Gene Selection and Classification of Human Lymphoma from Microarray Data

Joarder Kamruzzaman[1], Suryani Lim[1], Iqbal Gondal[1], and Rezaul Begg[2]

[1] Faculty of Information Technology, Monash University, Australia-3842
[2] Centre For Ageing, Rehabilitation, Exercise & Sport,Victoria University, Australia-8001

Abstract. Experiments in DNA microarray provide information of thousands of genes, and bioinformatics researchers have analyzed them with various machine learning techniques to diagnose diseases. Recently Support Vector Machines (SVM) have been demonstrated as an effective tool in analyzing microarray data. Previous work involving SVM used every gene in the microarray to classify normal and malignant lymphoid tissue. This paper shows that, using gene selection techniques that selected only 10% of the genes in "Lymphochip" (a DNA microarray developed at Stanford University School of Medicine), a classification accuracy of about 98% is achieved which is a comparable performance to using every gene. This paper thus demonstrates the usefulness of feature selection techniques in conjunction with SVM to improve its performance in analyzing Lymphochip microarray data. The improved performance was evident in terms of better accuracy, ROC (receiver operating characteristics) analysis and faster training. Using the subsets of Lymphochip, this paper then compared the performance of SVM against two other well-known classifiers: multi-layer perceptron (MLP) and linear discriminant analysis (LDA). Experimental results show that SVM outperforms the other two classifiers.

1 Introduction

The DNA microarray provides information of thousands of genes, which could be harnessed for different purposes. One common use is to separate cancerous from healthy cells using either unsupervised or supervised classifiers [1–3, 17, 21]. Alizadeh *et al.* [1] used unsupervised classifier to group genes having similar expression patterns in order to separate healthy from cancerous cells. In recent years, supervised methods have also been used for this classification task; for example, decision tress, linear discriminant analysis (LDA), multi-layer perceptron (MLP), support vector machines (SVM) and many others [5, 10, 12, 14]. In general, supervised methods have been shown to perform better than unsupervised methods.

Using the microarray in Alizadeh's study, Valentini [21] showed that a supervised method can achieve a significantly higher classification accuracy than that reported by Alizadeh *et al.* [1]. In his study, Valentini trained SVM using all

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 379–390, 2005.

4026 genes in the microarray to separate normal from malignant cells. However, use of such high feature dimensions reduces the efficiency of SVM.

The purpose of this paper is to test whether, by adopting gene (feature) selection techniques in conjunction with SVM, the same level of accuracy can be achieved using only a subset of the total number of genes. Fewer number of genes require less computational time for SVM as an added advantage. Identifying the contributing genes in this process also enables biologists to concentrate on few genes to explore their roles in malignancy development in greater details. We repeated Valentini's experiment by training SVM using the same microarray, and in addition, we trained more SVMs using only about 10% of the original microarray; the subset genes were derived from feature selection techniques. As MLP and LDA have been previously used for classifying microarray, we also compared the performance of SVM with these two methods on the selected subsets.

This paper is organised as follows. Section 2 discusses the methods for obtaining the subsets. Section 3 describes the three classifiers investigated. Section 4 describes the experimental set up, and Section 5 contains the results and discussion of results. Section 6 concludes the paper and provides direction for future work.

2 Gene Selection

Feature selection obtains a subset from a complete set of features and can increase the efficiency of the classifier by reducing redundant and irrelevant features. It can be formally defined as follows. Let S be a subset of X and $S = \{s_1, s_2, \cdots, s_n | s_i \in X, n << ||X||\}$, where n is the number of features in the subset. The feature selection function F selects s_i from X, that is $F : X \to S$. In general, feature selection can be broadly classified into three sub-areas: embedded, filter and wrapper [15]. In this paper, we concentrated on filter-based feature selection.

The filtering method reduces redundancies within the data by selecting only relevant features. However, the definition of relevance is domain dependant, and it has been known that although irrelevant features are less useful for classifiers, not all relevant features are necessarily useful [6].

In this paper, more generic approaches were adopted for feature selection using two statistical methods: the t-test and Significant Analysis of Microarrays (SAM). The rationale for using statistical tests is that they are often used to validate the significance of different treatments to influence an outcome. Therefore, by performing statistical tests on the microarray features, it is possible to reduce redundancy by excluding features that are not statistically significant. The following sections briefly describe these two statistical methods.

2.1 Standard *t*-Test

The standard t-test is defined as:

$$t\text{-test} = \frac{\bar{x}_{i,1} - \bar{x}_{i,2}}{\hat{\sigma}_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (1)$$

where $\overline{x}_{i,1} - \overline{x}_{i,2}$ is the difference of the means between the two classes, n_1 and n_2 are the number of samples in the two classes, and $\hat{\sigma_i}$ is within-class standard deviation for gene i [9].

2.2 Significance Analysis of Microarrays (SAM)

The standard t-test was proposed for testing the significance of any data, while SAM was proposed specifically for testing the significance of genes in microarrays. Tusher *et al.* [20] argued that the t-test may discover many significant genes by chance, and subsequently proposed the development of SAM. SAM assigns each gene a score calculated on the basis of change in gene expression relative to standard deviation of repeated measurements.

The statistical test in SAM is given by $d(i)$, the "relative difference" of gene i and $s(i)$, the "gene-specific scatter" of gene i. Tusher *et al.* defined $d(i)$ as [20]:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \tag{2}$$

where $\bar{x}_I(i)$ is the average level of expression for gene (i) in states I, $\bar{x}_U$ in state U, s_0 is a data dependent constant, and

$$s(i) = \sqrt{a \left\{ \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2 \right\}} \tag{3}$$

where $\sum_m$ is the summation of the expression measurements in state I, $\sum_n$ in state U, $a = \frac{1/n_I + 1/n_U}{n_I + n_U - 2}$, and n_I is the number of measurements in state I, and n_U in state U. The genes are then ranked in order of the magnitude of $d(i)$ and those larger than a threshold value are considered significant.

3 Formulation of the Three Classifiers Studied

This section describes the formulation of SVM, MLP and LDA, the three supervised classifiers used in this paper.

3.1 Support Vector Machines (SVM)

Support vector machine introduced by Vapnik [22] has attracted much research attention in recent years due its demonstrated improved generalization performance over other techniques in many real world applications including the analysis of microarrays [5, 21]. It has been used in classification as well as regression tasks. The main difference between this technique and many other conventional classification techniques including neural networks is that it minimizes the structural risk instead of the empirical risk. The principle is based of the fact that minimizing an upper bound on the generalization error rather than minimizing the training error is expected to perform better. The generalization error rate is bounded by the sum of training error rate and a term that depends on Vapnik-Chervonenkis (VC) dimension [13]. VC dimension is a measure of complexity of the dimension space. Support vector machines find a balance between the empirical error and the VC-confidence interval. SVMs perform by nonlinearly mapping the input data into a high dimensional feature space by means of a kernel function and then do classification in the transformed space.

Consider a data set consisting $D = (\mathbf{x}_i, y_i)_{i=1}^{L}$ of L, with each input $\mathbf{x}_i \in \Re^n$ and the associated output $y_i \in$ {-1,+1}. Searching an optimal separating hyperplane (OSH) in the original input space is too restrictive in most practical cases. In SVM, each input $\mathbf{x}$ is first mapped into a higher dimension feature space $\mathcal{F}$ by $\mathbf{z} = \phi(\mathbf{x})$ via a nonlinear mapping $\phi : \Re^n \rightarrow \mathcal{F}$. Considering the case when the data are linearly separable in $\mathcal{F}$, there exists a vector $\mathbf{w} \in \mathcal{F}$ and a scalar b that define the separating hyperplane as: $\mathbf{w}.\mathbf{z} + b = 0$ such that

$$y_i(\mathbf{w}.\mathbf{z}_i + b) \geq 1, \forall i. \tag{4}$$

SVM constructs an OSH for which the margin of separation between the two classes is maximized. This margin is $2/||\mathbf{w}||$ according to its definition. Hence the unique hyperplane that optimally separates the data in $\mathcal{F}$ is the one that

$$min \quad \frac{1}{2}\mathbf{w}.\mathbf{w} \tag{5}$$

under the constraints of Eq. (4). When the data is linearly non-separable, the above minimization problem must be modified to allow classification error. This is done by generalizing the previous analysis with the introduction of some non-negative variables $\xi_i \geq 0$, often called *slack variables*, such that

$$y_i(\mathbf{w}.\mathbf{z}_i + b) \geq 1 - \xi_i, \forall i. \tag{6}$$

Only the misclassified data points x_i yield nonzero ξ_i. The term $\sum_{i=1}^{L} \xi_i$ can be regarded as a measure of misclassification. Thus the OSH is determined so that the maximization of the margin and minimization of training error is achieved by adding a penalty term to Eq. (5):

$$min \ \frac{1}{2}\ \mathbf{w}.\mathbf{w} + C\sum_{i=1}^{L} \xi_i \tag{7}$$

$$subject\ to \ \ y_i(\mathbf{w}.\mathbf{z}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \forall i$$

where C is a constant parameter, called regularization parameter, that determines the trade off between the maximum margin and minimum classification error. Minimizing the first term corresponds to minimizing the VC-dimension of the classifier and minimizing the second term controls the empirical risk.

Searching the optimal hyperplane in Eq. (7) is a Quadratic Programming (QP) problem that can be solved by constructing a Lagrangian and transforming in a dual. The optimal hyperplane can then be shown as the solution of

$$min \ \ W(\alpha) = \Sigma_{i=1}^{L}\alpha_i - \frac{1}{2}\Sigma_{i=1}^{L}\Sigma_{j=1}^{L}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

$$\textit{subject to} \;\; \Sigma_{i=1}^{L} y_i \alpha_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \forall i$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ is the non-negative Lagrangian multiplier. The data points $\mathbf{x}_i$ corresponding to $\alpha_i > 0$ lie along the margins of decision boundary and are support vectors (sv). The kernel function $K(.)$ describes an inner product in the D-dimensional space as described later and satisfies the Mercer's condition [8].

Having determined the optimum Lagrange multipliers, the optimum solution for the weight vector $\mathbf{w}$ is given by

$$\mathbf{w} = \Sigma_{i \in sv} \alpha_i y_i \mathbf{z}_i \tag{9}$$

where sv are the the support vectors. For any test vector $\mathbf{x} \in \Re^n$, the output is then given by

$$y = sign(\mathbf{w}.\mathbf{z} + b) = sign(\Sigma_{i \in sv} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x} + b)) \tag{10}$$

The generalization performance (i.e. classification accuracy in this study) depends on the parameters C and kernel type. In this study, we used the following kernel functions which are commonly used:

Linear $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i.\mathbf{x}_j$
Polynomial $K(\mathbf{x}_i, \mathbf{x}_j) = \{a(\mathbf{x}_i.\mathbf{x}_j) + b\}^d$
Radial basis (RBF) $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$
Sigmoid $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\mathbf{x}_i.\mathbf{x}_j + c)$

3.2 Multi-layer Perceptron (MLP)

A three layer MLP has an input layer, a hidden layer and an output layer. Successive layers are fully connected by weights. An input vector $\mathbf{x}$ is presented to the network and multiplied by the weights. All the weighted inputs to each unit of upper layer are then summed up and produce an output governed by $\mathbf{h}$ and $\mathbf{y}$, which are defined as

$$\mathbf{h} = f(\mathbf{W}_h.\mathbf{x} + \boldsymbol{\theta}_h) and \tag{11}$$

$$\mathbf{y} = f(\mathbf{W}_o.\mathbf{h} + \boldsymbol{\theta}_o) \tag{12}$$

where $\mathbf{y}$ is the output vector produced by the network, $\mathbf{W}_h$ and $\mathbf{W}_o$ are the hidden and output layer weight matrices, respectively, $\mathbf{h}$ is the vector denoting the response of hidden layer, $\boldsymbol{\theta}_h$ and $\boldsymbol{\theta}_o$ are the output and hidden layer bias vectors, respectively and $f(.)$ is the sigmoid activation function. The standard Backpropagation training algorithm [18] uses gradient descent techniques to minimize the sum of squared error measured at the output layer. In this study, we used an Levenberg and Marquardt technique to accelerate learning speed [16].

3.3 Linear Discriminant Analysis (LDA)

In LDA an n-dimensional data is projected onto a line according to a given direction $\mathbf{w}$. The choice of the projection direction is determined by different criteria. The Fischer's linear discriminant aims at maximizing the ratio of between-class scatter to within-class scatter [4]. Let $I_y = \{i : y_i = y\}, y \in \{-1, +1\}$ be the sets of indices of training vectors belonging to the first and second class, respectively. The class separability in a direction $\mathbf{w} \in \Re^n$ is found by maximizing the function $J(\mathbf{w})$, which is defined as

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_b \mathbf{w}}{\mathbf{w}^t \mathbf{S}_w \mathbf{w}} \tag{13}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-class and within-class scatter matrices, respectively.

4 Experimental Setup

We used the Lympochip (a DNA microarray developed at Stanford University School of Medicine [1]); 24 samples were collected from healthy cells and 72 samples collected from malignant lymphocytes cells, and each sample consists of 4026 different genes. We then used the t-test and SAM implemented in BRB-ArrayTools [19] to derive two subsets of Lymphochip, so in total there are now three sets of Lymphochip (see the Appendix for the BRB-ArayTools parameters). For convenience, we refer to the complete set of genes as Lymphochip, the subset selected by t-test as L_{t-test} and the subset by SAM as L_{SAM}. The total number of genes in Lymphochip, L_{t-test} and L_{SAM} are 4026, 387 and 418, respectively. Note that both subsets have only about 10% of the total number of genes in Lymphochip.

We trained SVM classifiers using four kernels (linear, RBF, polynomial and sigmoid) for all three data sets using libsvm [7]. As SVM is sensitive to training parameters such as the regularisation parameters (C) and the parameters for each kernel type, we generated 1856 SVM by varying the values of C from 2^{-31} to 2^{27} with an increment of (2^2). The parameters a in the polynomial kernel, $\gamma = (1/2\sigma^2)$ in RBF and c in sigmoid were all varied from 2^{-31} to 2^9 with an increment of 2^2. The parameter d in the polynomial kernel was varied from 2 to 11 with an increment of 1.

We also trained LDA and MLP using L_{t-test} and L_{SAM} and compared the best performing SVM kernel (on L_{t-test} and L_{SAM}) with LDA and MLP. The MLP was trained using 5 to 13 hidden nodes with an increment of two nodes and the training was stopped when the mean square error reached 0.01 or smaller. As MLP settles down to different set of weights depending on initial weights and learning parameters producing different results on each run, it is a common practice to generate multiple runs of MLP and use the average results for comparison. In this experiment, MLP was trained 20 times using different initial weights and learning parameters. LDA is the simplest classifier of the three and requires no parameters setting. Like SVM, it is also a non-stochastic process, so it is adequate to run it only once.

The 6-fold cross-validation technique was used in all experiments, and each fold contained 4 healthy and 12 malignant samples. The performance of all classifiers was then analysed using the average accuracy, sensitivity and specificity of the 6-fold data measured as

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \tag{14}$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \tag{15}$$

$$Specificity = \frac{TN}{TN+FP} \times 100\% \tag{16}$$

where TP is the number of true positives i.e. the number of malignant cells labelled as malignant; TN is the number of true negatives i.e. the number of healthy cells labelled as healthy; FP is the number of false positive i.e. the number of healthy cells labelled as malignant; and FN is the number of false negative i.e. the number of malignant cells labelled as healthy. Accuracy measures the overall detection for both healthy and cancerous cells. Sensitivity measures the ability of a classifier in recognizing malignant cells whereas specificity measures the ability of a classifier for not failing to detect healthy cells.

To further analyse how well the classifiers are able to generalize on unseen data, we employed Receiver Operating Characteristics (ROC) curves and calculated the area under the curves. ROC curve plots sensitivity against (1-specificity) as the threshold level of the classifier is varied. ROC analysis is commonly used in medicine and healthcare to qualify the accuracy of diagnostic test and evaluate performance of intelligent system [11].

5 Results and Discussion

This paper first discusses the performance of SVM using four kernels in Lymphochip, L_{t-test} and L_{SAM} (Section 5.1) and then compared the best performing SVM kernel against MLP and LDA using L_{t-test} and L_{SAM} (Section 5.2)

5.1 Performance of SVM in Lymphochip, L_{t-test} and L_{SAM}

The average sensitivity, specificity and accuracy for 6-fold cross validation of the four SVM kernels using the three data sets are presented in Table 1. We first compare the effectiveness of the different SVM kernels within each data sets, then across different data sets. For polynomial kernel the second degree ($d = 2$) produced the best results and is referred as Poly2 in the table.

It is clear that for Lymphochip, the accuracy of linear, RBF and sigmoid kernels are equally good and are also the best whereas the two-degree polynomial has lower specificity and accuracy. While for L_{t-test}, only the RBF and polynomial kernels are equally good, followed by the linear and sigmoid kernels. It is difficult to say whether the linear kernel is better than the sigmoid kernel because although its sensitivity is better than that of the sigmoid kernel,

Table 1. The average sensitivity, specificity and accuracy for 6-fold cross validation of four SVM kernels using the three datasets: Lymphochip, L_{t-test} and L_{SAM}

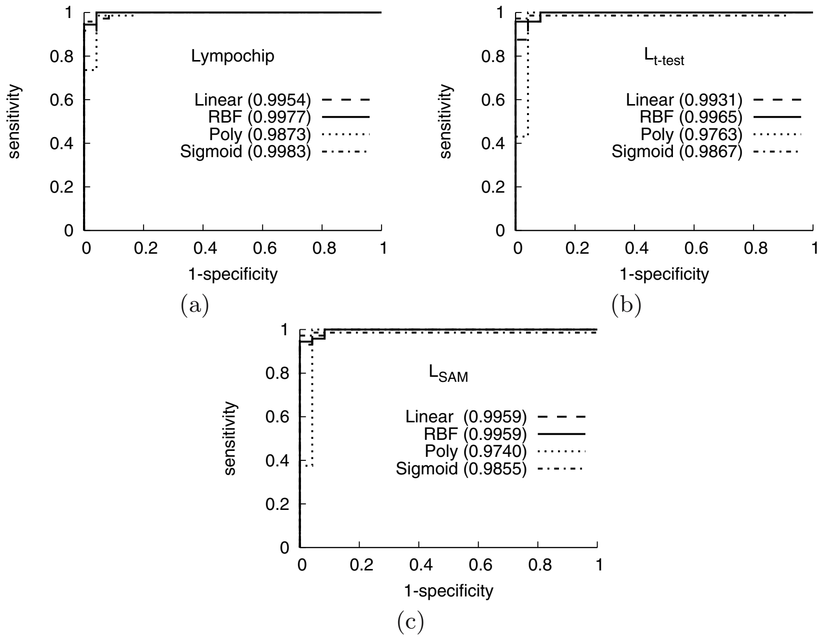
Kernel	Lympochip			L_{t-test}			L_{SAM}		
	Sens.	*Spec.*	*Acc.*	*Sens.*	*Spec.*	*Acc.*	*Sens.*	*Spec.*	*Acc.*
Linear	100.0	91.67	97.92	98.61	91.67	96.88	98.61	91.67	96.88
RBF	100.0	91.67	97.92	100.0	91.67	97.92	100.0	91.67	97.92
Poly2	100.0	79.17	94.79	100.0	91.67	97.92	100.0	87.5	96.88
Sigmoid	100.0	91.67	97.92	97.22	95.83	96.88	97.22	95.83	96.88

its specificity is worse than that of the sigmoid kernel. Interestingly, the specificity of Poly2 in L_{t-test} is better than in Lympochip. This means that having higher number of genes does not necessarily improves performance. For L_{SAM}, the RBF kernel performs the best, followed by the linear and sigmoid kernels and lastly, the Poly2 kernel. Again, it is difficult to rank the last three kernels because each performs well in different measures. As in L_{t-test}, the specificity of Poly2 is slightly better than in Lymphochip.

After comparing the performance of kernels within each data set, we now compare the accuracy of the best performing kernel across all data sets. The results suggest that the RBF kernel is the most suitable kernel for this microarray as it performs no worse than any of the other kernels across all data sets. The observation also suggests that it is possible to achieve the same level of accuracy using reduced subsets, L_{t-test} and L_{SAM}, especially by choosing appropriate kernel type. This is significant because both subsets have only about 10% of the number of genes in *Lymphochip* and they were derived without using any domain knowledge. The reduced number of genes reduces the training time of SVM by about 90% on average. More importantly, this added advantage comes without sacrificing the accuracy, enabling biologists to further explore the influence of the subset genes in malignancy development.

For further analysis, we studied the ROC curves and the area under the curves. The area under the ROC curve (AUC) summarizes the quality of the classification and is used as a single measure of accuracy [11]. A maximum attainable value of AUC is 1.0 and the higher value is more desirable. Figure 1 shows the ROC curves and AUC of all four kernels using the three data sets. It can be seen that apart from Poly2, the curves of the other three kernels remain reasonably close. It is interesting to note that Poly2, despite having a lower specificity in Lymphochip than in L_{t-test} and L_{SAM}, it actually has better AUC in Lymphochip than in L_{t-test} and L_{SAM}.

For the Lymphochip data set, it appears that the performance of all kernels except Poly2 is comparable. It is only at the subsets that we see more variation, most notably in the Poly2 kernel and, to a lesser degree, in the sigmoid kernel. This finding confirms the observation made earlier on the sensitivity, specificity and accuracy in that the performance of the kernels in Lympochip is more uniform than in the subsets. Notably, the AUC for RBF kernel in all the datasets are comparable. This result shows that RBF kernel is most suitable for



(a) (b) (c)

Fig. 1. ROC curves and areas under ROC curves (AUCs) of the four SVM kernels evaluated using (a) Lympochip and the subsets (b) L_{t-test} and (c) L_{SAM}

lymphoma microarray data, as it performs well across three data sets, and it is also consistent with the analysis based on sensitivity, specificity and accuracy measures. Given that the performance of L_{t-test} and L_{SAM} are comparable to using all genes, we can reasonably conclude that it is possible to use SVM-RBF most effectively in conjunction with gene selection techniques.

The next section compares the performance of SVM with other well known classifiers (MLP and LDA) using the L_{t-test} and L_{SAM} subsets.

5.2 Comparing SVM Against MLP and LDA in L_{t-test} and L_{SAM}

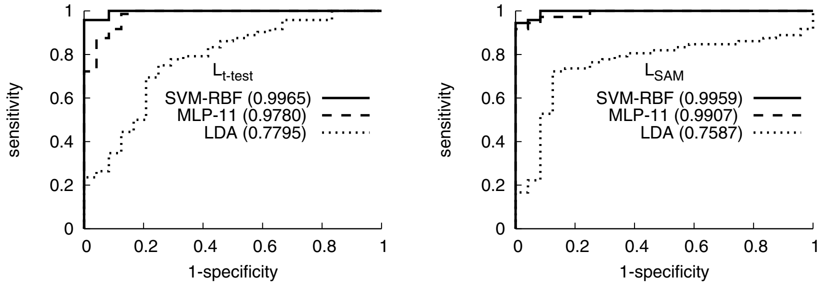
This section compares the best performing SVM kernel with MLP and LDA. To recap, the MLP were trained using 5 to 13 hidden nodes and for convenience, they are labelled from MLP-5 to MLP-13. Table 2 compares the sensitivity, specificity and accuracy of the SVM-RBF, MLP and LDA; the results for all MLP are the average from 20 runs. It can be seen from Table 2 that increasing the number of hidden nodes does not necessarily increase its effectiveness: in both L_{t-test} and L_{SAM}, MLP-11 has higher specificity and accuracy than MLP-13.

It is clear that SVM-RBF outperforms both MLP and LDA. As LDA is only suitable for linearly separable classification, the poor results seem to suggest that the classification on the reduced dataset is non-linearly separable. This would also explain why MLP performs better than LDA; MLP is known to perform well for non-linearly separable classifications.

Table 2. The sensitivity, specificity and accuracy of SVM-RBF (the best performing kernel), MLP with different hidden nodes and LDA

Classifiers	L_{t-test}			L_{SAM}		
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.
SVM-RBF	100.0	91.67	97.92	100.0	91.67	97.92
MLP-5*	95.49	87.71	93.54	92.85	91.67	92.55
MLP-7*	96.18	86.25	93.70	96.11	86.04	93.59
MLP-9*	96.11	90.42	94.69	95.21	90.42	94.01
MLP-11*	96.74	90.00	95.05	97.43	92.71	96.25
MLP-13*	96.88	88.96	94.90	97.22	88.54	95.05
LDA	76.39	70.83	75.0	72.22	83.33	75.0

*average of 20 runs



(a) (b)

Fig. 2. ROC curves and AUC of the SVM-RBF, MLP-11 and LDA using the subsets (a) L_{t-test} and (b) L_{SAM}

Figure 2 compares the ROC curves and AUC of SVM-RBF, MLP-11 and LDA in L_{t-test} and L_{SAM}. Note that the AUC for MLP is the average of 20 runs. In both datasets, SVM-RBF perform better than the other two classifiers. As the MLP was run 20 times, we only show the curve of one run: the run having the area closest to the average (for L_{t-test} the area was 0.9792, for L_{SAM} was 0.9896). For both datasets, SVM-RBF yields greater AUC followed by MLP-11 and lastly LDA.

6 Conclusions and Future Works

This paper compared the performance of SVM in four different kernels (linear, RBF, polynomial and sigmoid) using all 4026 genes in Lymphochip and two reduced subsets extracted by employing gene selection techniques, where each set has only about 10% number of genes from the Lymphochip. The performance was measured in terms of sensitivity, specificity, accuracy and ROC analysis. This paper showed that the performance of SVM using RBF, the most suitable kernel for lymphoma microarray data, on small subsets of genes are comparable to the results of using all genes. The advantage of using only small subsets is that it requires less training time for SVM without sacrificing accuracy. Importantly, these reduced subsets were obtained using generic approaches i.e. without using any domain knowledge. The reduced sets will help biologists to concrete on fewer genes to identify their roles in malignancy development.

This paper then compared SVM with MLP and LDA using the two subsets. Experimental results showed that SVM outperforms the other two classifiers. Future experiments will involve SVM in gene selection process by determining the relative influence of selected genes for further reduction of significant gene set and further classification of malignant cells into main types of lymphoma.

Appendix

The parameters used in generating L_{t-test} using BRB-ArrayTools were as follows. P-value was 0.001, multivariate permutation: maximum number of false discover was 5, maximum portion of false discover was 0.01, confidence level was 95% and the number of permutations was 2000.

The parameters used in generating L_{SAM} using BRB-ArrayTools were the following: median proportion of false discovery was 0.001 and the number of permutations was 500.

References

1. A A Alizadeh and M B Eisen et al. Distinct types of diffuse large B-cell lyumphoma identified by gene expression profiling. *Nature*, 403:503–511, Feb 2000.
2. U Alon, N Barkai, and D A Notterman et al. Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *PNAS*, volume 96, pages 6745–6750, Washington, DC, 1999. National Academy of Sciences.
3. A Ben-Dor, L Bruhn, N Friedman, I Nachman, M Schummer, and Z Yakhini. Tissue classification with gene expression profiles. In *4th Intl Conf on Comptnl Molecular Bio*, Tokyo, 2000. Universal Acad. Press.
4. C M Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.
5. M P S Brown, W N Grundy, D Lin, N Cristianini, C Sugnet, M Agnes Jr, and D Haussler. Support vector machine classification of microarray gene expression data. Technical report, U. California (Santa Cruz), 1999.
6. R A Caruana and D Freitag. How useful is relevance? Technical report, Fall'94 AAAI Symposium on Relevance, New Orleans, 1994.
7. C C Chang and C J Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
8. V Chercassky and P Mullier. *Learning from Data, Concepts, Theory and Methods.* John Wiley, 1998.
9. Jay L Devore. *Probability and Statistics for Engineering and the Sciences.* Brooks/Cole, 1987.
10. S Dudoid, J fridlyand, and T Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical report, University of California, Berkeley, 2000.

11. L. Lukas et al. Brain tumor classification based on long echo proton mrs signals. *Artificial Intelligence in Medicine*, 31:73–89, 2004.
12. T R Golub, D K Slonim, and P Tamayo et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, Oct 1998.
13. S Haykin. *Neural Network - A Comprehensive Foundation.* Prentice Hall, 1999.
14. J Khan, J S Wei, M Ringnér, L H Sall, M Ladanyi, and F Westermann. Classification and diagnostic prediction of cancers using gene expression profiling and aritifical neural networks. *Nat Med*, 7(6):673–679, 2001.
15. L C Molina, L Belanche, and A Nebot. Feature selection algorithms: A survey and experimental evaluation. In *ICDM'02*, 2002.
16. H.B. Demuth M.T. Hagan and M.H. Beale. *Neural Network Design.* PWS Publishing, Boston, MA, 1996.
17. J De Risi, V Iyer, and P Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–6, 1997.
18. D E Rumelhart and the PDP Research Group. *Parallel Distributed Processing.* MIT Press, New York, 1986.
19. Richard Simon and Amy Peng Lam. BRB ArrayTools v 3.2.
http://linus.nci.nih.gov/BRB-ArrayTools.html, 2004.
20. Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. In *Proc Natl Acad Sci*, volume 98, pages 5116–5121, 2001.
21. G Valentini. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. *Artificial Intelligence in Medicine*, 26:281–304, 2002.
22. V N Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.