International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Combining Supervised Attribute Clustering And Ga-Svm Classifier For Microarray Sample Classification

Jayapriya J
PG Scholar
Dept. of Computer Science And Engineering,
Anna University,Regional Center
Coimbatore

D. Palanikkumar
Assistant Professor,
Dept. of Computer Science And Engineering
Anna University,Regional Center
Coimbatore

Abstract

Microarray technique is used to record the expression levels of thousands of genes simultaneously.Only a small part of these genes will be useful for performing a certain diagnosis.A supervised attribute clustering is used to find such initial groups of clusters.[1][2] One of the main tasks with the expression data is to find groups of tightly coupled genes that is more functional or meaningful in nature After forming the clusters, a GA-SVM based predictor is used to accurately classify the generated clusters.[4]With the help of Genetic Algorithm it helps optimizing the clusters and also to avoid the spurious genes that are added in the ealier stages of clustering.The supervised attribute clustering acts as an aid for the actual microarray classification. Thereby it helps to increase the classification and predictive accuracyof the GA-SVM based classification .It also retains the informative structure of the data and provides excellent predictive capability for accurate medical diagnosis.[10]

Index terms—Microarray Analysis, attribute clustering, Genetic Algorithm, SupportVector Machine, Classification.

1. Introduction

DNA microarrays also known as Gene Chips or DNA chip.They are used to measure and record the expression levels of large numbers of genes simultaneously.A microarray gene expression data set is represented by an expression table, where each row depicts one particular gene,and each column to a sample, and each entry of the matrix is the recorded expression level of a particular gene in a particular sample. Eventhough there are large amount of genes in a any dataset,only a small number of them will be effective for performing a certain classification.Also the high dimensional feature of gene expression data is one of the major problem microarray based classification.It impose a high computational cost as well as the risk of "overfitting" at the time of classification.[7].A gene cluster is a group of one or more genes which encodes to form a same products.It helps to track the evolutionary history of an individual.In gene expression data analysis, the existing clustering methods such as bayesian clustering,hierarchical clustering, k-means algorithm,self-organizing map and principal component analysis ,group a subset of genes that are interdependent or correlated with each other.[1][11][12]. Henceforth, genes or attributes in a cluster are more correlated among themselves, whereas genes in different clusters are less correlated. The attribute clustering helps to reduce the search dimension of a classification algorithm and constructs the model using a tightly correlated subset of genes rather than using the entire genes. [1]

One of the main tasks with the gene expression data is to find groups of genes whose combined expression is strongly associated with the sample categories or response variables[2][5][6].So in the proposed method a supervised attribute clustering in combination with GA-SVM Predictor is used to find such groups of genes by incorporating the information of sample categories. A quantitative measure, based on mutual information, is used to calculate the similarity between attributes and cluster formation.[1][8][9].After forming the clusters, a GA-SVM classifier is used for the reduced feature set generation and also to evaluate the accuracy of the generated clusters.The use of Genetic Algorithm helps to remove the spurious genes that were incorrectly added to the cluster at earlier stages.

The supervised attribute clustering acts as an aid for microarray classification.It tries to cluster

genes such that the discrimination of different tissue types is as simple as possible. With the use of this classifier along with supervised attribute clustering,it is possible to enhance the performance of the overall microarray classification.By considering supervised clustering as an aid for the classification it retains the informative structure of the data and provides excellent predictive capability for accurate medical diagnosis. The proposed method reduces the dimentionality,avoids the noise sensitivity problem and increases the classification accuracy of microarray data.

2. Related Work

The microarray technology is an aid for simultaneous control of thousands of genes for each sample. The gene clustering and classification of the samples is often performed separately, or in a directional (one as an aid for the other). However, the separation of these two tasks can not include a reporting structure in the data. A model selection criterion based on new Rissanen's MDL (minimum description length) principle was developed for simultaneous clustering and classification.[3] MDL code length is given for the two explanatory variables (genes) and response variables (labels class samples). The final output of the algorithm is a sparse classification rule and interpret based on cluster centroids or the closest to the centroid genes.

Another method is based on the correlation based GA-SVM.[4]The microarray can be used to measure slight changes in the expression of many genes simultaneously. Feature selection is a way to reduce the dimensionality. Effective screening can be performed based on the correlation between attributes. Therefore a correlation algorithm based wrapper feature selection using genetic algorithm (GA) and Support Vector Machines with kernel functions are used for classification. The basic idea of GA-SVM method is to remove those features which are less suitable. The features that have high fitness value high classification accuracy and are preserved during evolution

3. Proposed Work

Most of the techniques used for microarray classification deals with either enhancing the performance of clustering method or enhancing the accuracy of the classifier.In the proposed method the performance enhancement focus on both the clustering and classifier so as to improve the overall microarray classification.The Existing supervised attribute clustering algorithm is used to find coregulated clusters of genes whose combined expression is strongly associated with the sample categories or class labels.The similarity between the attributes is computed by using a quantitative measure based on mutual information.This measure incorporates the information of sample categories while measuring the similarity between attributes or genes.

Henceforth, it helps to identify functional groups of genes that are of special interest in sample classification and discrimination of sample categories. The supervised attribute clustering method uses this measure to reduce the redundancy among genes.It includes partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are tightly coupled with strong association to the sample categories. After forming the clusters, a GA-SVM classifier is used to evaluate the accuracy of the generated clusters and feature subset selection.The supervised attribute clustering acts as an aid for microarray classification.Thereby it helps to increase the classification and predictive accuracy of the correlation based GA-SVM

The proposed method reduces the dimentionality,avoids the noise sensitivity problem and increases the classification accuracy of microarray data and more number of infected cells can be found out that are unable to find out using SVM.Also it helps for early disease identification.The proposed system deals with different operations on the microarray data such as preprocessing,attribute clustering ,GA-SVM based classification and the performance evaluation with the existing system. **T**he hierarchical structure of the propsed system is given below.

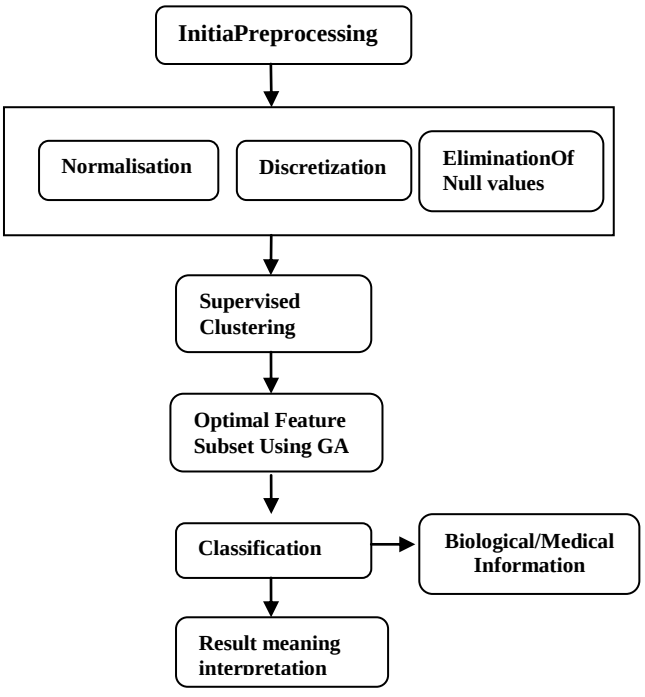

Fig.1 Structure of the Proposed System

3.1. Preprocessing

Todays real world databases are highly susceptible to noisy,missing, inconsistent data due to their typical huge size.Low quality data will result in low quality mining results.Therfore Preprocessing is one of the important task in Data mining.There are a number of data preprocessing techniques.Data in the real world is dirty, Incomplete data due to lacking values of an attribute, lacking certain attributes that are of special interest, or containing only aggregate data. Inconsistent data may come from Different data sources. Preprocessing is an indispensible part of Microarray classification as dimensionality reduction is one of its major challenge.The following are the some of the methods performed prior to gene clustering.Identifying the missing value,Log operations,Normalization,Discretization by Threshold or Percentage.Missing values are replaced by the attribute mean.Normalization is performed inorder to scale the values to fall within a specific range. z-score normalization is calculated by the following formula.

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Where μ_A is the mean and σ_A is the standard deviation.Discretization can be performed either using a Threshold value or by using a percentage. It divides the range of a continuous attribute into intervals. Data size can be reduced by discretization .Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.The label of these intervals can then be used to replace actual data values.

3.2. Supervised Attribute Clustering

The similarity measure used for clustering incorporates the information of the tissue categories so that the cluster results in more functional genes.The supervised attribute clustering algorithm relies on mainly two different factors they are,determining the relevance of each attribute and forming the cluster around each relevant attribute incrementally by adding one attribute after the other. The growth of a cluster is repeated until the cluster stabilizes, and then the clustering algorithm starts to generate a new cluster. The relevance uses information about the class labels and is thus a criterion for supervised clustering . The supervised clustering incorporates the information regarding the tissue and hence it helps to identify more functional gene clusters .

3.3. GA-SVM Based Classification

1) Genetic Algorithm

A genetic algorithm (GA) is a general optimization search methodology based on a direct analogy to Darwinian natural selection theory and genetics in biological sciences. It has been proved to be an important alternative to traditional heuristic methods on the basis of the Darwinian principle of "survival of the fittest", GA strats with a set of candidate items called a population and obtains the optimal solution after a series of iterative computations.GA evaluates fitness of each individual item, which represents the quality of the solution, through a fitness function.The crossover and mutation functions are the mainly used operators that randomly transform the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable and favourable results are obtained.On behalf of the characteristics of exploitation and exploration search, GA can deal with large set of genes efficiently, and henceforth it has minimum chance to get local optimal solution than other algorithms.

2). Support Vector Machine

Support Vector Machines (SVM) is a classification system derived from statistical learning theory.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; 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 usually called the optimal hyperplane, and the data points that are closer to the hyperplane are called support vectors.These support vectors are the critical and most important elements of the training set.The SVM can be converted to become a nonlinear classifier through the use of nonlinear kernels. There are different types of SVM classifier kernel functions such as linear, radial basis function (RBF), sigmoid and polynomial.

The basic idea of the GA-SVM method is to remove the features which are less fit. The features that have high fitness value and high classification accuracy are retained for the evolution.This is achieved by an iterative algorithm. After forming the clusters, a GA-SVM classifier is used for the reduced feature set generation and also to evaluate the accuracy of the generated clusters .The use of Genetic Algorithm helps to remove the spurious genes that were incorrectly added to the cluster at earlier stages and more number of infected cells can be found out.It tries to cluster genes such that the discrimination of different tissue types is as simple as possible.

4. Proposed Algorithm

1. Perform the initial preprocessing such as removing null values,normalization, descretization etc.

2. Calculate the supervised similarity measure.

3.Generate the clusters using the supervised attribute clustering algorithm .

4.Repeat the steps 5 to 7 for each clusters.

4. Run the Genetic Algorithm for each clusters and calculate the fitness value of each individual in the cluster. 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 and remove the spurious genes that are incorrectly added in the earlier stages of clustering.

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. Run SVM for classification to predict whether a particular gene is infected or not.

5. Results And Discussion

The proposed method is implemented in Java in NetBeans Environment.It deals with performance enhancement by increasing the accuracy of both the clustering and classification of microarray data.It can identify the better cluster patterns from an input gene expression dataset that resembles more to the sample categories.This clusters can be used for input to the GA-SVM classifier which will result in an improved classification and predictive accuracy.The genes in a cluster are more similar in their expression patterns.The expression profiles of the genes in a sample cluster is shown below.

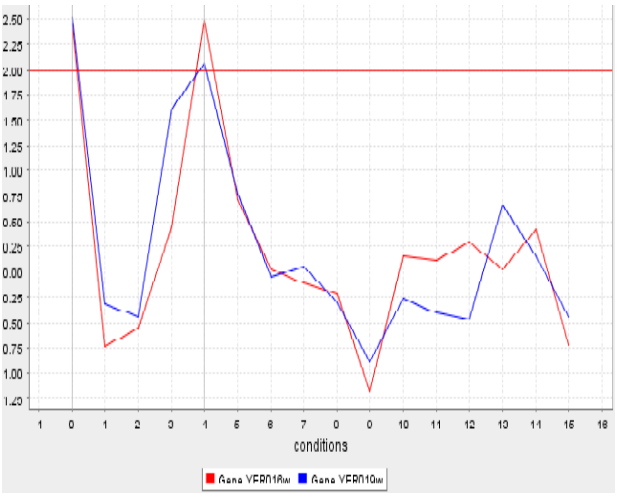

Fig .2 Expresssion Profiles Of a Genes In a Cluster

With the use of GA-SVM based predictor we are able to achieve more accurate results and hence we can identify more no of infected cells that are not identified using SVM.The comparison of the results are shown below.

	SVM Results	GA-SVM Results
YNL289w		
YPL209c	YNL289w -> infected cell	YNL289w -> infected cell
YJR043c	YPL209c -> infected cell	YPL209c -> infected cell
YKR013w	YJR043c -> infected cell	YJR043c -> infected cell
YDR013w	YKR013w -> normal cell	YKR013w -> infected cell
YPL208w	YDR013w -> normal cell	YDR013w -> infected cell
YNL300w	YPL208w -> normal cell	YPL208w -> normal cell
YER019w	YNL300w -> normal cell	YNL300w -> normal cell
	YER019w -> normal cell	YER019w -> normal cell
	SVM Accuracy = 34.818850199566164%	GA-SVM Accuracy = 69.466760176789 72%

TABLE 1
RESULT SUMMARISATION

Data Sets	Supervised clustering with SVM Predictor	Supervised Clustering with GA-SVM Predictor
Breast Cancer	79.98%	87.76%
Colon Cancer	81.025	92.59%

Fig:4 Comparison of SVM and GA-SVM Predictor

6. Conclusion And Future Work

The proposed method can identify the better cluster patterns from an input gene expression dataset that resembles more to the sample categories.This clusters are used for input to the GA-SVM classifier which will result in an improved classification and predictive accuracy of each cell. More number of infected cells can be found out using the GA-SVM based predictor. The future works can be extended at handling the classification of gene expression dataset with improved classification performance by using different kernel functions of SVM.

References

1. Pradipta Maji (2012),"Mutual Information Based Supervised Attribute Clustering For Microarray Sample Classification". *IEEE Transactions On Knowledge And Data Engineering*,Vol 24,NO 1,January 201

2. Marcel Dettling and Peter Biihlmann , "Supervised Clustering of Genes**",***Genom Biology 2002,3(12):research*0069.1-0069.15
3. Rebecka Jornsten and Bin Yu,"SimultaneousGene Clustering and Subset Selection for Sample Classification via MDL *Bioinformatics* 19(9)research (2002)
4. BinitaKumari,."FeatureSubsetSelectioninLargeDimensionalityUsing Correlation based GA-SVM," *IJCS Volume No.6* May 2012
5. **P**radipta Maji (2011) **"**Fuzzy-Rough Supervised Attribute Clustering Algorithm And Classification Of Microarray Data**" ,** *IEEE Transactions On Knowledge And Data Engineering*,Vol 24,NO 1,January
6. Mihajlo_Grbovic_,NemanjaDjuric_SlobodanVucetic (2011)_"Supervised Clustering of Label Ranking Data ."
7. B.Hari Babu1, N.Subash Chandra2 & T. Venu Gopal."Clustering Algorithms For High Dimensional Data – A Survey Of Issues And Existing Approaches".
8. Zheng Yun and Kwoh Chee Keong, April 2011 "A Feature Subset Selection Method Based On High-Dimensional Mutual Information",*Transactions on Entropy,2011.*
9. Feifei X, Jingsheng LEI, Lai WEI, "Improved Mutual Information-basedGene Selection with Fuzzy Rough Sets"**,** *Journal of Computational Information Systems,2011.*
10. Khalidraza,**(**2007**),** "Application Of Data Mining In Bioinformatics**",** *Indian Journal of Computer Science and Engineering* Vol 1 No 2, 114-118
11. Venkatadri.M, Dr. Lokanatha C. Reddy, *February 2011,* **"**A Review on Data mining from Past to the Future"**,** *International Journal of Computer Applications*
12. Jacinth Salome J, R M Suresh, PhD,June (2012) **"**Efficient Clustering for Gene Expression Data"**,** *International Journal of Computer Applications*,2012.
13. E. Domany,(2003) "Cluster Analysis of Gene Expression Data,"*J. Statistical Physics*, vol. 110, nos. 3-6, pp. 1117-1139, 2003.s.