Feature Selection for Cancer Classification on Microarray Expression Data

Hui-Huang Hsu and Ming-Da Lu
Dept. of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
E-Mail: h_hsu@mail.tku.edu.tw, 490191771@s90.tku.edu.tw

Abstract

Microarray is an important tool in gene analysis research. It can help identify genes that might cause various cancers. In this paper, we use feature selection methods and the support vector machine (SVM) to search for the disease-causing genes in microarray data of three different cancers. The feature selection methods are based on Euclidian distance (ED) and Pearson correlation coefficient (PCC). We investigated the effect on prediction results by training the SVM with different numbers of features and different kinds of kernels. The results show that linear kernel is the fittest kernel for this problem. Also, equal or higher accuracy can be achieved with only 15 to 100 features which are selected from 7129 or more features of the original data sets.

Keywords: Cancer Classification, Microarray, Support Vector Machine, Feature Selection, Pearson Correlation Coefficient

1. Introduction

Microarray experiment is an important tool which can help biologists to understand gene expression. It is a high-throughput method to show gene expression data. In general, the data sets have numerous features and it is hard to analyze the data sets efficiently. Therefore, bioinformatics plays a very important role in this problem, especially the data mining technique [1][2]. In the paper of G. Piatetsky-Shapiro and P. Tamayo, it is shown that there are three main issues in microarray data mining tasks [3].

The three main issues are listed as follows:

- Classification: Classifying diseases and predicting results by microarray expression data.
- Gene selection: In data mining, it can be called feature selection. Selecting the most expressed genes in microarray data can help predict faster and more accurate.
- Clustering: Finding new biological classes or refining existing ones.

Microarray data classification and gene selection tasks are closely-related problems in microarray experiment data analysis. In this paper, we will discuss these two problems and find out a method that can help to classify microarray expression data correctly and reliably. We used two feature selection methods, namely, the distance-based method and the correlation-based method, to filter out features for the classification task. Also, various kernels with the SVM are tested for this problem. From the experimental results, we conclude that distance feature selection method is a better one for the SVM. And the linear kernel performed best with the microarray expression data. Also, with feature selection, the number of features used for classification can be dramatically reduced.

The rest of this paper is organized as follows. Section 2 describes a brief background and related research. Section 3 explains our feature selection methods and the SVM. Section 4 presents the results of our experiments. Section 5 draws the final conclusion.

2. Background and related work

In the past few years, numerous information technologies have been developed to help solve biology tasks. Microarray data analysis is one of the interesting problems in bioinformatics. Among them, one of the classification tasks is to analyze gene expression data to judge which disease is with the patient. In order to obtain higher accuracy and analyze faster, feature selection is needed.

Feature selection is a method to filter junk information from other useful features. It is a data mining technique. The major task of feature selection is finding out the best subset which has the minimum features. In general, the best subset means that one gets the highest or higher accuracy compare with original set [4][5][6]. In M. Dash and H. Liu's paper, they gave a different definition. The definition is in the following:

Definition: Feature selection attempts to select the

978-0-7695-3382-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ISDA.2008.198

minimally sized subset of features according to the classification accuracy. The condition is that the accuracy does not significantly decrease and the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features [7].
It shows that feature selection should focus on finding minimal subsets.

There are three different approaches of feature selection:

- Measure method: dependence, consistency, information, distance, classifier error rate.
- Generation method: forward, backward, weight, compound, random.
- Search method: heuristic search, complete search, random search.

These three are the characterization of feature selection, and we can catalog the feature selection methods by them.

The support vector machine is a popular supervised learning method for classification and regression in recent years. It was proposed by Vapnik in 1995 [8]. In 2002, Vapnik used the SVM to investigate gene selection problem and it was found that 16 to 64 genes can get the best accuracy in acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) cancer classification problems. In 2002, S. Cho and J. Ryu compared seven classification and seven feature selection methods in AML and ALL data sets. They selected 30 genes from 7129 genes and the accuracy was 68.5~94.1% [9]. In 2003, J. Zhang, R. Lee, and Y. J. Wang investigated in microarray expression data set without feature selection. They listed nine advantages and limitations of the SVM on this problem [10]. In 2007, W. Fujibuchi and T. Kato discussed three classifiers and six kernels in AML and ALL problem. Their method can reach 97.8% accuracy with a complete feature set. After feature selection, their maximum accuracy is around 87.5% [11]. In 2007, S. Cho and H. Won used another classifier to predict the same problem, and they found that the same feature numbers - around 25 to 30, as the paper they proposed earlier [12], can get the best accuracy 97.1%, too [13].

The above-mentioned studies show some success for microarray expression data classification. However, further improvements are still in need. Here, we examine different kernels for the SVM, different feature selection measure methods, and different microarray expression data sets in this problem.

3. Feature Selection

3.1 Experiment design

We got the microarray expression data sets from Kent Ridge Bio-medical Data Set Repository and we last accessed it on Dec. 27 2007 [14]. It is a public cancer microarray data set repository and all of the original experiment data sets are from Broad Institute Cancer Program Data Sets [15]. The data sets were used by [9],[10],[11],[12], and [13], too .

We process the data sets from the microarray expression data into Libsvm data input format. Libsvm is a popular open source SVM prediction tool. We will describe it in more detail in Subsection 3.3 [16].

Table 1. Experimental procedures

Normalization Feature Selection	No	Yes
No	1	2
Yes	3	4 F N 5 N F

After preprocessing, there are two decisions to be made. The two decisions are feature selection and normalization as shown in Table 1. Five different procedures can be tried. Among them, FN means feature selection is performed before normalization; NF means normalization is performed before feature selection. In normalization, we scale all the features into the range of -1 to 1.The method is defined by the following equation.

$$N_{ij} = \frac{f_{ij} - \min(f_{ij})}{\max(f_{ij}) - \min(f_{ij})} \qquad (1)$$

In equation 1, N_{ij} is the output normalized value, f_{ij} is the feature value of sample j in feature i. If there are k samples, we need to find out the minimum and maximum feature values of feature i from $j=0$ to $j=k-1$.

For feature selection, we use Euclidian distance and Pearson correlation coefficient methods to measure the features. We will discuss this in detail in Subsection 3.2.

After feature selection we tested the data sets by Libsvm with 5-fold cross-validation. Cross-validation is a statistical method. One fifth of the samples are selected randomly to be a subset. The subset is called the validating set, and the remaining samples are the training set which is used to generate the testing model. The process is repeated for five times. And the accuracy is the average of the five experimental results. It is used to avoid biases on training and testing data selection.

3.2 Feature selection

Generally speaking, feature selection algorithms include two parts. The first part is feature subset generation and the other is evaluation part. Fig. 1 is the flowchart of feature selection.

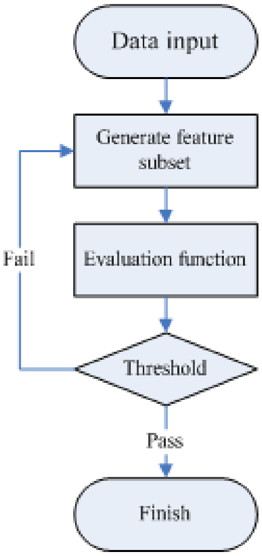

Fig. 1 Feature selection flowchart

In general, feature selection method is divided into two categories: filter and wrapper. The difference of the two categories is on the evaluation functions. The filter method is used in this research. In our method, the evaluation function is by some statistics, for example, distance measurement or dependence measurement. The evaluation function of the wrapper method depends on classifier predict accuracy.

There are two things to be noted in the generation step: one is the generation method; the other is the search method. In this step, we need to generate feature subsets first. The generation method might remove features one by one from all features. We call this method as the forward method. Another one adds features one by one from an empty set. We call that the backward method. Besides these two methods, there are still three other methods: random, compound, and weight. The random method is easy to understand. It generates subsets randomly. The compound method merges the forward method and the backward method. The weight method gives each feature a weighting label.

Under these kinds of generation methods, we need to use a search method to pick out one feature from the original feature set. Three main kinds of search methods are list as follows. The easiest one is complete search. Complete search make a exhaustive search, and its time complexity is *$O(2^n)$*. Random search picks feature randomly, and this method usually sets a counter to stop searching. The time complexity is less than *$O(2^n)$*. Heuristics search use the heuristic algorithm. Its time complexity is *$O(N^2)$*.

The evaluation function step in Fig. 1 uses statistics to measure the subset rationalization. In this paper, we use Euclidian distance (ED) and Pearson correlation coefficient (PCC). The Euclidian distance function is presented in equation 2.

$$ED = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} (X_i - Y_j)^2} \qquad (2)$$

In this function, X_i is the selected feature of sample i in class X, Y_j is the same feature of sample j in class Y, n is the number of class X, and m is the number of class Y. In each iteration, we add one feature into the subset, which ED value is the maximum from all features. This is repeated until our setting iteration stops.

PCC is a popular statistic parameter. And the function is well-known as equation 3.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad (3)$$

We discuss the relation between classes and features. The bigger the absolute value of ρ is, the stronger relationship between feature Y and class X. In general, there are three kinds of situations. They are absolute correlation, positive correlation, and negative correlation, and. All three correlations are tested in our experiments.

3.3 Support Vector Machine

The support vector machine is a supervised learning method for regression and classification. The main concept of SVM is finding out a hyperplane which separates binary class samples into their own groups. The idea is based on the linear separability. For example, assuming a training data set is {(2 0, B), (0 2, B), (2 2, B), (0 -2, R), (-2 0, R), (-2 -2, R)}. There are six balls in two colors B and R. This example is illustrated in Fig. 2. We can find a maximum margin by the points (2,0), (0,2), (-2,0), and (0,-2). There might be lots of lines produced by those points. And we can find that H_3, H_4, H_5, and H_6 cannot separate balls of two different colors. Only the lines passing points (2,0) & (0,2) and (-2,0) & (0,-2) can separate the balls of two colors. H_1 and H_2 are the two that can make the maximum margin. And we can find a hyperplane H by these two lines, which is in the middle of these two lines. H is the classifier of the two color balls problem. If we add a new ball without a label, we can test by this hyperplane and get the result.

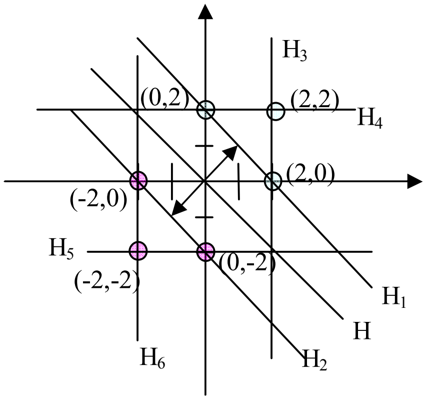

Fig. 2 An example to illustrate SVM

The SVM concept can be understood as follows. The SVM constructs a hyperplane to classify by support vectors. According to the support vectors, The SVM finds out the decision boundaries where the support vectors are located. And the boundaries judge the hyperplane, the distance from each boundary to hyperplane is the same. This is the training stage. After the training stage, testing data are input to the SVM. The SVM accords to the hyperplane to judge how the input data should be classified.

In real cases, many problems might not be linearly separable. They are nonlinear problems. In this case, we need to use a nonlinear kernel to transfer the feature space into another feature space. In the new feature space, the SVM can train as in a linear feature space. Fig. 3 shows the concept.

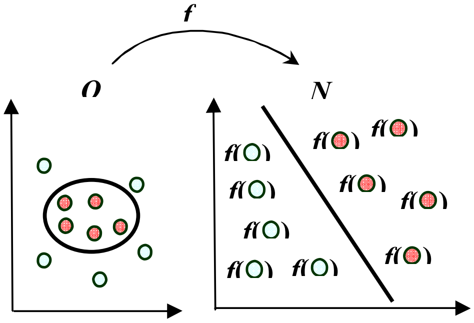

Fig. 3 Kernel functions are used to project current data points to a higher dimension. A nonlinear problem is thus transferred to a linear problem.

In this paper, we discuss about three well-known kernel functions. They are linear kernel, polynomial kernel, and radial basis function (RBF) kernel. Each equation of them is listed in the following.

Linear kernel:

$$K(X,Y)=X\bullet Y \qquad (4)$$

Polynomial kernel:

$$K(X,Y)=(X\bullet Y+1)^p \qquad (5)$$

RBF kernel:

$$K(X,Y)=\frac{e^{-\|X-Y\|^2}}{2\sigma^2} \qquad (6)$$

The open source SVM tool Libsvm is employed in our experiments. Before performing the experiments, we need to transfer the data set into Libsvm data input format. The first column is the class label and the following columns are the features which marked with feature order number. The feature value should be separated with feature number label by colon and feature data type should be in double type. In our experiments, the parameter *p* is set to 3 for polynomial kernel. That is the default value of Libsvm.

3.4 Data sets

The data sets in Kent Ridge Bio-medical Data Set Repository were fixed by the administrator. The data on that site have the experimental values and the gene names. But some of them do not keep the feature names, so we need to find the feature names from the original microarray experiment data in Broad Institute Cancer Program Data Sets. Broad Institute Cancer Program Data Sets collect some of MIT's microarray experiment data and those data are referred by lots of microarray researchers.

We selected the three most referenced data sets from there. They are the AML & ALL data set, the Lung cancer data set, and the Prostate data set. There are 72 samples in the AML & ALL data set, each with 7129 features. 47 of them are ALL data, and 25 are AML data. In the Lung cancer data sets, there are 181 samples with 12533 features. 31 of them are MPM data, the other 150 samples are ADCA. The Prostate cancer data sets have 136 samples and 12600 features in each sample. 77 samples are from patients and the remaining 59 samples are from normal people.

4. Experimental results

First we perform the classification task by SVM with all the features. Different kernel functions are tested here. The original results of each data set are shown in Table 2 where w/o and w/ mean without and with normalization, respectively. The results show that the RBF and polynomial kernels cannot get good results with large features data sets comparing to the linear kernel. This means the linear kernel could be the best choice for this task.

Table 2. Results without feature selection

Kernel & Norm. \ Data sets		AML&ALL	Lung	Prostate
Linear kernel	w/o	96%	100%	89%
	w/	97%	99%	92%
Polynomial kernel	w/o	97%	99%	90%
	w/	65%	83%	57%
RBF kernel	w/o	65%	83%	57%
	w/	68%	87%	63%

Next, experiments with feature selection and/or normalization are performed (as in Table 1). Part of the experimental results is shown here (see Fig. 5 to Fig. 8). In the figures, the feature number is in the *x* axis, and the accuracy is shown in the *y* axis. For comparison with related work, we set the program to select the number of features from 15 to 100. It shows that the polynomial kernel got lower accuracy with normalization. Fig. 8 shows that the best feature numbers of PCC measure method is around 50 to 65. This matches with Vapnik results. But in the ED measure method, it needed at least 60 features to get better results.

In our experiments, most of the results are not as good as the ones with the original data sets, especially when normalization is performed before feature selection. Nevertheless, the results show that the ED feature selection method can get higher accuracy than any PCC feature selection method. And the results of the ED feature selection method cannot get better or equivalent results than the original data set only in the prostate cancer classification. The accuracy of the ED method in AML and ALL is better, which reached 99% with the polynomial kernel as shown in Fig. 5. And in the lung cancer data set, we can get the same accuracy as the original with the linear kernel.

From the results, we think that using the SVM is the main reason for the results of the distance measure method to be better than the dependence measure one. Because the SVM is also analyzed by distance, feature selection task using the distance method can get a better result. All the feature selection results without normalization are shown in Table 3. They are the prediction accuracies with selected features. In the table, PCC used absolute correlation; P. PCC used only positive correlation, and N. PCC used only negative correlation. By comparing the percentages in Table 2 and Table 3, we can say that number of features does not influence the result with the RBF kernel. Also, the results show that the linear kernel is better than others in microarray classification.

To further confirm that the linear kernel is better for this problem. We tried the polynomial kernel with parameter *p* from 1 to 5. It is found that the smaller the *p* is, the better the accuracy. The result is shown in Table 4. Notice that when *p* equals 1, the polynomial kernel is reduced to the linear kernel. Finally, the results of normalization before feature selection are terrible. This shows that the values of microarray experiment data are in a regularization standard, thus no normalization is needed.

Table 3. Results with feature selection

Kernel & F.S. \ Data sets		AML& ALL	Lung	Prostate
Linear kernel	ED	85~96%	94~100%	79~87%
	PCC	68~89%	83~100%	66~84%
	P. PCC	71~89%	85~97%	63~89%
	N. PCC	75~92%	83~100%	72~81%
Poly. kernel	ED	86~99%	93~98%	76~84%
	PCC	69~88%	82~99%	68~78%
	P. PCC	56~92%	81~97%	63~83%
	N. PCC	72~88%	81~99%	63~82%
RBF kernel	ED	65%	83%	57%
	PCC	65%	83%	57%
	P. PCC	65%	83%	57%
	N. PCC	65%	83%	57%

Table 4. Results of ED method and polynomial kernel of different powers.

Power \ Data Sets	AML&ALL	Lung	Prostate
p=1	85~96%	94~100%	79~87%
p=2	86~97%	93~99%	76~85%
p=3(Default)	86~99%	93~98%	76~84%
p=4	65~89%	83~84%	52~54%
p=5	65~89%	83~84%	52~54%

In Table 5, we compare the best result with other methods in the AML and ALL data set. We got the best result with the ED feature selection method.

Table 5. Comparison of best results

Result \ Methods	[11]	[12]	[13]	Proposed Method
Best result	97.8%	94.1%	97.1%	99%

5. Conclusion

In this paper, we employed two feature selection methods and three types of kernel functions to analyze microarray expression data. Our experiments show that the linear kernel and the ED feature selection method are the best match for this problem. And the results show that the distance measure method is better for feature selection with the SVM classifier.

References

[1] Margaret Gardiner-Garden and Timothy G. Littlejohn, “A Comparison of Microarray Databases,”

Briefings in Bioinformatics, Vol. 2, No 2, May 2001, pp. 143-158.

[2] W. S. Noble, “Support Vector Machine Applications in Computational Biology,” http://noble.gs.washington.edu/papers/noble_support.pdf (last accessed May 8, 2008).

[3] G. Piatetsky-Shapiro and P. Tamayo, “Microarray Data Mining: Facing the Challenges,” ACM SIGKDD Explorations Newsletter, Vol. 5, No. 2, Dec. 2003, pp. 1–5.

[4] H. Liu, “Evolving Feature Selection,” IEEE Intelligent Systems, Vol. 20, No. 6, Nov. 2005, pp. 64 -76.

[5] L.C. Molina, L. Belanche, À. Nebot, “Feature Selection Algorithms: A Survey and Experimental Evaluation,” http://www.lsi.upc.es/~belanche/research/R02-62.pdf(last accessed May 8, 2008).

[6] H. Liu, H. Motoda “Feature Selection for Knowledge Discovery and Data Mining,” Kluwer Academic Publishers, 1998.

[7] M. Dash, H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, Vol. 1, No. 3, Mar. 1997, pp. 131-156.

[8] C. J. C. BURGES, “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, Vol. 2, No. 2, Jun. 1998, pp. 121-167.

[9] V. Vapnik, I Guyon, J. Weston, S. Barnhill, “Gene Selection for Cancer Classification using Support Vector Machines,” Machine Learning, Vol. 46, No. 1-3 Jan. 2002, pp. 389-422.

[10] J. Zhang, R. Lee, Y. J. Wang, “Support vector machine classifications for microarray expression data set,” IEEE International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2003), 27 -30 Sep. 2003, pp. 67-71.

[11] W. Fujibuchi and T. Kato, “Classification of heterogeneous microarray data by maximum entropy kernel,” BMC Bioinformatics 2007, Vol. 8, Jul. 26 2007, pp. 267-277.

[12] S. Cho and J. Ryu, “Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features,” PROCEEDINGS OF THE IEEE, Vol. 90, No. 11, Nov. 2002, pp. 1744-1753.

[13] S. Cho and H. Won, “Cancer classification using ensemble of neural networks with multiple significant gene subsets,” Applied Intelligence, Vol. 26, No. 3, Jun. 2007, pp. 243-250.

[14] Kent Ridge Bio-medical Data Set Repository, http://sdmc.lit.org.sg/GEDatasets/Datasets.html (last accessed Dec. 27, 2007).

[15] Broad Institute Cancer Program Data Sets, http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi(last accessed May 8, 2008).

[16] LIBSVM - A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed May 8, 2008).

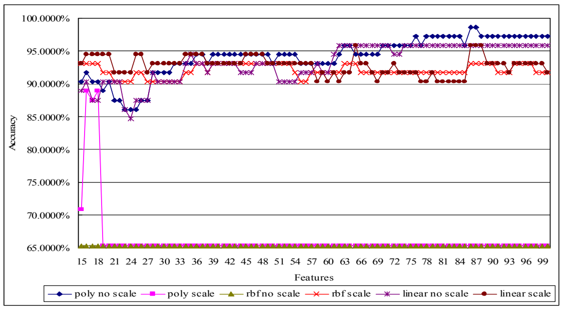

Fig. 5 AML and ALL by ED method without normalization.

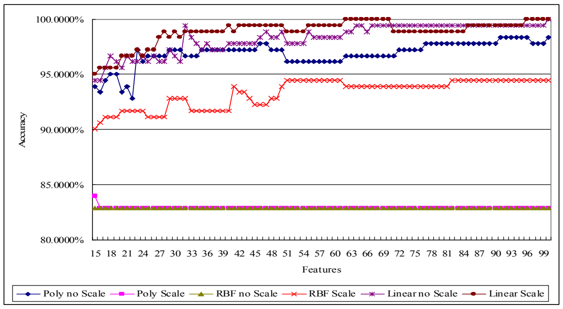

Fig. 6 Lung cancer by ED method without normalization.

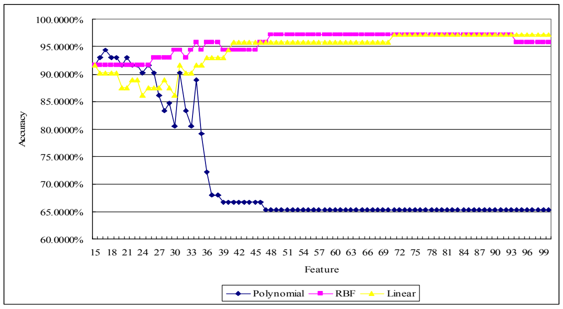

Fig. 7 AML and ALL by ED method with normalization.

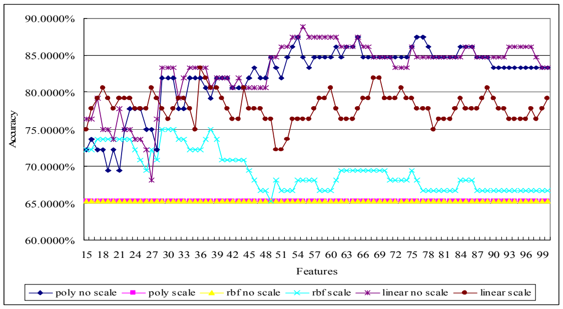

Fig. 8 AML and ALL by PCC method without normalization.