification (TNoM) which calculates a minimal error decision boundary and counts the number of misclassifications done with this boundary or the Park non-parametric scoring algorithm [15], (which is identical to the Wilcoxon / Mann-whitney [10] rank sum statistic) are also being used as a gene selection technique for microarray data.

In this paper, different combination of gene selection techniques and classification algorithms for tissue classification were studied. They are the four gene selection techniques including Fisher Criterion, Golub Signal-to-Noise, traditional t-test, and Mann-Whitney rank sum statistic and the two classification algorithms, including SVMs with several kernels and *k*-nn. The performance of the combined techniques was validated by using leave-one-out cross validation (LOOCV) technique and the

receiver operating characteristic (ROC) was used to analyze the results. The numbers of genes being selected for the experiment were from 1 to 100. Colon data set was used for the case study as it is one of the leading causes of death in Malaysia [16]. Results show that a better classification performance is achieved if important genes are selected prior to classification task.

2.0 DATA

The colon data set were applied. It is a collection of expression levels from colon biopsy samples reported by Alon *et al.* [17]. The data set consisted of 62 tissue samples of colon epithelial cells. These samples were collected from colon-tumor patients. The tumor biopsies were collected from tumors, and the normal biopsies were collected from healthy parts of the colons of the same patients. Each sample contains more than 6,500 human genes measured using high density oligonucleotide arrays. 2000 genes with the highest minimal intensity across the 62 samples were chosen for the analysis.

3.0 METHODOLOGY

The experimental design and the gene selection techniques are discussed in Section 3.1 and 3.2. Section 3.3 discusses the classification algorithms and the validation and evaluation methods are discussed in Section 3.4.

3.1 Overall Experimental Design

The overall experimental design is shown in Figure 1. The design is interpreted from left to right, from A to H. The microarray experiment was designed and being carried out (in A) to get different images. These images were then converted to numerical data (in B) which was arranged in a table with all the genes included. Different gene selection techniques were applied to the full table (in C) and only the

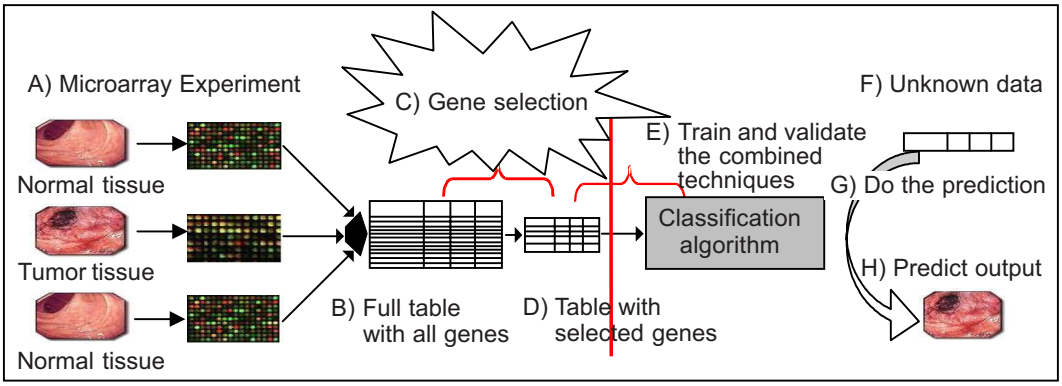

Figure 1 Overall experimental design

selected genes (in D) were used for training and validation (in E). Finally, an unknown tissue sample (in F) can use the trained classification algorithm for the prediction (in G) and the classification algorithm will predict the unknown tissue sample as tumor or normal tissue (in H). The research starts from B and ends with training and validation in E.

3.2 Gene Selection Techniques

The Fisher Criterion [13], Golub Signal-to-Noise [14], traditional t-test [10] and Mann-Whitney rank sum statistic [10] were applied to calculate the statistical score, S, for the genes. In these techniques, each gene was measured for correlation with the class according to some measuring criteria in the formulas. The genes were ranked according to the score, S, and the top ranked genes were selected.

The Fisher Criterion, *fisher*, is a measure that indicates how much the class distributions are separated. The coefficient has the following formula:

$$fisher = \frac{(\mu_1 - \mu_2)^2}{(v_1 + v_2)} \tag{1}$$

where μ_i is the mean and v_i is the variance of the given gene in class i. There were two tissue classes in this experiment, i.e. the tumor tissue and the normal tissue. The statistic gives higher scores to genes whose means differ greatly between the two classes, relative to their variances.

Golub used a measure of correlation that emphasizes the "Signal-to-Noise" ratio, *signaltonoise*, to rank the genes. It is very similar to the Fisher Criterion but use another related coefficient formula as shown below:

$$signaltonoise = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + \sigma_2)} \tag{2}$$

where μ_i is the mean and σ_i is the standard deviation of the gene in class i.

Traditional t-test, *ttest*, assumes that the values of the two tissues variances are equal. The formula is as follows:

$$ttest = \frac{(\mu_1 - \mu_2)}{\sqrt{(v_p / n_1) + (v_p / n_2)}} \tag{3}$$

whereμ_i is the mean of the gene in class i and v_p is the pooled variance,

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \text{ and } \frac{s_i^2 = \sum_{i=1}^{n_i} (x_i - \bar{x})^2}{n_i - 1}$$

The Mann-Whitney, *mann*, has the following formula:

$$mann = \frac{n_1 * n_2 * (n_1 + 1)}{2 - r_1} \tag{4}$$

where n_i is the sizes of class i, and r_1 is the sum of the ranks in class 1.
The score, S, for each gene is thus the score calculated by using the formula in these statistical techniques.

3.3 Tissue Classsfication

Two classification algorithms were used to evaluate the validity of the selected genes. They are the SVMs [18] with different kernels and the *k*-nn [19].

3.3.1 Support Vector Machines (SVMs)

SVMs are relatively new types of classification algorithms. An SVM expects a training data set with positive and negative classes as an input (i.e. a binary labeled training data set). It then creates a decision boundary (the maximal-margin separating boundary) between the two classes and selects the most relevant examples involved in the decision process (the so-called support vectors). The construction of the linear boundary is always possible as long as the data is linearly separable. If this is not the case, SVMs can use kernels, which provide a nonlinear mapping to a higher dimensional feature space.

The dot product has the following formula:

$$K(x, y) = (x \cdot y + 1)^d \tag{5}$$

where x and y are the vectors of the gene expression data. The parameter d is an integer which decides the rough shape of a separator. In the case where d is equals to 1, a linear classification algorithm is generated, and in the case where d is more than 1, a nonlinear classification algorithm is generated. In this paper, when d is equals to 1, it is called the SVM dot product, when d is equals to 2, it is called the SVM quadratic dot product and when d is equals to 3, it is called the SVM cubic dot product. The radial basis kernel is as follows,

$$K(x, y) = \exp\left(\frac{-|x - y|^2}{2\sigma^2}\right) \tag{6}$$

where σ is the median of the Euclidean distances between the members and non-members of the class.

The main advantages of SVMs are that they are robust to outliers, converge quickly, and find the optimal decision boundary if the data is separable [7]. Another advantage is that the input space can be mapped into an arbitrary high dimensional

working space where the linear decision boundary can be drawn. This mapping allows for higher order interactions between the examples and can also find correlations between examples. SVMs are also very flexible as they allow for a big variety of kernel functions.

3.3.2 k-nearest Neighbor

The *k*-nn classification algorithm is a simple algorithm based on a distance metric between the testing samples and the training samples. The main idea of the method is, given a testing sample *s*, and a set of training tuples *T* containing pairs in the form of (t_i, c_i) where t_i'*s* are the expression values of gene *i* and c_i is the class label of gene *i*. Find *k* training sample with the most similar expression value between *t* and *s*, according to a distance measure. The class label with the highest votes among the *k* training sample is assigned to *s*. The main advantage of *k*-nn is it has the ability to model very complex target functions by a collection of less complex approximations. It is easy to program and understand. No training or optimization is required for this algorithm. It is robust to noisy training data.

3.4 Validation and Evaluation Methods

LOOCV was used to validate the combined techniques. Under LOOCV, assuming that there are *n* samples, the combined technique is successively trained on *n*-1 samples and tested on the left out sample. This was performed through the entire dataset, leaving out different sample each time. The ROC score was used to analyze the entire performance. ROC score is the area under the ROC curve, which takes into account both false negative and false positive errors and it reflects the robustness of the classification. Perfect classification gives an ROC score of one, whereas a random classification has an expected ROC score close to 0.5.

4.0 RESULTS

In this section, the impact and importance of gene selection to the classification performance was first studied. This was carried out by comparing the classification performance by using all genes and gene selected by statistical techniques in Section 4.1. The classification performance for each classification algorithm is discussed in Section 4.2. The effectiveness of each statistical technique to the best classification algorithm is discussed in Section 4.3.

4.1 Importance of Gene Selection Technique Prior to Tissue Classification

Figure 2 shows the classification performance by using all genes and gene selected by using statistical techniques. The ROC scores recorded for the gene selection

techniques in the figure are the average ROC scores for number of genes selected from 1 to 100.

From the figure, by using all genes, the best performance was obtained by using SVMs with radial basis function while *1*-nn, *2*-nn and *5*-nn have the worst performance. *3*-nn and *4*-nn were comparable to each other when all genes were used. The performances of the classification algorithms improved after genes were selected by gene selection techniques especially for *k*-nn classification algorithm. This shows the importance of applying gene selection techniques to select important genes prior to the classification task. Applying gene selection techniques in selecting genes helps in removing a large number of irrelevant genes which improves the classification performance. Since one of the advantages of SVMs is it is robust to outliers and allows nonlinear classification to be done, gene selection techniques does not give big impaction to its performance, but, a better performance still can be obtained after applying gene selection techniques, which can be seen from the figure. One might ask why there is still a need to do gene selection if the classification performance using SVM has little difference while using all the genes in the data set compare to the selected subset of genes. One reason for this is that selecting sub set of genes not only help biologists to identify the potential genes rather than swimming in the huge data set, it also helps the classification algorithm to build a better and simple rule for classifying future unknown data.

This figure shows that a better classification performance can be achieved if genes are first selected by the gene selection techniques. However, which combination of

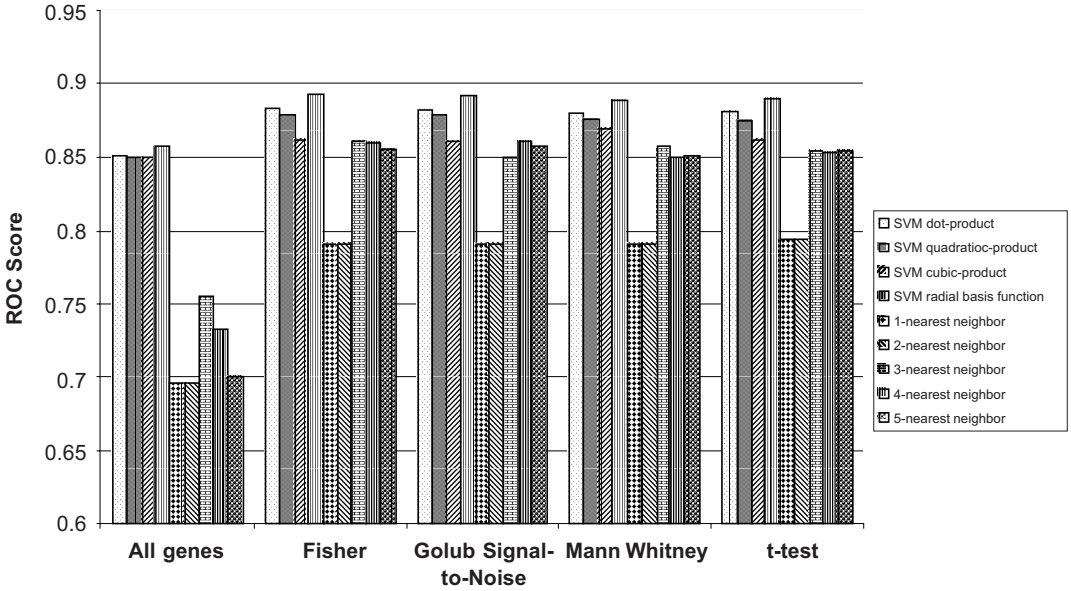

Figure 2 Classification performances by using all genes and genes selected by statistical techniques

statistical techniques and classification algorithm and how many genes are needed for the best performance? The next section will answer this question.

4.2 Classification Performance Between Different Classification Algorithms

Figure 3 shows the classification performance using SVMs with different kernels after gene selection by using statistical techniques.

Figure 3 shows that SVM radial basis function performs the best. Of the three product kernels, dot-product and quadratic product have better ROC score than cubic-product. These results indicate that overfitting causes the misclassification for the cubic-product kernel. If more samples are obtained and they are not separable

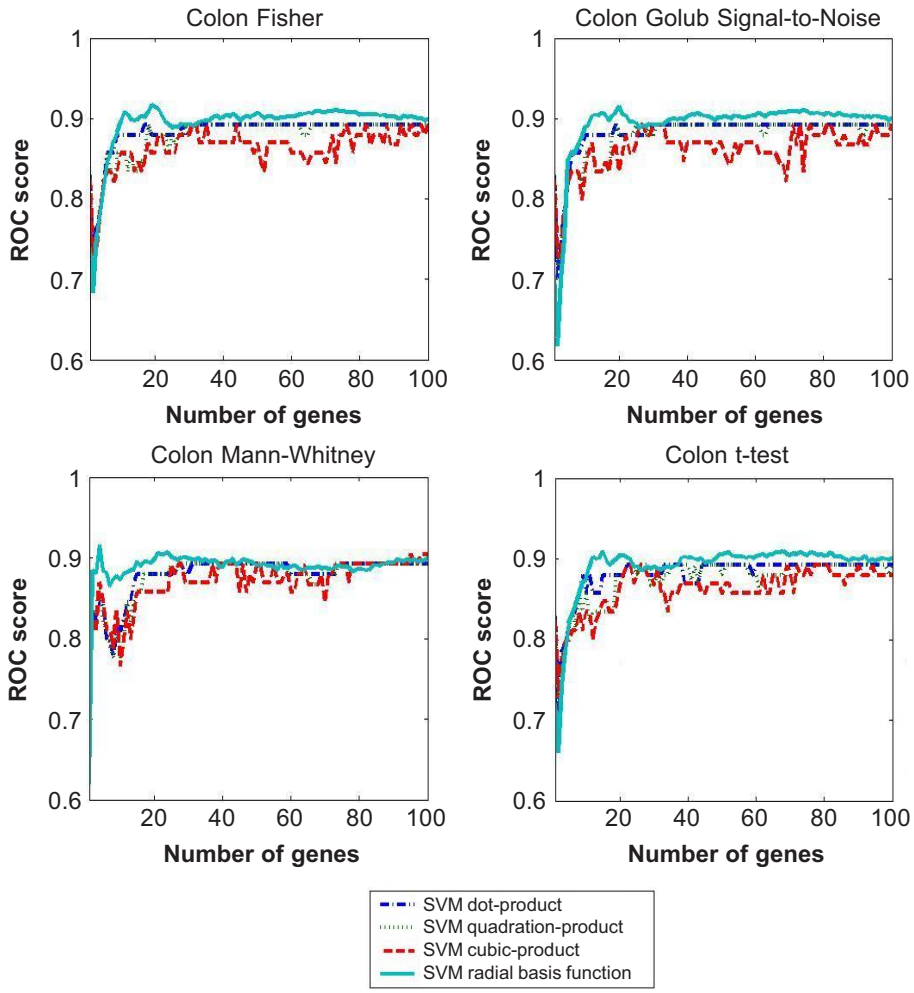

Figure 3 Classification performance using SVMs with different kernels after gene selection by using statistical techniques

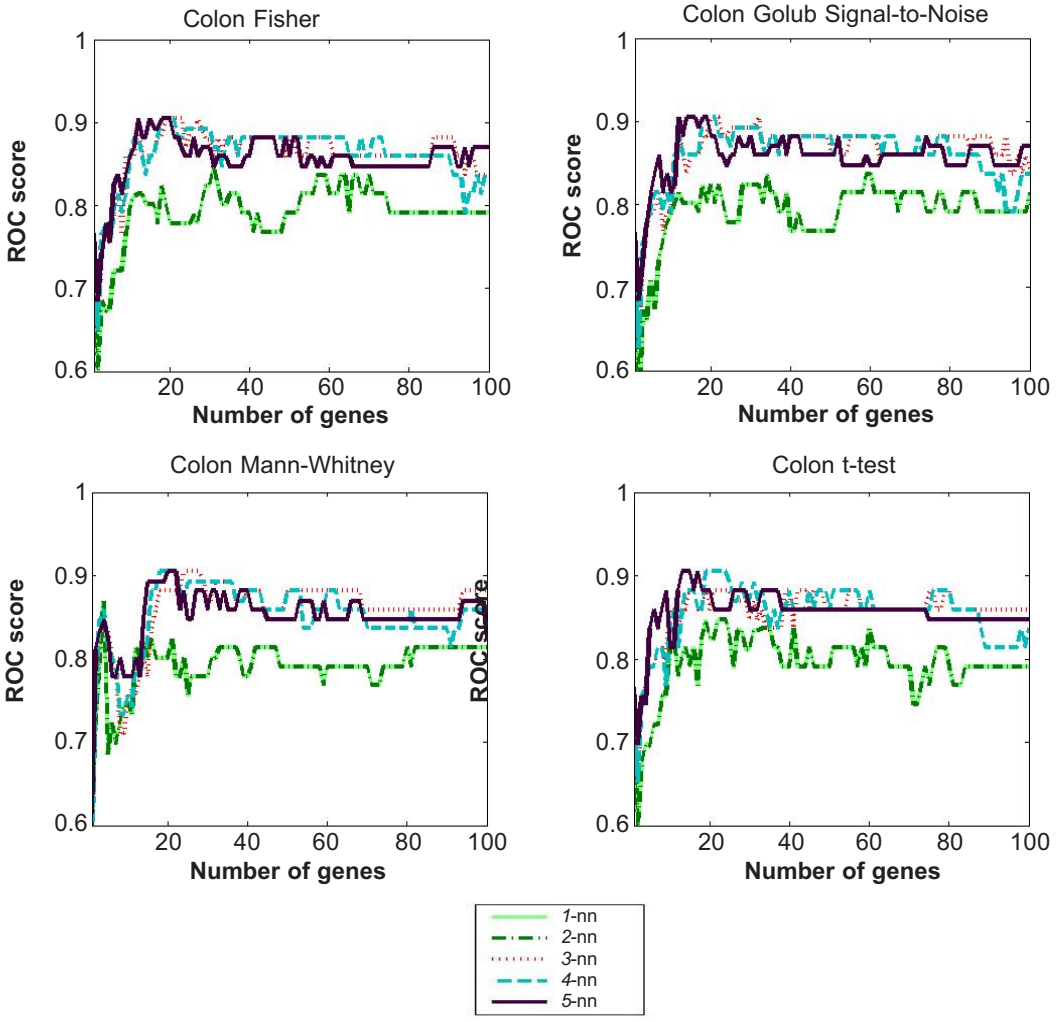

Figure 4 Classification performance using different *k*-nn after gene selection by using statistical techniques

linearly, nonlinear classification may perform well [20]. Figure 4 shows the classification performance by using *k*-nn with different number of *k* between 1 and 5.

Figure 4 shows that *k*-nn with *k* more than 2 outperform *k* which is equals to 1 and 2. One of the reasons for this to happen is that in the case of mislabeled training samples, it will have much greater effect on the classification result of *1*-nn since one mislabel will result in misclassifying the test sample. *3*-nn and *4*-nn is less prone to bias in the data and more tolerable to noise since it makes use of several training samples to determine the class of a test sample.

Figure 5 shows the classification performance between different classification algorithms after gene selection using statistical techniques (the best classification

algorithm is selected from SVM and *k*-nn based on their ROC score from the experiments).

Figure 5 shows that SVM with radial basis function as the kernel function always produced higher ROC score than *3*-nn. SVM produced most stable results when the genes are greater than 15. Generally, the results have lower ROC score with fewer genes for both classification algorithms. Lowest scores always drop between the numbers of genes from 1 to 15 except for Mann-Whitney Rank Sum Statistic. One reason for the lower scores might due to the characteristic of genes itself where genes do not act alone, but they interact with other genes for certain functions [21]. For example, if genes A and B are in the same function it could be that they have

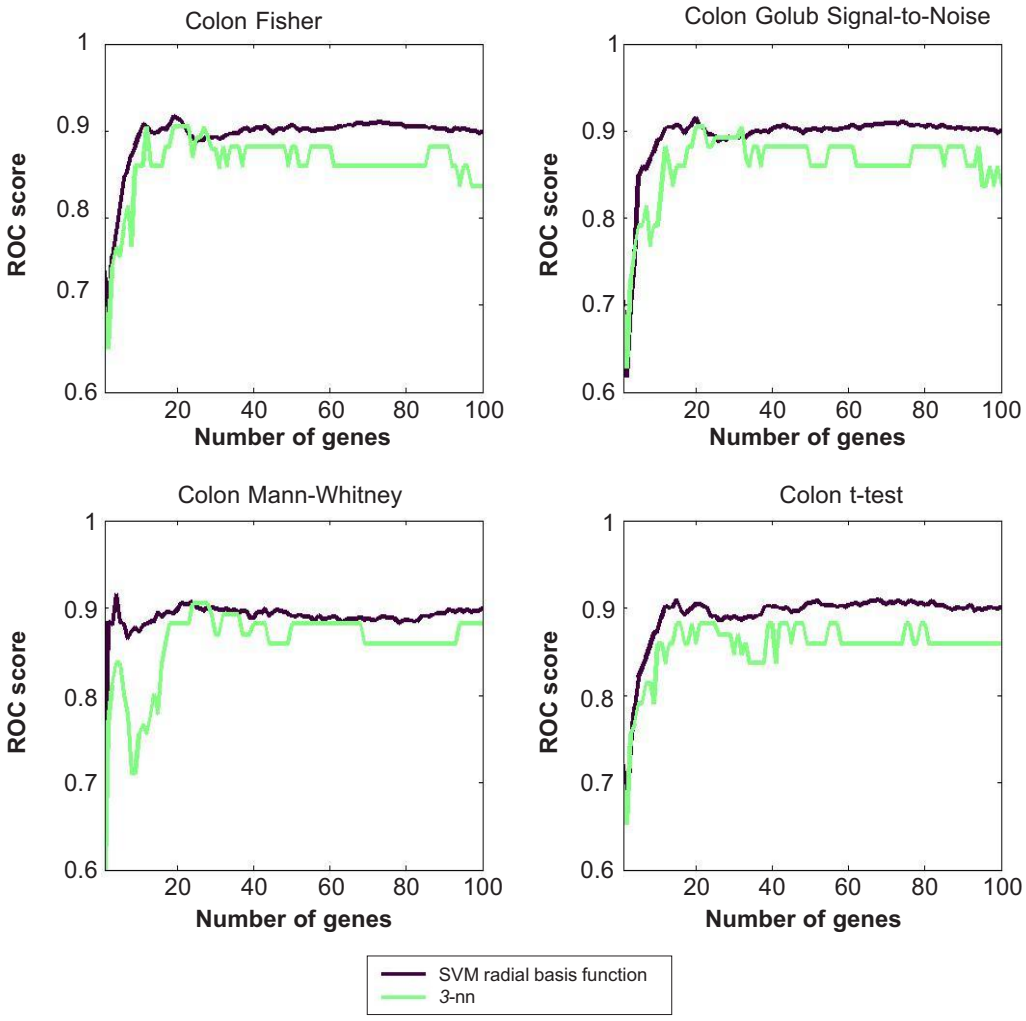

Figure 5 Classification performance between different classification algorithms after gene selection using statistical techniques (the best classification algorithm is selected from SVM and *k*-nn)

similar regulation and therefore, similar expression profiles. If gene A has a good discriminative score it is highly likely that gene B will, as well. Hence the statistical techniques are likely to include both genes in a classification algorithm, yet the pair of genes provides little additional information compared to either gene alone. If there are 5 functions in the dataset, 10 genes for each function, and if the genes in first function score highest in the gene selection score, so these 10 genes might be selected for the classification. In this case, the genes being selected are highly redundant, and provide little additional information.

4.3 Classification Performance Between Different Statistical Techniques

Figure 6 shows the classification performance between different statistical techniques for the number of genes between 1 and 30 where the best classification algorithm from Section 4.2 is used.

Figure 6 shows that by using Mann-Whitney Rank Sum Statistic as the statistical technique, higher ROC score can be obtained for the lesser number of genes, i.e. 2 to 5, compared to other statistical techniques. An explanation might be that Mann-

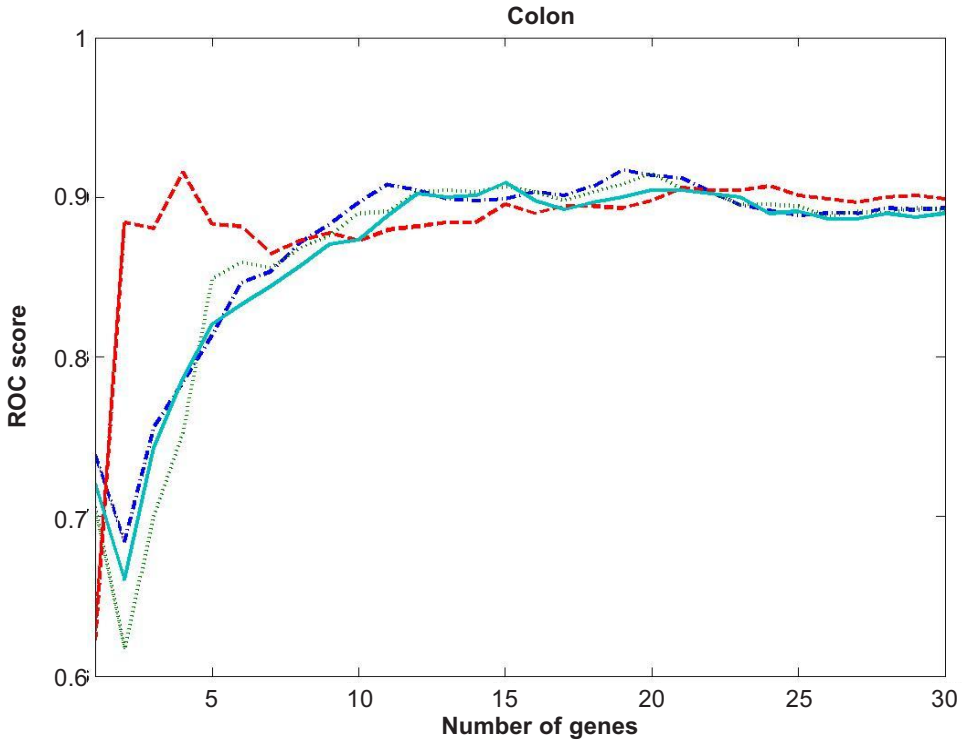

Figure 6 Classification performance between different statistical techniques (the best classification algorithm is used)

Whitney Rank Sum Statistic, as a non-parametric test, extracts less correlated genes and therefore, does a good job in selecting genes from different pathways. However, the ROC score for Mann-Whitney Rank Sum Statistic drops when the genes are more than 5 while other gene selection techniques have better ROC score for genes more than 5. One explanation is when Mann-Whitney Rank Sum Statistic selects genes from different pathways where some of them are irrelevant for the sample tissues, this can disturb the classification performance. While for other statistical techniques, inclusion of more genes produce better results because when 2 or 3 genes are selected, these genes might come from the same pathway and give little information for the classification algorithm but with more genes, it gives more information. More stable results are obtained when the number of genes is more than 15.

5.0 CONCLUSION

This paper reports the application of different combination of gene selection techniques and classification algorithms to the colon data set. The statistical score of each gene is first calculated and ranked. The top number of genes are selected and used for training the classification algorithms. ROC score of different combination of gene selection techniques and classification algorithms are obtained for analysis. From the observations, the best combination for better performance is obtained by using Mann-Whitney Rank Sum Statistic as the gene selection technique and SVM with radial basis function as the kernel. The best ROC score achieved for this combination is at 0.91.

ACKNOWLEDGEMENTS

We wish to thank Jochen Jaeger, Nazar Zaki and Hany Taher for their technical advice and contributions.

REFERENCES

[1] Lander, E. S. 1999. Array of Hope. *Nature Genetics Supplement*. 21: 3-4.

[2] Stears, L., T. Martinskt, and M. Schena. 2003. Trends in Microarray Analysis. *Nature Medicine*. 9:140-145.

[3] Brown, P. O., and D. Botstein. 1999. Exploring the New World of the Genome with DNA Microarrays. *Nat Genet*. 21(1 Suppl): 33-7

[4] Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. 1996. Parallel Human Genome Analysis: Microarray-based Expression Monitoring of 1000 Genes. Proc. Natl. Acad. Sci. USA, 93:10614-10619.

[5] Dudoit, S., J. Fridlyand, and T. P. Speed. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*. 97(576): 77-87.

[6] Li, L., C. R. Weinberg, T. A. Darden, and L. G. Pedersen. 2001. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity of Choice of Parameters of the GA/*k*-NN Method. *Bioinformatics*. 17(12): 1131-1142.

[7] Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler. 1999. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. Proc. Natl. Acad. Sci. USA. 97: 262-267.

[8] Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. 1996. Parallel Human Genome Analysis: Microarray-based Expression Monitoring of 1000 Genes. Proc. Natl. Acad. Sci. USA. 93: 10614-10619.

[9] Chen, Y., E. R. Dougherty, and M. L. Bittner. 1997. Ratio-based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics.* 2(4): 364-374.

[10] Devore, J. L. 1995. *Probability and Statistics for Engineering and the Sciences.* 4th edition. California: Duxbury Press.

[11] Callow, M. J., S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. 2000. Microarray Expression Profiling Identifies Genes with Altered Expression in HDL-deficient Mice. *Genome Res.* 10(12): 2022-9.

[12] Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed. 2002. Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. *Statistica Sinica.* 12: 111-139.

[13] Pavlidis, P., J. Weston, J. Cai, and W. H. Grundy. 2000. Gene Functional Analysis from Heterogeneous Data. Submitted for publication.

[14] Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caliguiri, C. D. Bloomfield, and E. S. Lander. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science.* 286: 531-537.

[15] Park, P. J., M. Pagano, and M. Bonetti. 2001. A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. Pacific Symposium on Biocomputing. 6: 52-63.

[16] Lim, G. C. 2002. Overview of Cancer in Malaysia. *Japanese Journal of Clinical Oncology.* 32: S37-S42.

[17] Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Cancer Tissues Probed by Oligonucleotide Arrays. *PNAS.* 96: 6745-6750.

[18] Cristianini, N., and J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. New York: Cambridge University Press.

[19] Friedman, M., and A. Kandel. 1999. *Introduction to Pattern Recognition.* London: Imperial College Press.

[20] Domura, D., H. Nakamura, S. Tsutsumi, H. Aburatani, and S. Ihara. 2002. Characteristics of Support Vector Machines in Gene Expression Analysis. *Genome Informatics.* 13: 264-265.

[21] Albertson, D. G., C. Collins, F. McCormick, and J. W Gray. 2003. Chromosome Aberrations in Solid Tumors. *Nature Genetics.* 34: 369-376.