International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 6, November-December 2014 **ISSN 2278-6856**

Classification of Microarray Samples using Attribute Clustering

K.R.SARANYAA[1], N.SUNDARAM [2]

[1] M.E. Student, Computer Science And Engineering, Mahabharathi Engineering College.

[2] Assistant Professor, Computer Science And Engineering, Mahabharathi Engineering College.

Abstract

Clustering is a useful technique for the analysis of gene expression data. Clustering algorithms based on probability model offers an alternative to heuristic algorithms. In particular, model-based clustering assumes that the data is generated by finite mixture of underlying probability distributions such as multivariate normal distributions. The issues of selecting a 'good' clustering method and determining the 'correct' number of clusters are reduced to model selection problems in the probability framework. This paper presents an attribute clustering method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data. It can be used for gene grouping, and classification. By clustering attributes, the search dimension of a data mining algorithm is reduced. The reduction of search dimension is especially important to data mining in gene expression data because such data typically consist of a huge number of genes (attributes) and a small number of gene expressions. It is for the aforementioned reasons that gene grouping and selection are important preprocessing steps for many data mining algorithms to be effective when applied to gene expression data. This project defines the problem of attribute clustering and introduces a methodology to solving it. Our proposed method group's interdependent attributes into clusters by optimizing a criterion function derived from an information measure that reflects the interdependence between attributes. By applying our algorithm to gene expression data, meaningful clusters of genes are discovered. The grouping of genes based on attribute interdependence within group helps to capture different aspects of gene association patterns in each group. Significant genes selected from each group then contain useful information for gene expression classification and identification.

Keywords :- Supervised clustering, Data mining, Gene expression ,attributes

1. INTRODUCTION

Clustering is one of important topic in data mining research work. In a relational table, a conventional clustering algorithm group's tuples, each of which is characterized by a set of attributes, into clusters based on similarity. Intuitively, tuples in a cluster are more similar to each other than those belonging to different clusters. This shows that clustering is very useful in many data mining applications .When clustering approach is applied to gene expression data analysis, conventional clustering algorithms often encounter the problem related to the nature of gene expression data which is normally wide and shallow. Data sets usually contain a huge number of genes called as attributes and a small number of gene expression profiles called as tuples. This characteristic of gene expression data often compromises the performance of conventional clustering algorithms. In this paper, we present a methodology to group attributes that are interdependent / correlated with each other. We refer such kind of process as attribute clustering. In this approach, attributes in a cluster are more correlated with each other whereas attributes in different clusters are less correlated. Attribute clustering is able to reduce the search dimension of a data mining algorithm to effectuate the search of interesting relationships or for construction of models in a tightly correlated subset of attributes rather than in the entire attribute space. After attributes are clustered, one can select a smaller number for further analysis

1.1. Correlated Gene Distance Measure

Many data mining techniques are based on similarity measure among objects. There are essentially two ways to measure the similarity of the objects; fist they can be obtained directly from the objects, for example a marketing survey may ask respondents to rate pairs of objects according to their similarity. Alternatively, measuring the similarities may be obtained indirectly from vector of measurements or characteristics describing each object. We have a formal definition of either "similar" or "dissimilar".So we can easily define the other by applying a suitable monotonically decreasing transformation for example if s(i,j) denotes the similarity and d(i,j) denotes the dissimilarity between objects i & j.

$$D(i,j) = 1 - S(i,j) \quad (1)$$

$$D(i,j) = \sqrt{2(1 - S(i,j)} \quad (2)$$

The proximity is often used as a general term to denote to measure either similarity or dissimilarity. Two additional terms, distance and metric are often used in this context. The term distance is often used informally to refer to a dissimilarity measure derived from the characterizing and describing the object as Euclidean distance, defined below

$$d_E(i,j) = \left[\sum_{K=1}^{P} \left| xk(i) - xk(j) \right|^2 \right]^{1/2} \quad (3)$$

We have n data objects with p real value measurements on each object. We denote the vector of observation for the ith object by x(i)=x1(i),x2(i),… xp(i); 1<=i<=n; where the value of the kth variable for the ith object is nk(i). Metric on the other hand is a dissimilarity measure that satisfies three conditions d(i,j)>=0 for all the i and j , d(i,j)=0; if and only if i=j; d(i,j)=d(j,i) for all i & j d(i,j)<=d(i,k)+d(k,j) for all i, j and k. These three conditions are called as triangle inequality.

1.2. Techniques Applied- Supervised Clustering

This paper focus on microarray data where experiments monitor gene expression in different tissues and where each experiment is equipped with an additional response variable such as a cancer type and brain tumor etc. Hierarchical Clustering identifies sets of correlated genes with similar behavior across the experiments, but yields thousands of clusters in a tree-like structure. This makes the identification of functional groups very difficult. In contrast, Self-Organizing-Maps require a pre-specified number and an initial spatial structure of clusters, but this may be hard to come up with in real problems. These drawbacks were improved by a novel graph theoretical clustering algorithm, but as all other unsupervised techniques, it usually fails to reveal functional groups of genes that are of special interest in tissue classification. This is because genes are clustered by similarity only, without using any information about the experiment's response variables. This paper present a promising new method for searching functional groups, each made up of only a few genes whose consensus expression profiles provides useful information for tissue discrimination.

2. Related Work

2.1. Feature Selection Based On Mutual Information: Criteria of Max-Dependency, Max-Relevancy and Min-Redundancy

Large gene expression studies, such as using DNA arrays, It provide millions of different pieces of gene expression data. To address the problem of analyzing such data, then describe a statistical method, which they have called 'gene shaving'. This method identifies subsets of genes with coherent expression patterns and large variation across conditions. Gene shaving method which differs from hierarchical clustering and other widely used methods for analyzing gene expression studies in that genes may belong to one or more cluster. The technique can be 'unsupervised', that is, the genes and samples or cells are treated as unlabeled, or fully supervised by using known properties of the genes or samples to assist in finding meaningful groupings.

2.2 Using Mutual Information for Selecting Feature in Supervised Neural Net Learning

Large gene expression studies, such as using DNA arrays, provide millions of different pieces of gene expression data. To address the problem of analyzing such data, then describe a statistical method, which they have called 'gene shaving'. This method identifies subsets of genes with coherent expression patterns and large variation across conditions. Gene shaving method which differs from hierarchical clustering and other widely used methods for analyzing gene expression studies in that genes may belong to one or more cluster. The technique can be 'unsupervised', that is, the genes and samples or cells are treated as unlabeled, or fully supervised by using known properties of the genes or samples to assist in finding meaningful groupings.

3. System Design

The uploaded training dataset converted into the micro array data. Data preprocessing does cleansing, normalization, transformation, feature extraction and selection. Identify the two types of gene selection such as occurrence based selection and sequence based selection. In occurrence based selection, we provide the separate gene and show all gene which is provided by users. Then in sequence based selection, we provide sequence and identify all sequences with position information. Three types of periodic patterns are present in time series they are symbol, sequence and segment periodic.

3.1 Load the Dataset

We upload the datasets. The dataset may be microarray dataset. A microarray database is containing microarray gene expression data. The user uses the microarray database to store the measurement of data, and also to manage the searchable index, and make the data available to other applications for analysis and interpretation.

3.2 Preprocessing

Data pre-processing is an important step in the data mining process. The representation and quality of data is first checked before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery and maintaining the data during the training phase is more difficult.

3.3 Pattern Evaluation

In this module we identify the two types of gene selection such as occurrence based selection and sequence based selection. In occurrence based selection, we provide the separate gene and show all gene which is provided by users. In sequence based selection, we provide sequence and identify all sequences with position information.

3.4 Clustering approach

The proposed supervised attribute clustering algorithm relies on mainly two factors, namely, determining the relevance of each attribute and growing the cluster around each relevant attribute incrementally by adding one attribute after the other. A new supervised attribute clustering algorithm is proposed to find co regulated clusters of genes whose collective expression is strongly associated with the sample categories or class labels. A new quantitative measure, based on mutual information, is introduced to compute the similarity between attributes. The proposed supervised attribute clustering method uses this measure to reduce the redundancy among genes.

3.4 Coherent Index Selection

We can identify the gene positions. Then finding good clustering configurations which contain interdependence information within clusters and discriminative information for classification; 2) selecting from each cluster significant genes with high multiple interdependence with other genes within each cluster; and 3) yielding very high classification results on both of gene expression datasets using a small pool of genes selected from the clusters found by as the training set.

3.5 Evaluation criteria

In this module, the performance of the proposed supervised attribute clustering algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on five microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separately index and classification accuracy of naive baye's classifier, K-nearest neighbor rule, and support vector machine. To compute the classification accuracy, the leave-one-out cross validation is performed on each gene expression data set.

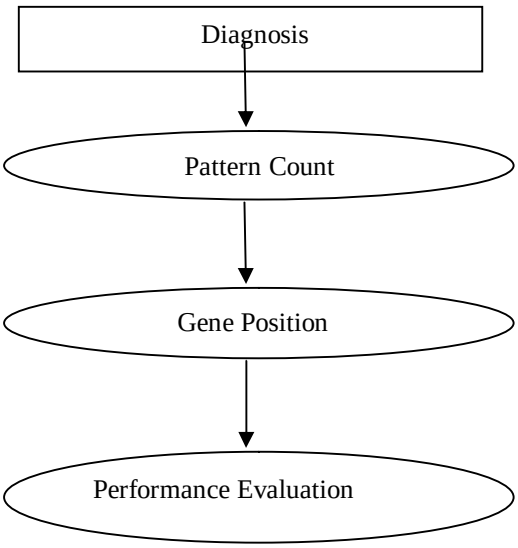

Figure 1 Pattern evaluation flow

4. Evaluation Result:

We evaluate the results by comparing the various techniques and the comparison can be shown as in the chart.

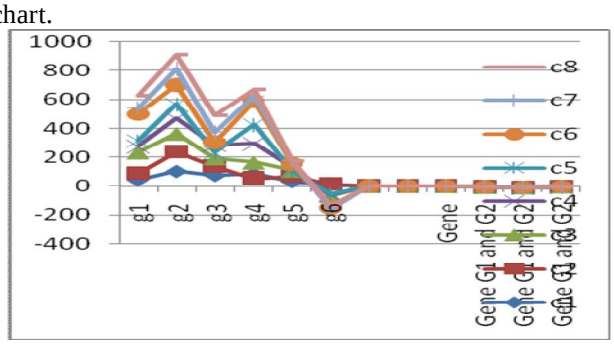

Figure 2 Evaluation chart

5. Conclusion

Modern Microarray technologies monitor transcription levels of tens and thousands of genes in concurrently. This approach reviewed both classical and recently developed clustering algorithms, which have been applied to gene expression data. The proposed supervised attribute clustering algorithm is based on measuring similarities between attributes using the new quantitative measure, whereby redundancy among the attributes is almost removed. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

References

[1] M. Ankerst, M.M. Breunig, H.P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure," Proc. SIGMOD, pp. 49-60.

[2] Z. Bar-Joseph, E.D. Demaine, D.K. Gifford, N. Srebro, A.M. Hamel, and T.S. Jaakkola, "K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data," Bioinformatics, vol. 19, no. 9, pp. 1070- 1078, 2003.

[3] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," J. Computational Biology, vol. 6, nos. 3-4, pp. 281-297.

[4] M. Blatt, S. Wiseman, and E. Domany, "Super-Paramagnetic Clustering of Data," Physical Rev. Letters, vol. 76, 1996

[5] Y. Cheng and G.M. Church, "Biclustering of Expression Data," Proc. Eighth Int'lConf. Intelligent Systems for Molecular Biology (ISMB), vol. 8, pp. 93-103, 2000.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Nat'l Academy of Sciences USA, vol. 95, no. 25, pp. 14863-14868.

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. Second Int'l Conf. Knowledge Discovery and Data Mining, pp. 226-231.

[8] Pradipta Maji,"Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification," Science IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012