Review on Feature Selection Techniques of DNA Microarray Data

Ammu P K
Sree Chitra Thirunal College of Engineering
Thiruvananthapuram
Kerala

Preeja V
Sree Chitra Thirunal College of Engineering
Thiruvananthapuram
Kerala

ABSTRACT
Feature selection from DNA microarray data is one of the most important procedures in bioinformatics. The huge dimensionality of the DNA microarray data becomes a problem when it is used for cancer classification. This problem can be alleviated by employing feature selection as a preprocessing step in classification.

This paper reviews some of the major feature selection techniques employed in microarray data and points out the merits and demerits of various approaches.

General Terms
Bioinformatics, Evolutionary algorithms

Keywords
Microarray, DNA, Feature Selection, Cancer classification.

1. INTRODUCTION
A DNA microarray is a collection of DNA spots attached to a solid surface. DNA microarrays are used to measure the expression levels of thousands of genes simultaneously. In DNA microarray experiments, the sample from which the expression levels of genes are to be measured is applied to the DNA microarray. Based on the amount by which the sample hybridize with the DNA spots on the microarray, the expression levels of genes are measured. The resultant dataset known as DNA microarray dataset is used for cancer classification. Various classifiers such as support vector machine (SVM), multi layer perceptron (MLP), K nearest neighbor (KNN) etc are employed for cancer classification. The overview of cancer prediction system is given in figure 1.

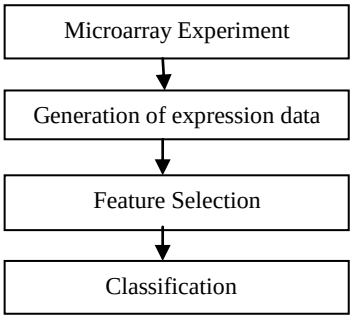
One of the main problems that exist with the microarray dataset is the 'curse of dimensionality'. The number of parameters in each sample is much smaller than the number of samples used for training the classifier. This may lead to over fitting of the classifier. This problem can be alleviated by employing feature selection. The main objectives of feature selection can be described as follows

1. To get rid of irrelevant and noisy genes from the input data set
2. To speed up the processing of data by reducing the dimensionality
3. To avoid over fitting of the classifier

This paper is organized as follows. In section 2 the various dimensionality reduction techniques are discussed. In section 3 the optimality of the various feature selection techniques are discussed.

Section 4 reviews some relevant feature selection approaches. Section 5 discusses the impact of various classifiers on feature selection and section 6 provides a comparison of various feature selection algorithms.

2. DIMENSIONALITY REDUCTION TECHNIQUES
One of the important objectives of feature selection is to reduce the dimensionality of data. The dimensionality reduction techniques are mainly classified as feature extraction and feature selection techniques. Feature extraction involves various techniques such as Principal Component Analysis [1], various clustering techniques such as K-means clustering [2] etc. Feature selection techniques can be broadly classified as univariate methods and multivariate methods



Uni vs multi variate

Figure 1. Various steps in the cancer prediction system

based on the criteria used for evaluating the genes. Univariate methods analyze a single variable at a time whereas multivariate methods analyze more than one variable at a time. Based on the classification approach used, feature selection techniques can be classified as filter, wrapper, embedded methods and hybrid methods. Filter methods can be either univariate or multivariate. Filter methods usually ranks the genes based on some univariate measure and selects the best genes among them. Filter methods can also be ranked into parametric and non parametric techniques. Examples of parametric techniques are information gain [3], Euclidian distance [4], signal to noise ratio [3] etc. Examples of nonparametric techniques are correlation coefficient [4], significant analysis of microarray [5] etc.

Filter, wrapper, hybrid methods

Filter

Wrapper methods, hybrid methods and embedded methods fall in the multivariate category. Wrapper methods makes search for a subset that is found to be optimal with respect to a subset evaluator such as a classifier. Examples of wrapper methods are various optimization algorithms such as Ant Colony Optimization (ACO) [6], Biogeography Based Optimization (BBO) [7], Genetic Algorithm (GA) [8] etc. Embedded methods incorporates the search algorithm into a classifier. They construct a model with the features and analyze the model to infer the importance of variables. An example for embedded method is SVM-RFE [9].

Wrapper

Embedded

Disadvantage of univariate filter

One of the major disadvantages of univariate filter method is that the features selected after preprocessing are redundant. This is due to the reason that the features are ranked based on same ranking criteria and only the highly ranked genes are selected after preprocessing. As a result similar features get selected. Redundancy elimination methods are a solution to this problem. Various redundancy elimination methods have been proposed in literature such as max relevance min redundancy method [10], redundancy based filter [11] etc. Hybrid methods on the other hand are a combination of several of the above said approaches.

The MRMR method proposed by Ding, C. et al [10] searches for subsets from the entire population satisfying minimum redundancy maximum relevance criteria. This method minimizes redundancy using various criteria such as minimizing inter correlation, maximizing Euclidian distance etc. Similarly relevance is maximized using various criteria such as maximizing mutual information with target phenotype. The features so obtained have the following characteristics

1. They represent broader spectrum of characteristics than those obtained through standard ranking methods.
2. . They are more robust.
3. They have good generalization capabilities

The taxonomy of feature selection techniques is given in figure 2. Table 1 compares the various feature selection algorithms and point out their advantages and disadvantages.

3. OPTIMALITY OF THE FEATURE SELECTION METHODS

The feature selection methods can be classified as optimal and suboptimal methods based on the optimality of the solutions obtained [12].
Selection methods involving exhaustive search, branch and bound search are included in the optimal selection method and selection methods involving evaluation of individual features, sequential forward selection, sequential backward selection

Optimal and suboptimal methods

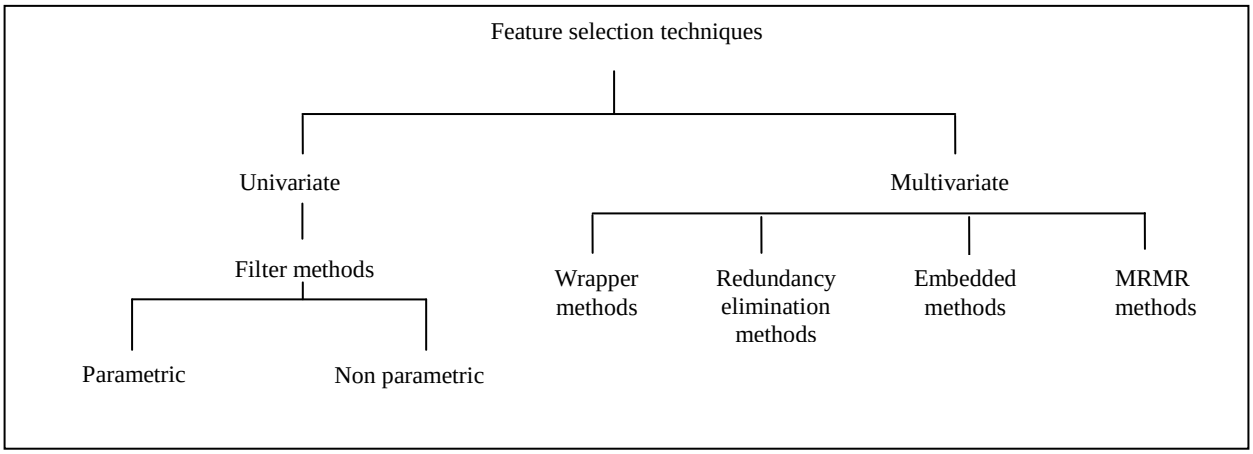


Figure 2. Taxonomy of feature selection techniques

etc are included in the suboptimal selection methods [13].

Since the wrapper methods includes exhaustive search, we can categorise it as optimal, filter methods as suboptimal since it includes evaluation of individual features. Similarly we can categorize the various feature selection methods as optimal and suboptimal based on the search procedure used.

4. REVIEW OF SOME FEATURE SELECTION METHODS

Filter methods >> Redundancy

The following subsections review some significant methods employed for feature selection. Filter methods pose the serious disadvantage of redundancy, hence only wrapper, hybrid and redundancy elimination methods are discussed here. Table 1 depicts the comparison of various feature selection algorithms.

4.1 Biogeography based informative gene selection

Biogeography based optimization (BBO) is an optimization algorithm introduced by Dan Simon [14]. The algorithm works on the basis of migration of species between different habitats and the process of mutation. BBO in combination with SVM has been observed to be a good wrapper approach for feature selection in DNA microarray [7]. The fitness function evaluates the cross validation accuracies of gene subsets using SVM. The work proposed in [7] however gives importance to the selection of informative genes during mutation. The information gain of a gene indicates how much informative the gene is for classification. Thus the algorithm tests the predictive power of data set and also help in retaining the informative genes. However one of the disadvantage of approach is that the method takes exponential time complexity, so is computationally expensive. The main advantages of the algorithm are as follows

1. The method is optimal.
2. The problem of redundancy doesn't exist.
3. The genes with high information gain are retained

Table 1. Comparison of various feature selection approaches

Methods	**Univariate Filter methods**	**Wrapper methods**	**Embedded methods**	**Minimum redundancy maximum relevance methods**	**Hybrid Methods**
Advantages	Time complexity is O(n), which is low as compared to other methods. Simple	Tests the predictive power of genes Carries out exhaustive search, generating optimal solutions	Tests the predictive power of genes Less computational complexity compared to wrapper method Less prone to overfitting	Eliminates redundancy Tests the relevance of genes in combination with other genes	Can combines the advantages of various approaches
Disadvantages	Creates redundancy Evaluates genes based on their individual scores, ignores their relevance in combination with other genes	Exponential time complexity Complex Doesn't take enough measures to eliminate redundancy Prone to overfitting	Heavily dependent on the model, so they can fail to fit the data well	Time complexity more as compared to filter methods $O(nm^2)$.	Time complexity may increase

4.2 Redundant gene selection based on particle swarm optimization (RGS-PSO)

Particle swarm optimization (PSO) is an algorithm proposed by Eberhart and Kennedy [15]. The algorithm works on the basis of movement of particles in a search space. RGS-PSO proposed in [16] is a hybrid of wrapper and MRMR approaches. The fitness function in PSO selects feature set s that maximizes the merit given in equation 1.

$$\text{Merit (s)} = \frac{\Sigma_{i=1:k} \text{ correlation of fi with target class}}{\Sigma_{i=1:k} \text{ correlation of fi with each feature in s}} \quad (1)$$

The numerator in equation 1 computes the maximum correlation between each feature fi and the class and the denominator computes the maximum inter correlation between every pair of features in set s. In other words the numerator of equation 1 tests the predictive power of the set and denominator computes the redundancy. The method is optimal, and eliminates redundancy.

4.3 Redundancy based filter (RBF)

Redundancy based feature selection approaches can be used to remove redundant genes from the selected genes as the resultant gene set can achieve a better representation of the target class. The redundancy based filter (RBF) proposed by Yu, L and Liu, H [10] is a redundancy based feature selection algorithm. The RBF works on the basis of finding and removing approximate redundant cover for each feature. A feature Fi forms an approximate redundant cover for Fj if and only if correlation (Fi, Class) >= correlation (Fj, class) and correlation (Fi, class) >= correlation (Fi, Fj, Class). The RBF algorithm is as shown below

1. Order the features in a list based on the decreasing order of predictability.
2. Select the first feature Fi from the list. Find and remove all features from the list for which Fi forms an approximate redundant cover.
3. Add Fi to the end of the list and repeat step 2 until the end of the list.

RBF is substantially faster compared to other feature selection methods because a large number of features get removed in each round. The best case time complexity of the algorithm is O(n) when only one feature is selected (n is the number of features) and the worst cast time complexity is $O(n^2)$ when all features gets selected.

4.4 IG-GA: A hybrid filter wrapper approach

Hybrid approaches can overcome the demerits of various above defined algorithms. A two stage hybrid filter wrapper method proposed by Karzynski, M .et al [8]. The two stages of the approach are as follows.

1. In the first stage a subset of the original feature set is obtained by applying information gain as the filtering criteria. This is done by ranking the genes based on Information gain, and those with

information gain values above a threshold value will go to the next stage.

2. In the second stage the genetic algorithm is applied to the set of filtered genes.

The proposed method in [8] employs KNN classifier for checking the cross validation accuracies of genes. The advantages of this method are as follows

1. Applying the filter approach before the wrapper approach fastens the wrapper approach since the dimensionality of the dataset is reduced, thereby reducing the computational complexity of the wrapper approach.

2. It may increase the classification accuracy.

One of the disadvantages of this approach is that the time complexity may increase. The time complexity of the approach is found to be O(n lg n + nmpg) where n represents the number of samples, m represents the dimension of the data sets, p represents the population size and g represents the number of generations.

4.5 Gene selection based on dependency of features

A hybrid feature selection approach that selects the features based on their dependency is proposed by Zhang, L. J. et al [17]. The features are classified as independent, half dependent and dependent features. Independent features are those features that doesn't depend on any other features and are essential for classification. Eg: consider Table 2

Table 2. Features of two samples and their catagory

Width	Eye color	Length	Class
3	Black	9	Salmon
3	Black	3	Seabass

Here length is the only feature that makes the two samples different. Such features that uniquely identify each class independent of any other features are known as independent features.

Half dependent features are more relevant in correlation with other features. Dependent features are fully dependent on other features. They are only relevant when they are with other features. The algorithm proposed in [17] first selects independent features into a set, then from the remaining features in the feature set, selects the half dependent features. Half dependent features have higher correlation with the class ,when with other features in the set than by itself. Sequential forward selection is employed for their selection. After the selection of half dependent features from the remaining features, the features having higher accuracy with other features in the set are selected, ie the dependent features are selected. A classifier is used for this purpose.

Let n be the number of features and m be the number of samples. The selection of independent features takes $O(m^2n)$ time complexity. The selection of half dependent features take O(n) time complexity and time complexity of the selection of dependent features depends on the classification algorithm chosen.

The main advantages of the algorithm are as follows.

1. The method is optimal.
2. The method combines the advantages of wrapper methods and embedded methods.
3. The relevant features are retained in this approach.

The main disadvantage of the approach is its time complexity.

1. The algorithm doesn't gives importance to elimination of redundant features.
2. The approach is computationally expensive.

5. IMPACT OF VARIOUS CLASSIFIERS ON FEATURE SELECTION

Various classifiers such as KNN, SVM, neural networks, random forests etc are employed for cancer classification. Among these, SVM is widely used for cancer classification. The work proposed by Karzynski, M. et al [8] makes use of KNN for cross validation. The major distinctive features of KNN pointed out in [8] are as follows

1. Simplicity
2. Easy implementation
3. KNN is not negatively affected when the training data set is large.
4. KNN is indifferent to noisy data.
5. Time complexity of KNN is smaller as compared to that of SVM. Time complexity of KNN is O (nm), where n is the number of features and m is the number of samples.

SVM on the other hand provides more classification accuracy as compared to KNN, back propagation neural network, decision trees etc. SVM is known to be the best method in classification of microarray gene expression data. The merits of SVM pointed out in [18] are as follows

1. Good generalization capabilities
2. Flexibility in choosing a similarity function
3. Ability to identify outliers
4. Ability to handle large feature spaces

However, while employing various pre processing methods before classification, it will be useful to consider the efficiency of the pre processing methods and classification algorithms together. The efficiencies of MLP, SVM, KNN, decision tree and self organizing map in combination with various pre processing steps such as pearson coefficient, spearman coefficient, Euclidian distance, Cosine coefficient, information gain, mutual information and signal to noise ratio have been studied by Ryu, J., and Cho, S.B [19].

However the various results obtained for KNN, SVM and MLP are produced here in Table 3 for the purpose of completion. Among the various methods MLP in combination with pearson coefficient shows the best performance.

Table 3. Classification accuracies of various classifiers in combination with various feature selection techniques

Classifier / Feature selection	MLP	SVM	KNN
Pearson coefficient	97.1	79.4	29.4
Spearman coefficient	70.6	88.2	32.4
Euclidian distance	97.1	58.5	32.4
Cosine Coefficient	79.4	94.1	23.5
Information gain	91.2	88.2	58.8
Mutual information	67.6	58.5	58.8
Signal to noise ratio	94.1	58.5	8.8

6. COMPARISON OF VARIOUS FEATURE SELECTION ALGORITHMS

A comparison of the various feature selection algorithms described in section 4 is provided in Table 4. In terms of time complexity, the filter approaches are the best and in terms of efficiency wrapper and hybrid approaches are the best.

7. CONCLUSION

The various approaches used for feature selection are discussed in this paper. However, it is not possible to say that one approach is universally better compared to other methods. Every method has its own advantages and disadvantages and behaves differently on different datasets. The various algorithms are compared here based on their common behaviours.

Table 4. Comparison of some major feature selection algorithms

Algorithm	**Optimality**	**Catagory**	**Selection process**	**Time complexity**
Biogeography based informative gene selection	Optimal	Wrapper	Exhaustive search	Exponential
Redundant gene selection based on particle swarm optimization	Optimal	Minimal redundancy maximum relevance method	Exhaustive search	Exponential
Redundancy based filter	Suboptimal	Redundancy elimination method	Sequential forward selection	O(n)
IG-GA	Optimal	Hybrid	Branch and bound search	O(n lg n + nmpg)
Gene selection based on dependency of features	Optimal	Hybrid	Sequential forward selection	Ist part- $O(m^2n)$ IInd part- O(n) IIIrd part-depends on the classification algorithm

8. References

[1] Raychaudhuri, S., Stuart, J. M., and Altman, R. B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. In Pacific Symposium on Biocomputing. 455-466.

[2] Shannon, W., Culverhouse, R., and Duncan, J. 2003. Analyzing Microarray Data using Cluster Analysis. Pharmacogenomics. 4(1). 41-52.

[3] Wang, Z. 2005. Neuro-Fuzzy Modeling for Microarray Cancer Gene Expression Data. In Proceedings of the Second International Symposium on Evolving Fuzzy Systems. 241 – 246.

[4] Cho, S. B., and Won, H. H. 2003. Machine Learning in DNA Microarray Analysis for Cancer Classification. In proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics. 189-198.

[5] Fung, Y. M., and Ng, V.T.Y. 2003. Classification of Heterogeneous Gene Expression Data. Article. 5(2). 69-78.

[6] Chiang, Y. M., Lin, S.Y. 2008. The application of ant colony optimization for gene selection in microarray-based cancer classification. In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. 12-15.

[7] Nikumbh, S., Ghosh, S., and Jayaraman, V. K. 2012. Biogeography-Based Informative Gene Selection and Cancer Classification Using SVM and Random Forests. In IEEE Congress on Evolutionary Computation. 1-6.

[8] Karzynski, M., Mateos, A., and Dopazo, J. 2003. Using a Genetic Algorithm and a Perceptron for Feature Selection and Supervised Class Learning in DNA Microarray Data. Artificial intelligence review. 20(1). 39 – 51.

[9] Yu, Y. SVM-RFE Algorithm for Gene Feature Selection. Technical report. University of Delaware.

[10] Ding, C., and Peng, H. Minimum Redundancy Feature Selection from microarray gene expression data. J Bioinform Comput Biol. 523—529.

[11] Yu, L., and Liu, H. 2004. Redundancy Based Feature Selection for Microarray Data. In Proceedings of SIGKDD. 737-742.

[12] Ali, M. L., 2005. Feature Selection of DNA Microarray Data. University of Windsor.

[13] Duin, R. P., Jain, W., Jain, A. K ., and Mao, J. 2000. Statistical Pattern Recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(1). 4-37.

[14] Simon, D. 2008. Biogeography Based Optimization. IEEE transactions on evolutionary computation. 12(6). 702-713.

[15] Kennedy, J., and Eberhart, R. C. 1995. Particle swarm optimization. In IEEE International Conference on Neural Networks. 1942 - 1948

[16] Chen, F., Zeng, X. Q., Li, G.Z. et al. Redundant Gene Selection based on particle swarm optimization. In Proceedings of the 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 10-16

[17] Zhang, L. J., Li, Z. J., and Hu, X.H. A Hybrid Gene Selection Method for Cancer Classification. In Proceedings of 2004 international conference on Machine Learning and Cybernetics. 2537 – 2542.

[18] George, G., and Raj, V. C. Review on feature selection techniques and the impact of svm for cancer classification using gene expression profile, International Journal of Computer Science & Engineering Survey. 2(3). 16-27.

[19] Ryu, J., and Cho, S.B. Towards Optimal Feature and Classifier for Gene Expression Classification of Cancer. In Proceedings of the 2002 AFSS International Conference on Fuzzy Systems. 310-317.