Feature Subset Selection in Large Dimensionality using Correlation based GA-SVM

Binita Kumari
Assistant Professor, Department of Computer Science
ITER, SOA University
Orissa, INDIA

ABSTRACT

The microarray can be used to measure the changes in the expression levels of thousands of genes simultaneously, to detect SNPs or to genotype or resequence mutant genomes. The high dimensional feature vectors of microarray impose a high dimensional cost as well as the risk of overfitting during classification. Feature selection is one way of reducing the dimensionality. Effective feature selection can be done based on correlation between attributes.

In this paper, we introduce a correlation based wrapper algorithm for feature selection using genetic algorithm (GA) and Support Vector Machines with kernel functions for classification. We compare our approach with existing algorithm.

General Terms

Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords

Microarray, Feature Selection, Wrapper method, support vector machine, Classification.

1. INTRODUCTION

A DNA microarray (also commonly known as gene chip, DNA chip, or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data Accurate disease diagnosis is vital for the successful application of specific treatments. The DNA microarray technology is providing great opportunities in reshaping the biomedical science. An orderly and computational analysis of microarray datasets is a motivating way to study and understand many aspects of underlying biological process. Machine learning methods to are used to analyse and understand the data generated by this new kind of experiments. The analysis involves class prediction (supervised classification), regression, feature selection, principal component analysis, outlier detection, discovering of gene relationships and cluster analysis (unsupervised classification) [1,3].

For most biological problems, information about type (class) of each cell line exists indicating whether the tissue is diseased or healthy. By means of the interesting class information, the DNA microarray analysis can be formulated as a classic supervised classification task.

Feature selection can be applied to both supervised and unsupervised learning; we focus here on the problem of supervised learning (classification), where the class labels are known beforehand.

Microarray technology is used to study the expression of many genes at a time. The high dimensional [2,5] feature vectors of microarray data often impose a high computational cost as well as the risk of “overfitting” at the time of classification. Thus it is necessary to reduce the dimensionality through ways like feature selection.

A microarray chip or data can be analyzed as shown in figure 1.First the microarray dataset is normalized so that there are no missing values and the data is scaled between a specific range. Then feature selection is done as a result of which we get the key genes. Then the classification or clustering is done and the output is interpreted to get the required biological information

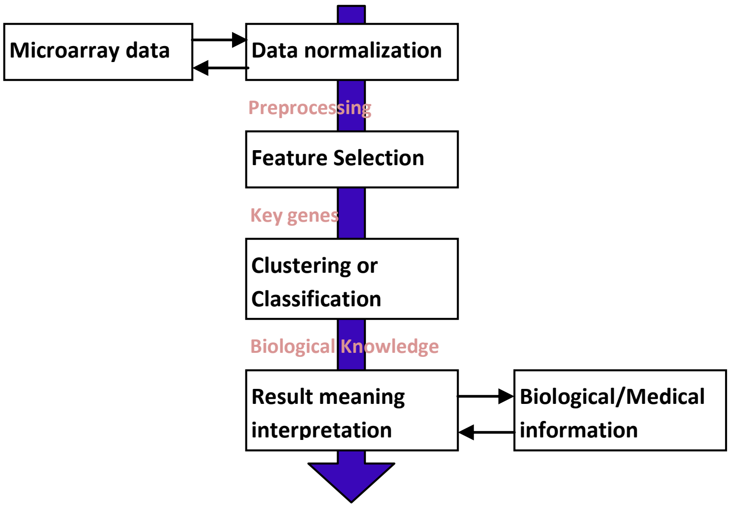

Fig.1 : Microarray chip analysis

The selection of relevant features and elimination of irrelevant ones is a great problem. Before an induction algorithm can be applied to a training dataset to make decisions about test cases, it must decide about which attributes to be selected and which to be ignored.

Irrelevant features increase the measurement cost, decrease the classification accuracy and add to making the computation complex. Obviously, one would like to use only those attributes that are relevant to the target concept.

The rest of the paper is organized as follows: a brief review of the existing techniques of feature selection, related work and proposed work, and comparative results followed by conclusion and future work.

2. RELATED AND PROPOSED WORK

Feature selection (also known as subset selection) entails choosing the feature subset that maximizes the prediction or classification accuracy. The best subset contains the least number of features that most contribute towards accuracy.

2.1 Feature Selection

Feature selection (also known as subset selection) is a process commonly used in machine learning, where a subset of features is selected from the available data for application of a learning algorithm [5]. So we prefer the model with the smallest possible number of parameters that adequately represent the data. Selecting the best feature subset is a NP complete problem. The task is challenging because first, the features which do not appear relevant singly may be highly relevant when taken with other features. Second, relevant features may be redundant so that omission of some of them will remove unnecessary complexity. An exhaustive search of all possible subsets of features will guarantee the best feature subset. The best subset contains the least number of features that most contribute towards accuracy.

There are two approaches of feature selection [2,4,5]:

Forward selection:

(i)Start with no variables.(ii)Add the variables one by one, at each step adding the feature that has the minimum error.(iii)Repeat the above step until any further addition does not signify any decrease in error.

Backward selection:

(i)Start with all variables.(ii)Remove the variables one by one, at each step removing the feature that has the highest error.(iii)Repeat the above step until any further removal increases the error significantly.

The two broad categories of feature subset selection have been proposed: filter and wrapper [4,5]. Filter techniques assess the relevance of features by looking at the intrinsic properties of the data. In filter criteria, all the features are scored and ranked based on certain statistical criteria. The features with the highest ranking values are selected and the low scoring features are removed.. Filter methods (fig 2) are fast and independent of the classifier but ignore the feature dependencies and also ignores the interaction with the classifier. They also easily scale to very high-dimensional dataset. As a result feature selection need to be done only once and then different classifiers can be evaluated. The common disadvantage of filter methods is that they ignore the interaction with the classifier and each feature is considered independently thus ignoring feature dependencies In addition, it is not clear how to determine the threshold point for rankings to select only the required features and exclude noise.

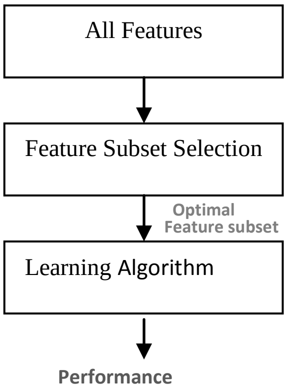

Fig.2 The feature filter approach

Wrapper methods embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm. To search the space of all feature subsets, a search algorithm is then 'wrapped' around the classification model. However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset.

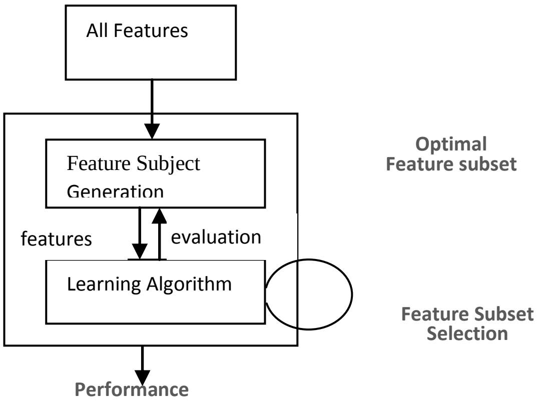

Fig.3 The feature Wrapper approach

Thus feature selection [4,3] is of considerable importance in classification as it (i)Reduces the effects of curse of dimensionality(ii)Helps in learning the model(iii)Minimizes cost of computation(iv)Helps in achieving good accuracy

2.2 Introduction to Support Vector Machine

Support Vector Machines (SVM)[3,5] is a classification system derived from statistical learning theory. It has been applied successfully in fields such as text categorization, hand-written character recognition, image classification, biosequences analysis, etc.

Support vector machine is a method of obtaining the optimal boundary of two sets in a vector space independently on the probabilistic distributions of training vectors in the sets. Its fundamental idea is very simple; locating the boundary that is most distant from the vectors nearest to the boundary in both of the sets. This simple idea is a traditional one, however, recently has attracted much attention again. This is because of the introduction of *kernel method,* which is equivalent to a transformation of the vector space for locating a nonlinear boundary. The SVM separates the classes with a decision surface that maximizes the margin between the classes. The surface is often called the optimal hyperplane, and the data points closest to the hyperplane are called support vectors. The support vectors are the critical elements of the training set. The SVM can be adapted to become a nonlinear classifier through the use of nonlinear kernels. While SVM is a binary classifier in its simplest form, it can function as a multiclass classifier by combining several binary SVM classifiers (creating a binary classifier for each possible pair of classes). There are different types of SVM classifier kernel functions, such as linear, polynomial, radial basis function (RBF), and sigmoid.

2.3 Introduction to Genetic Algorithm

A genetic algorithm (GA) is a general adaptive optimization search methodology based on a direct analogy to Darwinian natural selection and genetics in biological systems. It has been proved to be a promising alternative to conventional heuristic methods. Based on the Darwinian principle of 'survival of the fittest', GA works with a set of candidate solutions called a population and obtains the optimal solution after a series of iterative computations.

GA evaluates each individual's fitness, i.e. quality of the solution, through a fitness function. The fitter chromosomes have higher probability to be kept in the next generation or be selected into the recombination pool using the tournament selection methods. If the fittest individual or chromosome in a population cannot meet the requirement, successive populations will be reproduced to provide more alternate solutions. The crossover and mutation functions are the main operators that randomly transform the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable results are obtained. Associated with the characteristics of exploitation and exploration search, GA can deal with large search spaces efficiently, and hence has less chance to get local optimal solution than other algorithms.

2.4 Correlation based measures

In general, a feature is *good* if it is *relevant* to the class concept but is not *redundant* to any of the other relevant features. If we adopt the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not highly correlated to any of the other features.

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Under the first approach, the most well known measure is *linear correlation coefficient.*

There are several benefits of choosing linear correlation as a feature goodness measure for classification. First, it helps remove features with near zero linear correlation to the class. Second, it helps to reduce redundancy among selected features. It is known that if data is linearly separable in the original representation, it is still linearly separable if all but one of a group of linearly dependent features are removed

2.5 Proposed work

The basic idea of the GA-SVM [2] method is to remove the features which are less fit. The features having high fitness value and high classification accuracy are retained for the evolution. This is achieved by the following iterative algorithm:

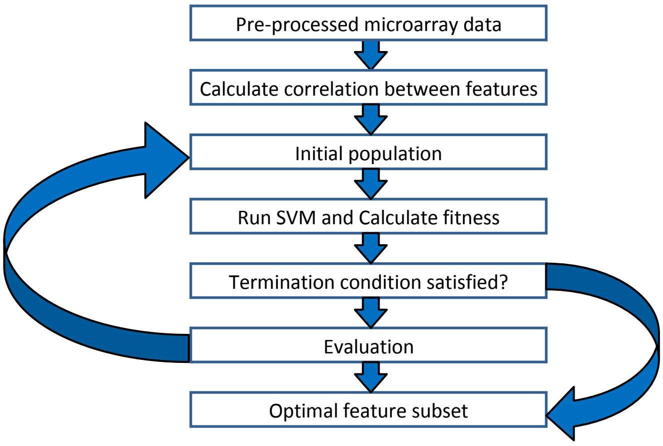
Pre-processed microarray data
Calculate correlation between features
Initial population
Run SVM and Calculate fitness
Termination condition satisfied?
Evaluation
Optimal feature subset

Fig.4 Flowchart of Corr-GA-SVM wrapper approach

Corr-GA-SVM Algorithm for Feature Selection

1. Data pre-processing and Model Selection i.e choose the kernel parameters.
2. Calculate the correlation between features.
3. Create an initial population of certain size.
4. Calculate the fitness value of each individual in the initial population by implementing SVM and finding out the classification accuracy. Rank the features according to their fitness i.e according to their classification accuracy.
5. Select a certain number of individuals with high fitness value to retain them in next generation.
6. Check whether termination conditions are satisfied. If so, then evolution stops and the optimal result is obtained else evolution continues giving rise to next generation by crossover and mutation.
7. Repeat from step 3 to 6.

3. RESULTS

3.1 Datasets

1. *Colon Tumor :* Contains 62 samples collected from coloncancer patients. Among them, 40 tumor biopsies are from tumors (labelled as "negative") and 22 normal (labelled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes

were selected based on the confidence in the measured expression

2. *BreastTissue :* Six classes of fleshy excised tissue were studied under electric impedance . It contains 106 samples. Among them, 21 are of carcinoma type, 15 are of fibro-adenoma type, 18 are of mastopathy, 16 are under glandular, 14 are under connective and 22 are under adipose.

3.2 Parameters

In this study, we chose the radial basis function kernel because it works well in most cases and has only two parameters.

We set the sigma=1.0, correlation threshold value=0.6, number of features to be selected (N)=2, generations=30, mutationrate=0.04, populationsize=20. The summary of he classification accuracy using GA-SVM and Corr-GA-SVM has been shown in table 1.

Table 1. The classification accuracy for HO-SVM and Corr-HO-SVM

Dataset	**GA-SVM**	**Corr-GA-SVM**
Breasttissue	83.7	85.6
Colon	85.6	88.3

The above results show that the correlation based HO-SVM works better than GA-SVM. The results are shown in graphical manner in figure 5.

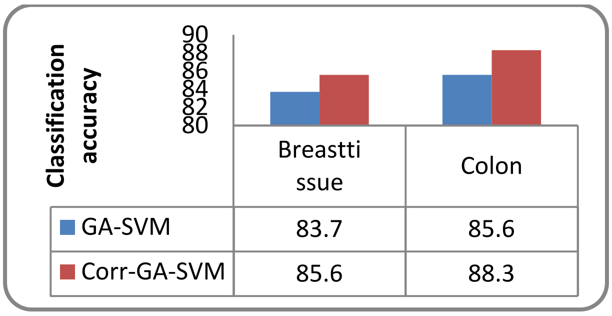

Fig.5 Comparison of GA-SVM and Corr-GA-SVM

4. CONCLUSION AND FUTURE WORK

We presented a wrapper approach for feature selection using SVM, genetic algorithm and correlation among features.

It gives better classification accuracy than the GA-SVM wrapper method which does not use correlation.

Future work can be done in various directions. It would be interesting to use the proposed wrapper technique for feature selection in combination with the variations of SVM, such as different Kernel functions and Support Vector Regression.

5. REFERENCES

[1] A. Rakotomamonjy, Variable selection using SVM-based criteria, Journal of Machine Learning Research 3 (2003) 1357–1370

[2] Yu Wanga, Igor V. Tetkoa, Mark A. Hallb, Eibe Frankb, "Gene Selection from microarray data for cancer classification-a machine learning approach", in *Proc. Computational Biology and Chemistry,* 29 (2005) 37–46

[3] Li Zhuo, Jing Zheng, Fang Wang, **A** genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine,In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing 2008*

[4] Jianping Hua, Waibhav D. Tembe, and Edward R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," , in *Proc Pattern Recognition,* 42 (2009) 409 -- 424.

[5] Iffat A.Gheyas, Leslie S.Smith, "Feature subset selection in large dimensionality domains", in *Proc. Pattern Recognition,* 43 (2010) 5 - 13