A Review on Micro Array Multiclass Classification Using SVM

Swetapadma Rath

Abstract— The fundamental power of microarrays lies in the ability to conduct parallel surveys of gene expression patterns for tens of thousands of genes across a wide range of cellular responses, phenotypes and conditions. Thus microarray data contain an overwhelming number of genes relative to the number of samples, presenting challenges for meaningful pattern discovery. This paper provides a comparative study of gene selection methods for multi-class classification of microarray data. We compare several feature ranking techniques, including new variants of correlation coefficients, and Support Vector Machine (SVM) method based on Recursive Feature Elimination (RFE). The results show that feature selection methods improve SVM classification accuracy in different kernel settings. The performance of feature selection techniques is problem-dependent. SVM-RFE shows an excellent performance in general, but often gives lower accuracy than correlation coefficients in low dimensions.

Keywords- Class, Gene expression, Kernel, Micro array, phonetypes , SVM.

1.INTRODUCTION

Bioinformatics is an emerging field that has its roots in molecular biology, mathematics and computer science. It deals with generation, management and analysis of the biological data, which is obtained from various experiments and techniques, often resulting in large data[1]. The analysis of such enormous biological data requires use of sophisticated algorithms, which can process the data and help in visualizing the data and extract information from it . This led to the evolution of Bioinformatics, an interdisciplinary field involving both biologist and computer scientists. Advancement in the field of bioinformatics has facilitated many researchers in analyzing the data and understanding the structural, comparative and functional properties. Some of the enhancements being analysis of genomes and proteins, identifying metabolic and signaling pathways which define the gene to gene relationships, development of microarray chip and conducting microarray experiments to measure the gene expression levels. The availability of the data on public websites and repositories made it easier to carry out the research. NCBI is one such database that includes DNA and protein sequence data and also facilitates researchers to contribute their sequences to the database. KEGG and EcoCyc are the databases that consists the 2 information about the pathways .To process the data, finely tuned algorithms were developed over the years and have been made publicly available. Some of them are BLAST, CLUSTALW algorithms that perform sequence comparison. Algorithms to perform phylogenetic analysis were also made available on the public websites. One of the major advancement made in the field of bioinformatics is the emergence of microarray technology. Microarray technology facilitates in determining the expression values of several genes simultaneously. The gene expression data is used for various analyses to understand the biological significance of the species or the tissue from which the genes were extracted for the experiment. One such analysis is classification of the sample based on the gene expression values that are obtained from the microarray experiment. This study focuses on analysis and calculation of distance measure and margin of a support vector machine classifier for microarray dataset. It also deals with studying the effect of margin value on the classification accuracy and relation between them. Before we proceed further, a brief introduction to gene expression and microarray technology is provided followed by a discussion of the support vector machine classifier.

The paper is divided into three sections. In section-1 a brief gene expression and microarray is given. Support Vector Machine (SVM) is presented in section-2 and conclusion is given in section-3.

1.1 Gene Expression and Microarray Technology

The characteristic features and behavior of a biological species largely depends on the genes and the proteins present in it. Proteins obtained from the genes vary depending upon the gene expression levels. Hence analyzing the expression levels of genes under various conditions will help us in identifying the reason behind abnormalities in diseased 3 species in addition to identifying the genes responsible for the abnormality. Microarray technology is used to study and record the gene expressions of thousands of genes simultaneously. Microarray is a chip on which biological substrates are bound to the probes present on the silicon chip or a glass slide. The biological substrates can be DNAs, proteins molecules or carbohydrates that decide the type of microarray chip. There are different types of microarrays such as DNA microarrays, protein microarrays, tissue microarrays and carbohydrate microarrays [9]. DNA microarrays are the commonly used ones to record the expression levels of genes.

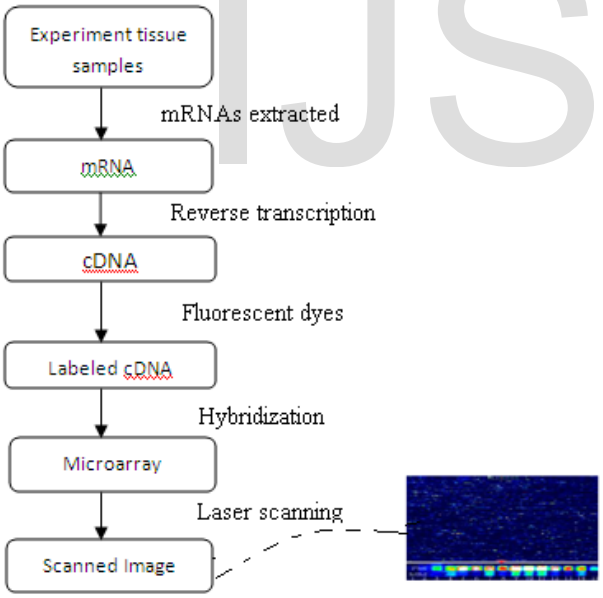

Figure 1.1 a typical microarray experiment.

The target mRNAs (messenger RNAs) of the species whose gene expressions are to be measured are reverse transcribed to cDNAs (complimentary DNAs). The cDNAs are labeled with fluorescent dyes or radioactive lasers and are hybridized on the microarray chip. The chip is left overnight to let it hybridize. During the process of hybridization cDNAs bind to their complementary strands present on the microarray chip using base pair bonding. Then the chip is washed to remove any non-specific DNA bindings present on it. It is then scanned to obtain a digital image. The image obtained is analyzed and processed using image processing and data normalization techniques to record the expression levels of thousands of genes. For a dual channel microarray chip, both control cell tissue samples and experiment cell tissue samples are extracted and are colored with different fluorescent dyes. Then they are reverse-transcribed to cDNAs and are hybridized on the dual channel microarray chip. After hybridization the chip is scanned to obtain an image which is further processed to obtain the gene expression levels of the experiment tissue samples. Microarrays have many applications in medical and biological fields. Various kinds of microarrays are used to obtain the expression levels of the biological entities. For example, the protein microarrays are used to understand the protein-protein, protein-drug and protein-DNA interactions. In medicine, DNA microarrays are used to identify the differentially expressed genes. In addition microarrays are used for drug discovery and to study the changes in the gene expression levels in response to the drugs. In cancer research, microarrays are used for determining mutation detection, gene copy number analysis, and cancer therapeutics and drug sensitivity.

1.2 Microarray Structure And Analysis Model

To standardize the analysis of microarrays, a commonly accepted form of the microarray data structure has evolved. The data structure is an (M×N) 2-D matrix of gene expressions of (M) genes for (N) samples. In some literature it is defined as the transpose of this definition, i.e. (M×N).

This data structure is usually referred to as (X):

$$X(t) = (x1(t) \ \ldots. \ xm(t)), \ t = 1,2 \ldots\ldots. N \quad \text{(Eq. 1)}$$

(Eq.1) shows the mathematical definition of the microarray. The expression $x_i(t)$ denotes the value of the gene (i) for the sample (t). In most of the times this set of data is associated with groups' labels vector y (t) which maps each sample's gene expression vector to a group label. Usually the labels are discrete numeric values that represent different groups. For example if some of the samples belong to cancer tumors y(t) and the others to normal tumors then might be either 1 or 0 denoting a

cancer sample or a normal sample respectively. (Eq.2) shows the mathematical mapping x (t) of y (t).

$$X(t) = [\, x(1)\ \ x(2)\ \ \ldots\ \ x(N)\,] \,,\ Y(t) = [\, y(1)\ \ y(2)\ \ \ldots\ y(N)\,] \quad \text{(Eq.2)}$$

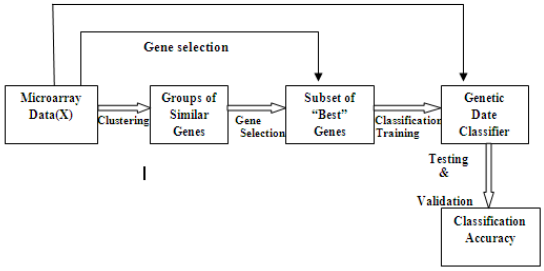

Figure 1.2: Analysis Model

This array results after what is known as preprocessing which turns the initial row data taken from the experiments into the standard form of sample-gene matrix (X) [3, 4]. There are many forms of analysis. It can be used directly to train a classifier using the entire gene set. Although this is possible, it has many disadvantages. It demands a lot of processing resources and the seriousness of this problem depends on the nature of the classifier to be used. In addition to this, using the entire set of genes misses the biological empirical observation which says that for a particular target issue most of the genes are irrelevant [1, 11]. The most common way in building classifiers using microarray data is to start with gene selection to select a subset of genes which is expected to contain the most relevant genes to the particular given phenotypic issue. Gene selection is usually applied over the sample-gene matrix directly and it tackles two main issues; discriminative genes and redundant genes. Discriminative genes are those genes whose profiles have strong statistical differences between different classes, so they are good genes to be used to differentiate between samples that belong to different classes. Redundant genes are those genes which have close profiles. Even if these genes are strongly discriminative, including all of them adds no value since one of them offers almost the same amount of information as all of them [1, 5, 6]. In some cases, to select the best set of genes, it can be useful to start by grouping genes in number of groups (clusters) that include genes of close profiles. This process is called clustering and it is applied using many different clustering methods. After clustering the genes into clusters, genes are selected for classification purposes by taking number of genes from each cluster in a way that these genes cover different spaces of the classification problem, i.e. are from different clusters [8-10]. After training the classifier with the samples using the selected subset of genes the classifier needs to be tested and assigned a numeric performance value. The most common metric to measure classifiers' performance is the classification accuracy, which is the percentage of the correctly classified test samples of the entire test sample set. Many methods have been introduced in the literature to perform testing and validation for the classifiers [2, 11, 12].

2. SUPPORT VECTOR MACHINE

Support vector machine (SVM) is gaining popularity for its ability to classify noisy and high dimensional data. SVM is a statistical learning algorithm that classifies the samples using a subset of training samples called support vectors. The idea behind SVM classifier is that it creates a feature space using the attributes in the training data. It then tries to identify a decision boundary or a hyper-plane that separates the feature space into two halves where each half contains only the training data points belonging to a category. This is shown in Figure 1.2

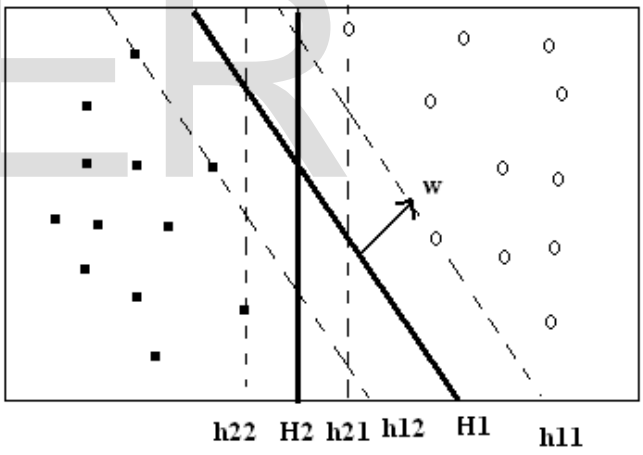

Figure 2 Example of SVM

In Figure the circular data points belong to one class and square points belong to another class. SVM tries to find a hyper-plane (H1 or H2) that separates the two six categories. There are two types of SVMs, (1) Linear SVM, which separates the data points using a linear decision boundary and (2) Non-linear SVM, which separates the data Points using a non-linear decision boundary. For a linear SVM the equation for the decision boundary is

$$w.\, x + b = 0 \qquad (2.1)$$

where, w and x are vector s and the direction of w is perpendicular to the linear decision boundary. Vector w is determined using the training dataset. For any set of data points (xi) that lie above the decision boundary the equation is

$$w \cdot xi + b = k, \quad \text{where } k > 0 \qquad (2.2)$$

and for the data points (xj) which lie below the decision boundary the equation is

$$w \cdot xi + b = k', \quad \text{where } k'<0 \qquad (2.3)$$

By rescaling the values of w and b the equations of the two supporting hyper planes (h11and h12) can be defined as

$$h11 : w.x + b = 1 \qquad (2.4)$$

$$h12: w.x + b = -1 \qquad (2.5)$$

The distance between the two hyper planes (margin "d") is obtained by

$$w \cdot (x1 - x2) = 2 \qquad (2.6)$$

$$d = 2/||w|| \qquad (2.7)$$

The objective of SVM classifier is to maximize the value of d. This objective is equivalent to minimizing the value of ||w||2/2. The values of w and b are obtained by solving this quadratic optimization problem under the constraints

$$w \cdot xi + b > 1 \text{ if } yi = 1 \qquad (2.8)$$

$$w \cdot xi + b < -1 \text{ if } yi = -1 \qquad (2.9)$$

Where yi is the class variable for xi. Imposing these restrictions will make SVM to place the training instances with yi = 1 above the hyper plane h11 and the training instances with yi = -1 below the hyper plane h12. The optimization problem can be solved using Lagrange multiplier method. The objective function to be minimized in the Lagrangian form can be written as:

$$Lp = \frac{1}{2}||w||x^2 - \sum_{i=1}^{N} \propto i(yi\,(w.xi + b) - 1) \qquad (2.10)$$

αi are Lagrange multipliers and N are the number of samples . The Lagrange multipliers should be non-negative (αi > 0). In order to minimize the Lagrangian form, its partial derivatives are obtained with respect to w and b and are equated to zero.

$$\frac{\partial Lp}{\partial W} = 0 \Rightarrow w = \sum_{i=1}^{N} \propto_i y_i x_i \qquad (2.11)$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \propto_i y_i = 0 \qquad (2.12)$$

The dual form of (2.10), by using Lagrangian equation is:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \alpha_i\, \alpha_j y_i y_j \mathbf{x_i x_j} \qquad (2.13)$$

The training instances for which the value of αi > 0 lie on the hyper plane h11or h12 are called support vectors. Only these training instances are used to obtain the decision boundary parameters w and b. Hence the classification of unknown samples is based on the support vectors. In some cases it is preferable to misclassify some of training samples (training errors) in order to obtain decision boundary plane with maximum margin. A decision boundary with no training errors but smaller margin may lead to over-fitting and cannot classify unknown samples correctly. On the other hand, a decision boundary with few training errors and a larger margin can classify the unknown samples more accurately. Hence there must be a tradeoff between the margin and the number of training errors. The decision boundary thus obtained is called as soft margin. The constraints for the optimization problem still hold good but need the addition of slack variables, which account for the soft margin. These slack variables correspond to the error in decision boundary. Also a penalty for the training error should be introduced in the objective function in order to balance the margin value and the number of training errors. The objective function for the optimization problem will be minimization of

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \alpha_i\, \alpha_j y_i y_j \phi(\mathbf{x_i})\phi(\mathbf{x_j})$$

$$\|\mathbf{w}\|^2/2 + C\,(\sum \xi_i)^k \qquad (2.14)$$

Where C and k are specified by the user and can be varied depending on the dataset. The constraints for the optimization problem will be

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq 1 - \xi_i, \quad \text{if } y_i = 1, \qquad (2.15)$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1 + \xi_i, \quad \text{if } y_i = -1. \qquad (2.16)$$

The Lagrange multiplier for soft margin differs from the Lagrange multipliers of linear decision boundary. αi values should be non-negative and also should be less than or equal to C. Hence the parameter C acts as the upper limit for error in the decision boundary [6]. Linear SVM performs well on datasets that can be easily separated by a hyper-plane into two parts. But sometimes datasets are complex and are difficult to classify using a linear kernel. Non-linear SVM classifiers can be used for such complex datasets. The concept behind non-linear SVM classifier is to transform the dataset into a high dimensional space where the data can be separated using a linear decision boundary. In the original feature space the decision boundary is not linear. The main problem with transforming the dataset to higher dimension is the increase in complexity of the classifier. Also the exact mapping function that can separate data linearly in higher dimensional space is not known. In order to overcome this, a concept called kernel trick is used to transform the data to higher dimensional space. If Φ is the mapping function, in order to find the linear decision boundary in the transformed higher dimensional space, attribute x in the Equation 1.13 is replaced with Φ(x). The transformed Lagrangian dual form is given by

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x_i})\phi(\mathbf{x_j}) \qquad (2.17)$$

The dot product is a measure of similarity between two vectors. The key idea behind the kernel trick is that it considers the dot product analogous in the original and the transformed space. Consider two input instance vectors xi and xj in the original space. it aids in computing the kernel function in the transformed space using the original attribute set. Hence the original nonlinear decision boundary equation in lower dimension space is transformed to an equation of linear decision boundary in higher dimensional space given by:

$$\mathbf{w} \cdot \phi(\mathbf{x}) + b = 0 \qquad (2.18)$$

2.1 Application of SVM:-

DATASET INFORMATION

For simulation work data sets are downloaded from the 'UCI Machine Learning Repository website. It maintains 187 data sets as a service to the machine learning community. The following data sets are chosen for Simulation work .These data sets don't contain any missing data. Description of each data set is given below:-

Dataset	No. of record	No. of attribute	No. of classes
Irish	150	4	3
Lung cancer	197	581	4

Table 1: Iris data set

Training percentage	Testing percentage	Accuracy achieved	Confusion matrix
50	50	77.33	24 1 0 0 21 4 0 12 13
60	40	80	19 1 0 0 19 1 0 10 10
70	30	75.55	14 1 0 0 13 2 0 8 7
80	20	80	9 1 0 0 10 0 0 5 5

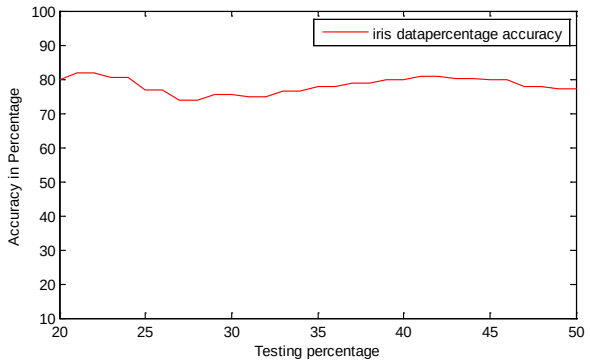

Figure 2.1 Accuracy plot for Iris data

Table 2 Lung cancer data set

Training percentage	***Testing percentage***	***Accuracy achieved***	***Confusion matrix***
50	50	86	65 1 2 2 1 8 0

			0 6 0 5 0 2 0 0 8
60	40	90	53 1 1 1 1 6 0 0 3 0 6 0 1 0 0 7
70	30	85.2459	37 1 1 3 1 5 0 0 1 0 5 1 1 0 0 5
80	20	80.4878	24 1 0 3 1 3 0 0 1 0 3 1 1 0 0 3
90	10	90.4762	13 0 0 1 0 2 0 0 1 0 2 0 0 0 0 2

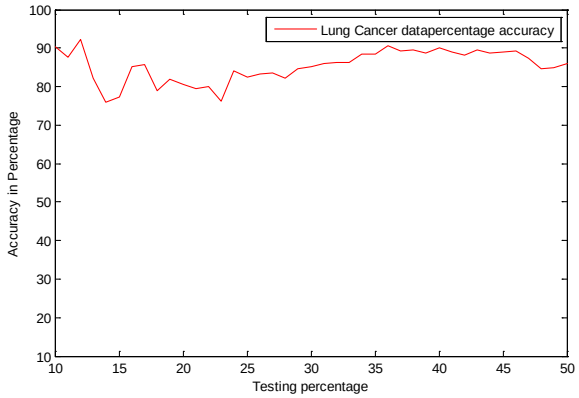

Figure 2.2 Accuracy plot for Lung Cancer data

3. CONCLUSION

Support Vector Machines are an attractive approach to data modeling. They combine generalization control with a technique to address the curse of dimensionality. The formulation results in a global quadratic optimization problem with box constraints, which is readily solved by interior point methods. The kernel mapping provides a unifying framework for most of the commonly employed model architectures, enabling comparisons to be performed. In classification problems generalization control is obtained by maximizing the margin, which corresponds to minimization of the weight vector in a canonical framework. The solution is obtained as a set of support vectors that can be sparse. These lie on the boundary and as such summarize the information required to separate the data. Microarray gene expression data are a very useful format of biological information.SVM new and very promising classification approach. The simulation result of two models is

presented for verification of the theory. A lot of research still to be done on Biological information processing using techniques developed in fields such as Machine Learning, Data Mining, etc.

4. REFERENCES

[1] Jawdat, D.; "The Era of Bioinformatics"; Information and Communication Technologies, 2006. ICTTA '06; 2nd , vol.1, no., pp.1860-1865.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H.Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfeld, E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, 286 (5439), pp 531-537, 1999.

[3] Sung-Huai Hsieh, Zhenyu Wang, Po-Hsun Cheng, I-Shun Lee, Sheau-Ling Hsieh, Feipei Lai; "Leukemia Cancer Classification based on Support Vector Machine", 8th IEEE International Conference on Industrial Informatics; pp 819-824, 2010.

[4] J. Phan, R. Moffitt, J. Dale, J. Petros, A. Young, M. Wang; "Improvement of SVM Algorithm for Microarray Analysis Using Intelligent Parameter Selection"; Engineering in Medicine and Biology Society, 2005, IEEE-EMBS 2005; 27th Annual International Conference of the IEEE; vol., no., pp.4838-4841, 2005.

[5] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer, David Haussler; "Support vector machine classification and validation of cancer tissue samples using microarray expression data"; Bioinformatics, Oxford Journals, Vol 16 no. 10, pp 906-914; 2000.

[6] Pang-Ning Tan, Michael Steinbach, Vipin Kumar; "Introduction to Data Mining"; Pearson Education, 2006; ISBN 0-321-42052-7; pp 256-276.

[7] Jawdat, D.; "The Era of Bioinformatics"; Information and Communication Technologies,. ICTTA '06; 2nd, vol.1, no., pp.1860-1865; 2006.

[8] Jacques Cohen; "Bioinformatics-an introduction for computer scientists"; ACM Computing Surveys; Vol 36, Issue 2, pp 122-158; 2004.

[9] Supratim Choudhary; "Microarrays in Biology and Medicine"; Journal of Biochemical and Molecular Toxicology; Volume 18, Issue 4; pp 171-179, 2004.

[10] D. A. Rew; "DNA microarray technology in cancer research"; European Journal of Surgical Oncology; Volume 27, Issue 5, pp 504-508; 2001.

[11] G. Piatetsky-Shapiro, P. Tamayo; "Microarray Data Mining: Facing the Challenges"; ACM SIGKDD Explorations Newsletter; ACM, NY, USA; Volume 5, Issue 2, pp 1- 5, 2003.

[12] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, T. Poggio; "Support Vector Machine Classification of Microarray Data"; AI Memo 1677, CBCL Paper No 182, MIT; 1998.