Microarray Data Classification Using Support Vector Machine

Seeja.K.R. seeja@jamiahamdard.ac.in
Department of Computer Science
Jamia Hamdard University
New Delhi, India

Shweta rajput.laksh@yahoo.co.in
Department of Computer Science
Jamia Hamdard University
New Delhi, India

Abstract

DNA microarrays allow biologist to measure the expression of thousands of genes simultaneously on a small chip. These microarrays generate huge amount of data and new methods are needed to analyse them. In this paper, a new classification method based on support vector machine is proposed. The proposed method is used to classify gene expression data recorded on DNA microarrays. The proposed method is tested by using benchmark datasets and it is found that the proposed method is faster than neural network and the classification performance is not less than neural network.

Keywords: Support Vector Machines, Microarray, Classification

1. INTRODUCTION

When a normal tissue becomes cancerous, the expression levels of many genes change. By identifying these changes in gene expression, the tissues can be classified as cancerous and normal. Microarray technology is a hybridization technique which allows monitoring the expression of thousands of genes in a single experiment on a small chip. The output of these microarray experiments are the expression levels of different genes and these data are publicly available. These datasets include a large number of gene expression values and need to have a good data mining method to extract knowledge from these microarray gene expression datasets. Support vector machine (SVM) is a supervised computer learning technique used for data classification. It performs classification by constructing an optimal hyper plane which separates the data into two classes.

Many researchers have developed and demonstrated different classification techniques for cancer classification based on micro array gene expression data. Feature selection techniques [1],[2] have been suggested before classification, which finds the top features that discriminate various classes. Kernel based techniques [3],[4] like SVM have already been used for binary disease classification problems. Gene selection[5] and neural networks[6] based classifications were also reported in microarray data analysis.

In this paper SVM is used for the cancer classification based on microarray gene expression data. SVM is trained using different kernels like Poly Kernel, Normalized Poly Kernel and RBF and found that SVM performs better or equal classification than Neural Network.

2. MATERIALS AND METHODS

2.1 DNA Microarray

DNA microarrays can be used to measure changes in expression levels of genes in different biological conditions. The principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence mean tighter non-covalent bonding between the two strands. After washing off of non-specific bonding sequences, only

strongly paired strands will remain hybridized. So fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the strength of the hybridization. Microarrays use relative quantization in which the intensity of a feature is compared to the intensity of the same feature under a different condition.

2.2 Support Vector Machine

Support vector machine[7] is a powerful data mining technique for classifying data. The support vector machine is a training algorithm for learning classification and regression rules from data. SVM was developed form statistical learning theory and was first suggested by Vapnik[8] in the 1960 for data classification. SVM classifies data in large data sets by identifying a linear or non-linear separating surface in the input space of a data set. The separating surface depends only on a subset of the original data known as a set of support vectors. A support vector machine constructs a hyper plane or set of hyper planes in a high dimensional space, which can be used for classification. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class, called functional margin. If this functional margin is large, then the generalization error of the classifier will be small. SVM models are built around a kernel function [9],[10] that transforms the input data into an n-dimensional space where a hyper plane can be constructed to partition the data.

2.3 Dataset Used

In this paper, the acute leukemia bench mark dataset described by Golub et al [1] is used for classification and it is downloaded from Broad Institute's website[11]. The leukemia data set includes expression profiles of 7,129 human DNA probes spotted on Affymetrix Hu6800 microarrays of 72 patients with either acute myeloid leukemia (AML) or acute lymphocytic leukemia (ALL). Tissue samples were collected at time of diagnosis before treatment, taken either from bone marrow (62 cases), or peripheral blood (10 cases) and reflect both childhood and adult leukemia. The gene expression profiles of the original data set are represented as log10 normalized expression values. This data set was used as a benchmark for various machine learning techniques. The data set is divided into training set containing 38 samples and a validation set containing 34 samples.

2.4. Feature Selection

The proposed SVM based classification method, uses a feature selection algorithm to find the top features, which classifies the data sets effectively. The F(x) score[2] helps to find features that discriminate between the two classes. In this application genes are the features. The feature selection algorithm described below identifies the genes whose expression shows great change in both the classes.

1. Obtain the mean of the expression values for each gene of ALL samples and mean of the expression values for each gene of AML samples.
2. Obtain absolute difference between the mean of ALL samples and the mean of AML samples.
3. Arrange the genes based on absolute difference in decreasing order.
4. Select Top 250 genes.
5. Apply the following formula on selected 250 genes.

   $$F(x_i) = (\mu(ALL) - \mu(AML)) / (\square(ALL) + \square(AML))$$

   where μ is the mean and $\square$ is the standard deviation.
6. Select 200 genes with highest absolute $F(x_i)$ scores as our top features.

2.5 SMO (Sequential Minimal Optimization) Algorithm

The learning task in SVM can be formulated as a convex optimization problem, which can be solved by using Lagrange Multiplier method. Sequential Minimal Optimization (SMO) [12] is a simple algorithm that can quickly solve the SVM QP problem without any extra matrix storage and without using numerical QP optimization. The advantage of SMO is its ability to solve the Lagrange multipliers analytically.

4. RESULTS AND DISCUSSION

We have used the WEKA version 3.6.4[13] software for performing the classification. WEKA contains an implementation of SMO algorithm which supports SVM. Feature selection algorithm is implemented in C#.

First SVM is trained by using the bench mark training set. After training, the classification accuracy is validated using the training set as well as testing set. The training dataset contains 38 training samples and all the samples were classified without error using poly kernel, Normalized poly kernel and RBF kernel during training as shown in Figure 1. On 10 fold cross validation of training dataset all the 38 samples were classified without error using poly kernel, Normalized poly kernel and one AML sample was misclassified using RBF kernel as shown in Figure 2. Then we applied 34 test data samples to the trained SVM, 2 AML samples were misclassified using RBF kernel and 3 AML were misclassified using Poly kernel and Normalized poly kernel. All other samples were classified correctly. Figure 3 shows this result.

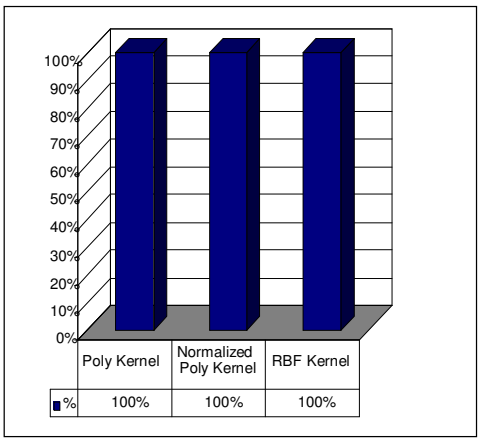

FIGURE 1: Classification accuracy on training set for different kernels

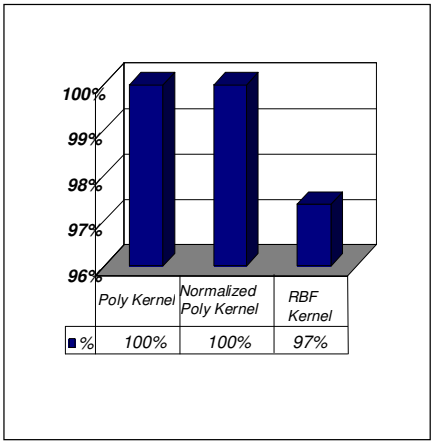

FIGURE 2: Classification accuracy on 10-fold cross validation of training set for different kernels

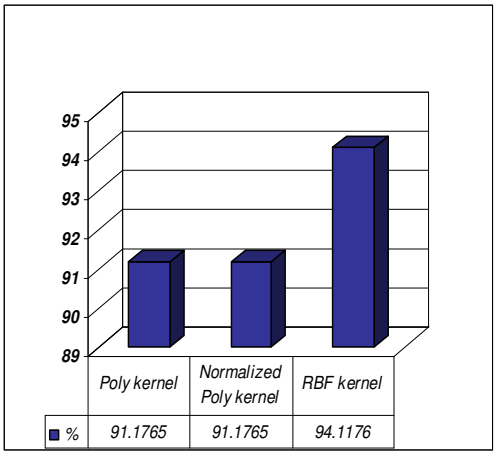

FIGURE 3: Classification accuracy on test dataset for different kernels

In order to evaluate the performance of SVM, we have applied the same dataset to the neural network learning algorithm available in WEKA. We found that both SVM and neural network classifies the data with same accuracy. But SVM is taking less time than neural network. Figure 4 and figure 5 show the comparison on time.

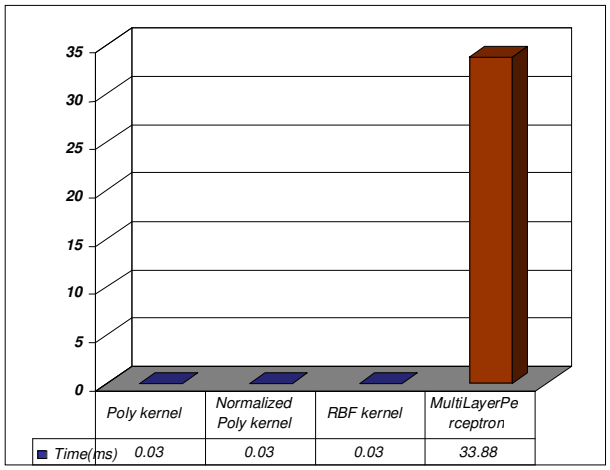

FIGURE 4: SVM Vs NEURAL NETWORK(Training)

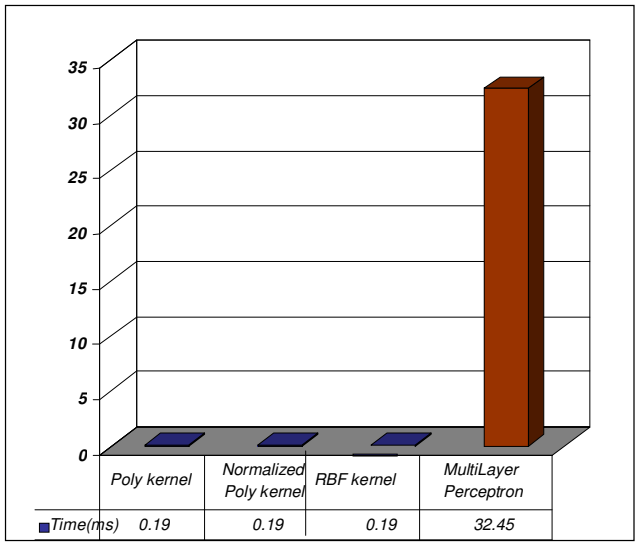

FIGURE 5: SVM Vs NEURAL NETWORK(Testing)

5. CONCLUSION

We have proposed an efficient and powerful method for microarray gene expression data classification and prediction using support vector machine. We applied SVM on ALL/AML dataset. In order to evaluate the performance of SVM, we have applied the same dataset to the neural network learning algorithm available in WEKA. We found that both SVM and neural network classifies the data with same accuracy. But SVM is taking less learning time than neural network.

6. REFERENCES

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, 286(15):531–537, 1999.

2. Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer, and David Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data ", Bioinformatics6(10): 906-914 , 2000

3. Zhang, X. and Ke, H.," ALL/AML cancer classification by gene expression data using SVM and CSVM approach", Genome Informatics, Universal Academy Press, pp. 237-239, 2000

4. Xin Zhao, Leo Wang-Kit Cheung, "Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data", BMC Bioinformatics.,8:67,2007.

5. Wenlong Xu, Minghui Wang, Xianghua Zhang, Lirong Wang, Huanqing Feng," SDED: A novel filter method for cancer-related gene selection", Bioinformation 2(7): 301-303,2008.

6. D.P. Berrar, C.S. Downes, W. Dubitzky, "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks", Pacific Symposium on Biocomputing 8:5-16, 2003.

7. Pang-Ning Tan, Michal Steinbach, Vipin Kumar, "Introduction to Data Mining.",Pearson Education Inc., pp. 256-276, 2009
8. Vapnik V , The nature of statistical learning theory. 2nd edition. Springer,1999

9. Joachims, T., “Making large-scale SVM learning practical”, Advances in Kernel Methods – Support Vector Learning, B. Schokopf et al. (ed.), MIT Press, 1999.

10. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G, “Support Vector Machines and Kernels for Computational Biology.”, PLoS Comput Biol 4(10), 2008.

11. ALL/AML Bench Mark Dataset:
12. www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

13. Platt, J. C.,” Fast training of support vector machines using sequential minimal optimization”. Advances in kernel methods: Support vector machines, B. Schokopf et al. (ed.), MIT Press, 1999.

14. WEKA software: www.cs.waikato.ac.nz/~ml/WEKA