Improved Feature Selection on Microarray Expression Data

Cheng-Wei Hsieh, Hui-Huang Hsu, and Ming-Da Lu
Dept. of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
E-Mail: 892190108@s92.tku.edu.tw, h_hsu@mail.tku.edu.tw, 490191771@s90.tku.edu.tw

Abstract

To identify the relationship between genes and cancers, microarray is always helpful. However, the number of microarray data is quite large, and it is not easy to find out the disease gene from all the microarray data. This paper presents an improved feature selection to filter out the most irrelative or redundant genes. By combining the benefits of "filters" and "wrappers" feature selection, we can not only reduce the processing time of feature selection, but also increase the classification accuracy. In the result, we make a successful result with only 70 genes from 7,129 genes and 70 genes from 12,533 genes in leukemia and Lung cancer microarray data sets. The classification accuracy of leukemia and Lung cancer microarray data are 98.61% and 100%, respectively.

Keywords: Feature Selection, Filter, Wrapper, Support Vector Machine, Microarray

1. Introduction

Microarray is a very useful tool for biologists to discover the gene expression. Because of its high-throughput, large genes expression data could be processed quickly. However, the data sets have numerous features and it is hard to analyze the data sets efficiently. Hence, several techniques are applied in this field to decrease the process complexity. For example, the machine learning technique is one of the most powerful tools to process the microarray data.

Three main applications are classification, gene selection and clustering [1]. In the classification process, a learning model could classify and predict the patient to be ill or not by microarray data. In the second application, gene selection, the main idea is to find out the critical disease genes from all the genes in the microarray. Hence, the patient could be diagnosed ills through those few critical disease genes. In the last application, the clustering could find new biological classes or refining existing ones.

In this paper, we focus on the second application, gene selection, and it also works as the feature selection which selects the most discriminated features from original feature set. In the previous studies, two main models are often applied which are "filters" and "wrappers" [2], but both of them have serious defects. The filters work fast but can not provide a good result. On the other hand, the wrappers' results are good but it works slowly. In our research, an improved feature selection model is provided to solve the above question.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the proposed new feature selection mechanism. Section 4 presents the leaning model SVM. Section 5 lists the experimental results. And finally Section 6 draws the final conclusion.

2. Related work

In recent years, biology gains ground greatly. One of the main reasons is the invention of microarray. It can not only discover proteins' conformation fast, but also could reveal the reaction between genes easily. One of the major researches is the diseases' classification task which can analyze gene expression data to judge which disease is with the patient. This topic is also related to another important field which is feature selection. The classification works without feature selection will affect not only the processing time, but also the classification accuracy.

The support vector machine is a useful supervised learning method in the field of classification. It was proposed by Vapnik in 1995 [3]. In 2002, Vapnik applied 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, Cho and Ryu compared seven classification and seven feature selection methods in AML and ALL data sets. They selected 30 genes from 7,129 genes and the

accuracy was 68.5~94.1% [4]. In 2003, Zhang, Lee, and Wang investigated in microarray expression data set without feature selection. They listed nine advantages and limitations of the SVM on this problem [5]. In 2007, Fujibuchi and Kato discussed three classifiers and six kernels in AML and ALL problems. Their method can reach 97.8% accuracy with a complete feature set. After feature selection, their maximum accuracy is around 87.5% [6]. In 2007, Cho and 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 [7], can get the best accuracy 97.1%, too [8].

The above-mentioned studies show some success for microarray expression data classification. However, further improvements are still in need. Here, we use our feature selection method to examine the same data set, AML&ALL.

3. A new feature selection mechanism

3.1. Filters vs. wrappers

For the current feature selection models, two kinds of theorems are most applied, "filters" and "wrappers". From the point of view of information theorem, the "information" of a set of feature could be calculated by various statistic methods, and that is the core of "filters" kind of feature selection methods. Because of the fast calculation, filters are often applied on high dimensionality feature selection. The complete procedure of "filters" is presented in Figure 1.

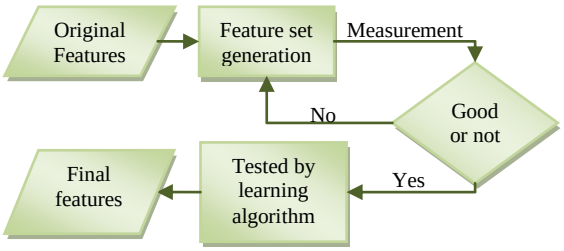

Figure 1. The filter

In Figure 1, three main procedures should be concerned. The first one is the "feature set generation" stage, and this part is also called the "searching" stage. By using different algorithms, the model should define the feature searching order to increase the probability of best feature set generation. If the best feature subset can be generated quickly, the processing time of the filter method would be saved. Next, the second part is the "measurement" stage. This is the step to measure the result of the previous "generation". If the result is not acceptable, it will go back to the "generation" step to generate a different feature set for "measurement". Several kinds of methods are provided to accomplish the measurement. For example, the information gain and mutual information can be applied. Until the result is satisfied, the feature set will be tested by a learning algorithm to show its result, and this is the third part which uses a learning algorithm to test the feature set. Finally, the best feature set will be reported.

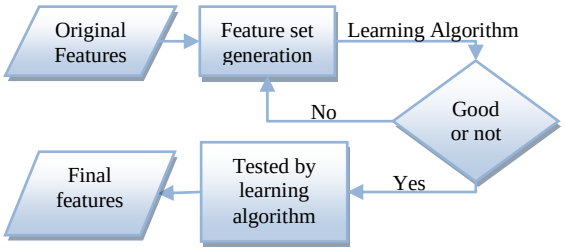

Figure 2. The wrapper

Figure 2 presents the wrapper method. Unlike the filter method, the wrapper method introduces a learning machine to measure the selected feature subset. The measure standard is based on the prediction error rate. Thus the testing result would be better than the filter method which only analyzes the redundancy or relevancy between features. On the other hand, because of the learning model, the processing time of the wrapper method will be long for training.

The key points of the wrapper method are on the feature generation and learning algorithm parts. There are several searching algorithms applied in this field, such as the brute force method, branch and bound, sequential backward/forward search, and the sequential floating search method [9]. The second key point is the learning algorithm. Neural networks, Bayesian networks, and the SVM [10] are often applied on different wrapper problems.

Figure 3. Comparison of filters & wrappers

Figure 3 describes the comparison of filters and wrappers. The processing time of filters is faster than wrappers, but the classification accuracy of filters is not always stable. On the contrary, the wrappers use the learning algorithm to find the best feature set. Hence it can guarantee the accuracy of classification. Therefore, in this paper we intend to take advantage of the merits of both filters and wrappers.

3.2 Combined feature selection

In this paper we combine both filter and wrapper to perform a new feature selection procedure. As described in the previous section, the filter method works fast, but its result is not always stable and the best classification result can not be guaranteed. However, we can consider it as a preprocessing procedure. It can remove part of redundant features. We use two different kinds of filter methods which are F-score [10] and information gain [11] to filter out most features, and then use the wrapper method to improve the classification accuracy. Figure 4 is the architecture of our system.

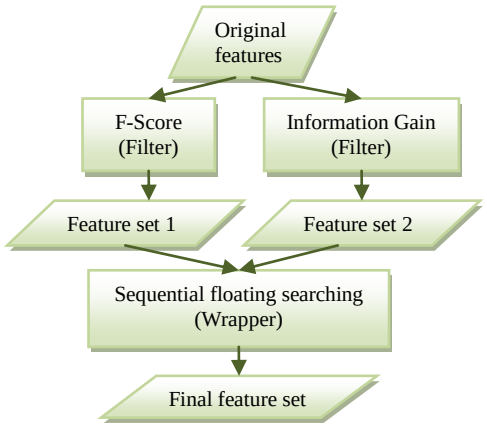

Figure 4. System Architecture

The detail algorithm is in the following:

Step 1: Use F-Score to generate feature set 1.

Step 2: Using information gain to generate feature set 2.

Step 3: Find out the intersection feature set 3 of feature set 1 and 2, and also mark the feature set 4 which is the XOR of feature set 1 and 2 (the concept is presented in Figure 5).

Step 4: Test the accuracy of feature set 3 and then uses the sequential floating search method (SFSM) which starts with feature set 3 and runs on feature set 4. That means the SFSM will pick a feature from feature set 4 and combine it with feature set 3 to test if the result is better than only with feature set 3. In addition, the SFSM will also remove a feature from current feature set to increase the accuracy. That is what SFSM does and it will process repeatedly until the result is best.

Step 5: Step 5 is a special case that if the feature set 3 is the best set then run the sequential backward search method to find out the minimal best feature set.

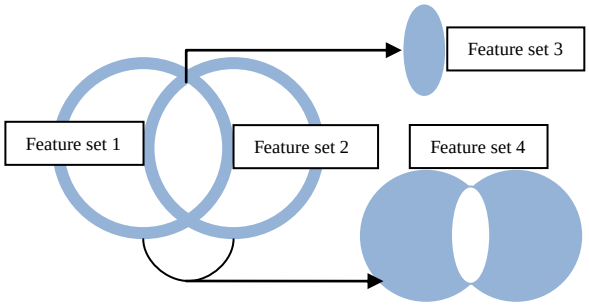

Figure 5. The feature sets' relationship.

In steps 2 and 3, the filters work as the preliminary screening procedure. It will remove most redundant or irrelative features. The reason that we use the F-score and information gain is that the F-score can calculate the degree of difference between the positive class and the negative class. By choosing the most different features, we can separate the classes easily. Equation 1 is the F-score.

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

where $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ and $\bar{x}_i$ are the average of the i_{th} feature of the positive, negative and whole data sets; n_+ and n_- are the number of positive and negative instances; and $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ are the i_{th} feature of the k_{th} positive instance and the i_{th} feature of the k_{th} negative instance.

However, from Equation 1 we can easily notice the weakness of F-score. The F-score only test the discrimination ability with a single feature. Figure 6 presents a situation that two features may perform good discrimination ability if they are both selected, but in this case, these two features would be removed because of the bad discrimination ability of each respective feature.

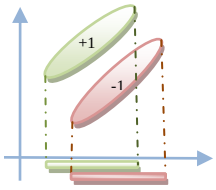

Figure 6. Two separable classes in the two dimension space may not be separated by only observing one feature's distribution.

To improve the filtering result of F-score, we also use information gain (IG). The information gain concerns how much information each feature can provide. Equations 2, 3 and 4 are the calculation steps of information gain. In Equation 1, P_i is the probability of class *i* which appears in total N data, and this equation calculates the information of classes. As for Equation 3, D_{ji} means that the *jth* feature contains *i* kind of different value. The Equation 4 derives the information gain of the *jth* feature by calculating the difference of Equation 2 and 3.

$$\text{Entropy}(N)=\sum_{i=1}^{k} P_i \log_k\left(\frac{1}{P_i}\right)=-\sum_{i=1}^{k} P_i \log_k P_i \quad (2)$$

$$\text{Entropy}(D_j)=\sum_{i=1}^{|D_j|}\frac{D_{ji}}{N}\times \text{Entropy}(D_{ji}) \quad (3)$$

$$IG(D_j)=\text{Entropy}(N)-\text{Entropy}(D_j) \quad (4)$$

When a feature is removed from the feature set, the prediction accuracy change can represent the amount of information of this feature related to the problem. A feature is more important if it contains more information. To remove the features with less information, the remaining feature subset might still be able to result in good prediction accuracy. Figure 7 compares the concept of the two filter methods.

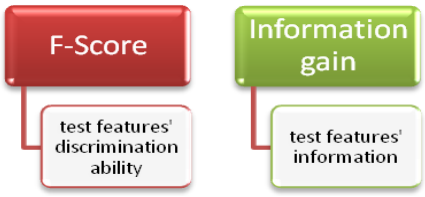

Figure 7. The consideration of F-score and information gain

Step 4 of our algorithm, a wrapper model is applied to improve the accuracy of classification. Here we use the SFSM which is presented in Figure 8. The SFSM is the combination of sequential forward search (SFS) and sequential backward search (SBS). As it presents in Figure 8, the SFSM will perform the SFS that is to select a feature from the candidate feature set and test its classification accuracy with a learning algorithm. It will perform repeatedly until it reaches the stop criterion. Then it will perform the SBS which select a "worst" feature from the current working feature set, and the "worst" means the accuracy could be improved if some feature is removed. To perform these two procedures repeatedly, the best feature set would be produced finally. Here, in order to reduce calculation time, we only test feature set 3 (intersection part of feature set 1 and 2) and feature set 4 (XOR part of feature set 1 and 2). In this step, we use the SFSM to perform the final wrapper procedure, and it only test the feature set 3 and 4. The reason why we choose feature set 3 is that the feature set 3 is the intersection part of F-score and information gain, and this part is the most confidence features. Besides, it also can reduce the execution time which starts with only 1 feature in the original SFSM concept.

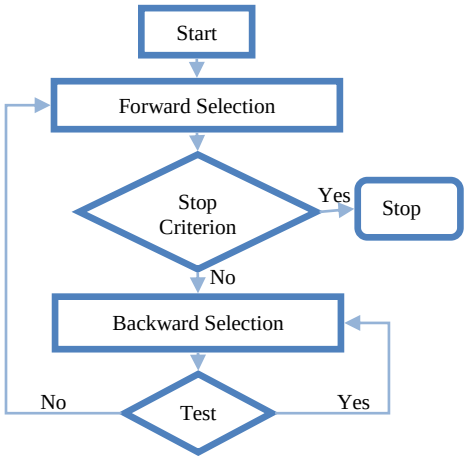

Figure 8. Sequential floating search method (SFSM)

Step 4 is performed searching for the best feature set for classification. Nevertheless in some specific cases, the classification accuracy of intersection part is higher than any combination of intersection and XOR part, and then the procedure will go to step 5 to reduce the feature number with the same or higher classification accuracy. In the step 5, we use the sequential backward searching (SBS) and Figure 9 describe the working flow of SBS which is performed with removing the "worst" feature each time to find out the best minimal feature set.

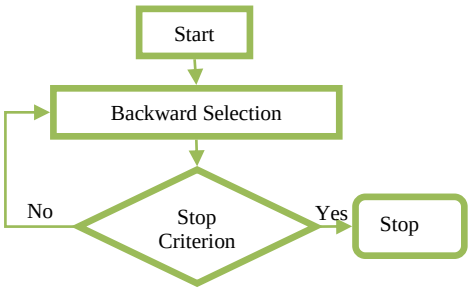

Figure 9. Sequential backward searching (SBS)

4. Learning model and data sets

In the previous section, we say that both the wrappers and filters need a learning model as their testing procedure. Besides, the wrappers also need a learning model to measure the selected features. Here we choose the support vector machine as the learning model. The SVM is based on the SV (support vector) learning. That means the SVM would not always compare the prediction target to all the existing training samples. In contrast, the SVM selects several samples as its SVs, and use these SVs to judge the label of prediction target.
In the testing stage, the SVM model would use the SVs to do the prediction. And also, these SVs would locate on the maximum margin of separation. The SVM is also rated an excellent classifier in practical applications. The SVM can handle more complex nonlinear problems. Hence, the SVM is chosen as the core of our learning model, and we use the RBF kernel as the SVM's kernel function.

4.1. Data sets

The Kent Ridge Bio-medical Data Set Repository [12] saves both experimental values and the gene names. Nevertheless, part of them loses the feature names. Hence, besides the previous database, we also map the feature names from the original microarray experiment data in Broad Institute Cancer Program Data Sets [13] which collects some MIT's microarray experiment data.

In our research, we selected the two most referenced data sets from above database. They are the AML & ALL data set and the Lung cancer data set. Totally 72 samples are in the AML & ALL data set, each with 7,129 features. 47 of them are ALL data, and 25 are AML data. In the Lung cancer data set, there are 181 samples, each with 12,533 features. 31 of them are MPM data, the other 150 samples are ADCA.

5. Experimental results

In the first filter procedure, we test two kinds of filters, F-score and information gain (IG). The result is listed in Table 1.

Table 1. Result of preliminary screening

Data set	Method	Threshold	Features	Accuracy (5-fold cross validation)
AML & ALL	-	-	7,129	68.06%
AML & ALL	F-score	50	873	98.61%
AML & ALL	IG	0.64	1,510	98.61%
Lung cancer	-	-	12,533	86.74%
Lung cancer	F-score	100	996	99.45%
Lung cancer	IG	0.455	1,571	99.45%

In Table 1, the threshold setting is resolved from a greedy process. Originally, the classification accuracy of AML & ALL and Lung cancer datasets are 68.06% and 86.74% respectively with the whole set of features 7,129 and 12,533. After the filter processes of F-score and IG, the AML & ALL feature set is reduced to 873 and 1,510 features, and the prediction accuracy is increased to 98.61%. As for the Lung cancer data set, the F-score and IG reduced the features from 12,533 to 996 and 1,571, and also raised the accuracy to 99.45%.

Next, Table 2 and Table 3 show the feature numbers with the F-score and IG filtering.

Table 2. Number of features (AML&ALL)

Relationship	Number
Total features	7,129
F-score ∩ IG	276
(F-score ∪ IG)- (F-score ∩ IG)	1,831

Table 3. Number of features (Lung cancer)

Relationship	Number
Total feature set	12,533
F-score ∩ IG	326
(F-score ∪ IG)- (F-score ∩ IG)	1,915

From Table 2, there are 276 features (F-score ∩ IG) which represent the confident features of F-score and IG, and only 1,831 features ((F-score ∪

IG)- (F-score∩IG)) would be tested in the SFSM procedure. This can greatly decrease the wrapper's processing time, and limit the number of testing features. In the end, the wrapper will work much faster. At the same time, Table 3 shows the confident part is with 326 features and only 1,915 features would be tested by SFSM.

Here, we list the prediction results of (F-score∩IG) and best feature set in Table 4. Because in these cases, the SFSM could not improve the prediction accuracy anymore, the confidence part will process the SBSM to reduce the feature number. Table 4 shows that the best feature sets of AML&ALL and Lung cancer are both 70, and the classification accuracy are 98.61% and 100%.

Table 4. The prediction accuracy after combination of feature subsets

Data set	Relationship	Dimension	Accuracy (5-fold cross validation)
AML & ALL	F-score∩IG	355	98.61%
AML & ALL	Best feature set	70	98.61%
Lung cancer	F-score∩IG	326	99.45%
Lung cancer	Best feature set	70	100%

Finally, Table 5 compares our proposed method with other existing feature selection methods on the AML&ALL data set. The result shows the success of our model.

Table 5. The comparison with other methods (AML&ALL)

Methods / Results	[6]	[7]	[8]	Proposed method
Accuracy	97.8%	94.1%	97.1%	98.61%
# of features	170	30	50	70

6. Conclusion

Microarray is always a useful tool to identify disease genes. Nevertheless, the data in microarray are quite large. In general, the biologists only want to know which genes are really related to the diseases. To remove most redundant and irrelevant genes is not an easy work. In this paper, we use feature selection to perform the gene selection. Unlike the previous research, we can not only improve the accuracy of "filters", but also reduce the working time of "wrappers". By combining the "filters" and "wrappers", we show the success of our model. Comparing to the previous researches, our classification accuracy of leukemia microarray data set is the highest.

References

[1] 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.

[2] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, Vol. 97, 1997, pp. 273-324.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15, No. 11, 1994, pp. 1119–1125.

[10] LIBSVM - A Library for Support Vector Machines,http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed April 3, 2008)

[11] J. R. Quinlan, *Discovering Rules from Large Collections of Examples: A Case Study*, In Michie, D. (Ed.), Expert Systems in the Microelectronic Age, Edinburgh, Scotland: Edinburgh University Press, 1979, pp. 168-201.

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

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