ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Semi Supervised Approach for Microarray Data Analysis

Priya.V[1], Shanmuga priya.S[2]

Student, Computer Science and Engineering, St.Joseph's College of Engineering and Technology, Thanjavur, India[1]

Asst.Prof., Computer Science and Engineering, St.Joseph's College of Engineering and Technology,Thanjavur,India[2]

Abstract: Clustering is a useful exploratory technique for the analysis of gene expression data. In particular, model-based clustering considers that the data is generated by a 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. Using clustering attributes, the search dimension of a data mining algorithm is dense. 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 OFS algorithm to gene expression data, important clusters of genes are exposed. The grouping of genes based on feature interdependence within group helps to capture different aspects of gene association patterns in each group. Important genes selected from each group then contain useful information for gene expression classification and identification.

Keywords: Feature Selection, Online Learning, Large-scale Data Mining, Classification.

INTRODUCTION

Clustering is an important topic in data mining research. Given 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. It has been shown that clustering is very useful in many data mining applications .When 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." 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 or correlated with each other. We refer to such a process as attribute clustering. In this sense, 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

- Data Gene expression data is obtained by extraction of quantitative information from the images/patterns resulting from the readout of fluorescent or radioactive hybridizations in an microarray chip. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Each element of this matrix represents the expression level of a gene under a specific condition, and is represented by a real number, which is usually the logarithm of the relative abundance of the mRNA of the gene under the specific condition.

- Gene expression matrices have been extensively analyzed in two dimensions: the gene dimension and the condition dimension. These analysis correspond, respectively, to analyze the expression patterns of genes by comparing the rows in the matrix, and to analyze the expression patterns of samples by comparing the columns in the matrix. A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a time-series during a biological process (e.g., the yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues).

In this paper, we will focus on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes". Similarly, we will uniformly refer to all kinds of experimental conditions as "samples" if no confusion will be caused. A gene expression data set from

- a microarray experiment can be represented by a real-valued **expression matrix** $M = \{w_{i,j} | 1 \leq i \leq n, 1 \leq j \leq m$ (Figure 1), where the rows form the expression patterns of genes, the columns represent the expression profiles of examples, and each cell is the measured expression level of gene i in sample j. Figure 1 (b) includes some notation that will be used in the following sections.

DOI 10.17148/IJARCCE.2015.4242

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

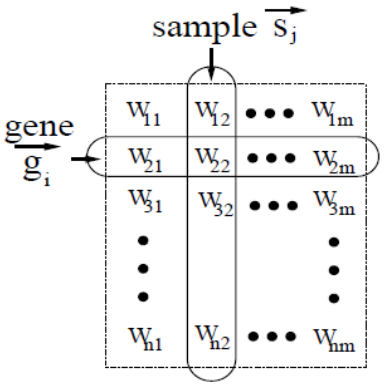

Figure 1 Gene Expression format

The original gene expression matrix obtained from a scanning process contains noise, missing values, and systematic variations arising from the experimental procedure. Data pre-processing is indispensable before any cluster analysis can be performed. Some problems of data pre-processing have themselves become interesting research topics. Those questions are beyond the scope of this survey; an examination of the problem of missing value estimation appears in, and the problem of data normalization is addressed.clustering approaches apply one or more of the following pre-processing procedures: filtering out genes with expression levels which do not change significantly across samples; performing a logarithmic transformation of each expression level; or standardizing each row of the gene expression matrix with a mean of zero and a variance of one. In the following discussion of clustering algorithms, we will set aside the details of pre-processing procedures and assume that the input data set has already been properly pre-processed.

RELATED WORK

The paper defines the problem of attribute clustering and introduces a methodology to solving it. The proposed method groups 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. To evaluate the performance of the proposed approach, then applied it to two well-known gene expression datasets and compared the results with those obtained by other methods. Experiments show that the proposed method is able to find the meaningful clusters of genes. By selecting a subset of genes which have high multiple-interdependence with others within clusters, significant classification information can be obtained. Thus a small pool of selected genes can be used to build classifiers with very high classification rate. From the pool, gene expressions of different categories can be identified.

Wolfgang Huber, Anja von Heydebreck, Martin Vingron et al [2] This article reviews the methods utilized in processing and analysis of gene expression data generated using DNA microarrays. This type of experiment allows determining relative levels of mRNA abundance in a set of tissues or cell populations for thousands of genes simultaneously. Naturally, such an experiment requires computational and statistical analysis techniques. At the outset of the processing pipeline, the computational procedures are largely determined by the technology and experimental setup that are used. Subsequently, as more reliable intensity values for genes emerge, pattern discovery methods come into play. The most striking peculiarity of this kind of data is that one usually obtains measurements for thousands of genes for only a much smaller number of conditions. This is at the root of several of the statistical questions discussed here.

n	number of genes
m	number of samples
M	a gene expression matrix
w_{ij}	each cell in a gene expression matrix
$\vec{g}_i$	a gene
$\vec{s}_j$	a sample
$G, G', G_0, \ldots$	a set of genes
$S, S', S_0, \ldots$	a set of samples

Table 1 Main Notations

III.FRAMEWORK FOR OFS

We now turn our attention to developing formalism and framework for online feature selection.

A. Regularized Risk Minimization

In recent years, a lot of attention has been given to the idea that certain forms of regularization may be used as an alternative to feature subset selection. This provides the foundation of our incremental approach. To develop the argument, we begin by considering the problem of deriving a good mapping, given a full set of features, as one of regularized risk minimization. That is, the criterion to be optimized, C, takes the form:

$$C = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f_i) + \Omega(f)$$

Where L () is a loss function, and Ω(f) is a regularization term that penalizes complex mapping functions. We have used f_i as shorthand for f (x_i).

B. Loss Functions

Different loss functions are appropriate for different types of learning problem. In this paper we will deal with binary classification problems, with y taking values of ±1, and so a suitable loss function is the binomial negative log-likelihood, used in logistic regression.

$$L_{bnll} = \ln(1 + e^{-yf(x)})$$

DOI 10.17148/IJARCCE.2015.4242

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

The BNLL loss function has several attractive properties. It is derived from a model that treats f(x) as the log of the ratio of the probability that y = +1 to the probability that y = −1, which allows us to calculate1 p(y = +1 | x) using the following relation:

$$p(y = \pm 1|X) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

The loss function is also convex in f(x), which has positive implications for finding a global optimum of C. Finally, it only linearly penalizes extreme outliers, which is important for robustness. We denote the mean loss over all training points as L_{bnll}. Most of what follows in this paper applies to other commonly used loss functions as well, and we indicate this by dropping the BNLL subscript, except where we need to be specific. A regression task, for instance, would be more likely to employ a sum of squared errors loss function.

C. Regularizes

The choice of regularizer in (1) depends upon the class of models used for f. Here, we will restrict ourselves to classes of models whose dependence on x is parameterized by a weight vector w. Linear models fall into this category, as do various kinds of multi-layer perceptions and radial basis function networks. A commonly used regularizer for these models is based on a norm of the weight vector:

$$\Omega_p(w) = \lambda \sum_{j=1}^{n} |w_j|^p$$

Where λ is a regularization coefficient, p is a nonnegative real number, and n is the length of w. This type of regularizer is the familiar Murkowski l_p norm raised to the p'th power, and so is usually called an l_p regularizer. If p = 2, then the regularizer is equivalent to that used in ridge-regression and support vector machines. If p = 1, then the regularizer is the "lasso". If p → 0 then it counts the number of non-zero elements of w. The p = 1 lasso regularizer has some interesting properties. Firstly, it is the smallest p for which Ωp is a convex function of w. This means that, if the loss function in (1) is also a convex function of weights, then optimizing C with respect to w using gradient descent is guaranteed to find the global optimum, since the sum of two convex functions is also convex. For our work, the second crucial property[2] of the l_1 regularizer is that there is a discontinuity in its gradient with respect to w_j at $w_j = 0$, which tends to force a subset of elements of w to be exactly zero at the optimum of C which is precisely what we require for a model that is sparse in features.

For these reasons we use the l_1 regularizer in our work here. Note that the model for f may have additional parameters, e.g. bias terms, which we do not include in the regularization. With the BNLL loss function and l_1 regularization, the learning optimization criterion becomes:

$$C = \frac{1}{m} \sum_{i=1}^{m} \ln\left(1 + e^{-y_i f(x_i)}\right) + \lambda \sum_{j=1}^{n} |w_j|$$

D. Normalization

The Ωp regularizer penalizes all weights in the model uniformly. This only makes sense if all the features used as input to the model have a similar scale, which can be achieved by normalizing all features as they arrive. A convenient and efficient normalization process is to linearly rescale each feature so that the mean of each feature (over all training data) is zero, and the standard deviation is one, i.e. we rescale incoming feature values x_j to normalized feature values x_j', using the relation:

$$x_j' = \frac{x_j - \bar{x_j}}{\sigma_{x_j}}$$

Where x_j is the mean raw feature value, and σ_{xj} is the standard deviation. It is obviously necessary to use the same rescaling when applying the learned model to new unseen data.

IV.PREPROCESSING

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc.

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

V.PATTERN EVALUATION

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.

VI. 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 measure incorporates the information of

DOI 10.17148/IJARCCE.2015.4242

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

sample categories while measuring the similarity between attributes. In effect, it helps to identify functional groups of genes that are of special interest in sample classification. The proposed supervised attribute clustering method uses this measure to reduce the redundancy among genes.

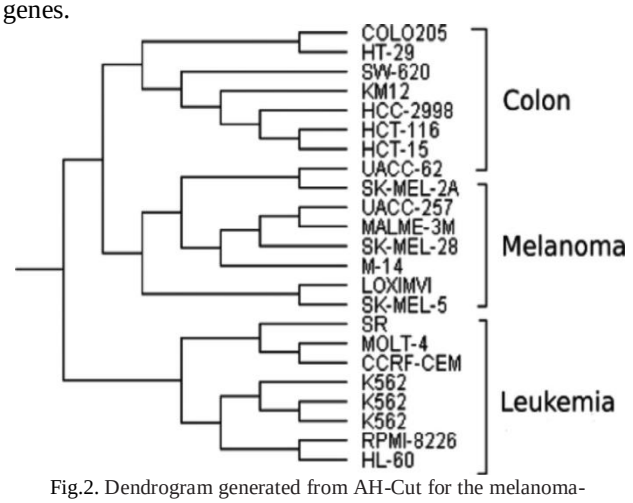

Fig.2. Dendrogram generated from AH-Cut for the melanoma-colonleukemia data set

It involves partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are highly co regulated with strong association to the sample categories while those in different clusters are as dissimilar as possible.

VII. COHERENT INDEX SELECTION

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

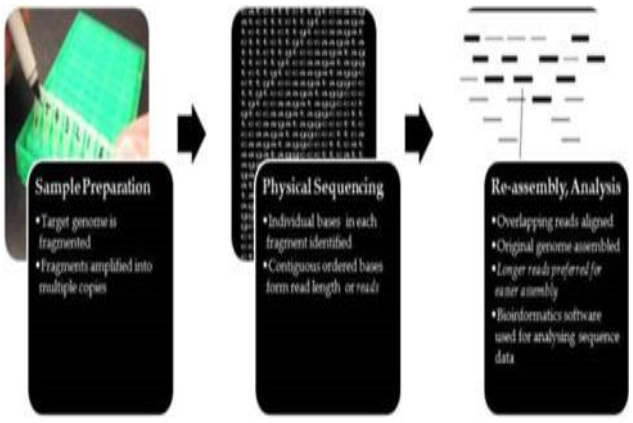

Fig.3. DNA Sequencing

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.

VIII.EVALUATION CRITERIA

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 separability index and classification accuracy of naive bayes 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

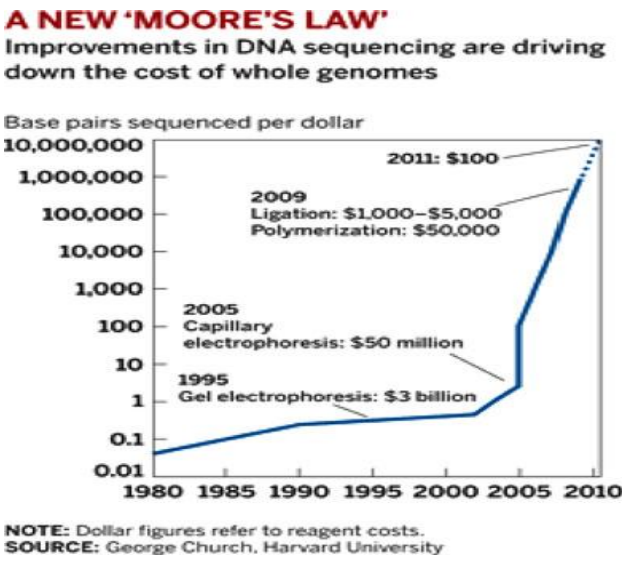

Fig.4.Graph of plummeting technological cost

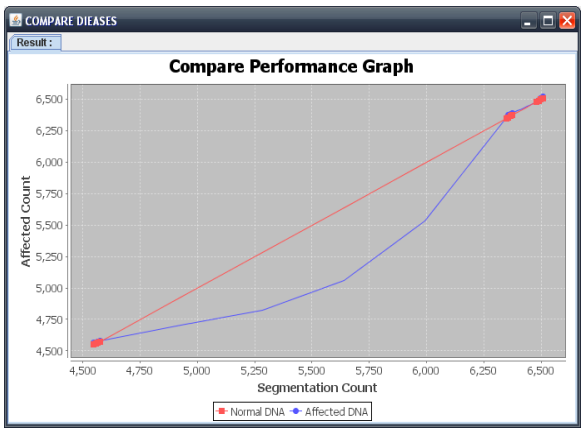

Fig.5.

DOI 10.17148/IJARCCE.2015.4242

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

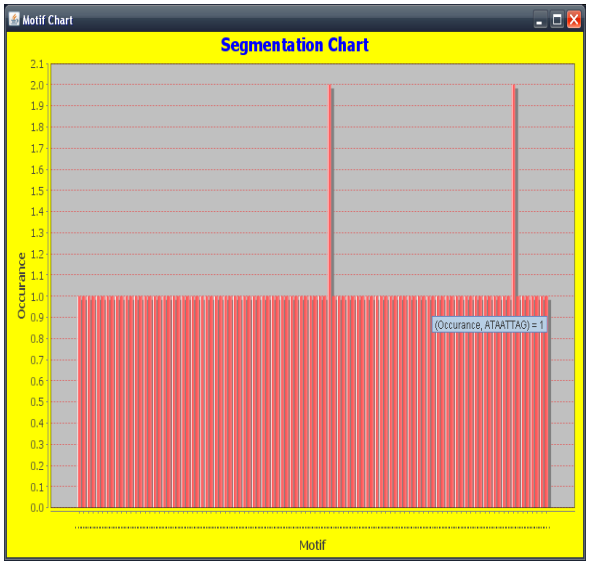

Fig.6.

IX. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a new approach Online Feature Selection that helps to select a small and permanent number of features for binary classification in online learning fashion. There are two different OFS task in different setting (i) OFS by learning with full inputs of all the dimensions /attributes, and (ii) OFS by learning with partial inputs of the attributes. To solve the OFS tasks using the OFS algorithms and offered theoretical analysis on the mistake bounds. We comprehensively examined their empirical presentation and applied the proposed techniques to solve two real-world applications: image classification and microarray gene expression analysis. The results show the proposed algorithm is effective for the feature selection task for online applications.

Future work could extend our framework to other settings, e.g., online multi-class classification and regression problems, or to help tackle other emerging online learning tasks, such as online transfer learning or online AUC maximization.

REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, no. 5439, pp. 531-537, 1999.

[2] E. Domany, "Cluster Analysis of Gene Expression Data," J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.

[3] J.G. Liao and K.-V. Chin, "Logistic Regression for Disease Classification Using Microarray Data: Model Selection in a Large p and Small n Case," Bioinformatics, vol. 23, no. 15, pp. 1945-1951, 2007.

[4] L. Wang, F. Chu, and W. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 1, pp. 40-53, Jan.-Mar. 2007.

[5] P.A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach. Prentice Hall, 1982.

[6] D. Koller and M. Sahami, "Toward Optimal Feature Selection," Proc. Int'l Conf. Machine Learning, pp. 284-292. 1996.

[7] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997.

[8] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification and Scene Analysis. John Wiley and Sons, 1999.

[10] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004

BIOGRAPHY

Priya.V received her B.Tech degree in InformationTechnology in Dhanalakshmi Srinivasan Engineering College ,Perambalur in 2006, Tamilnadu, India. Now she is doing her Master in Engineering in St.Joseph's college of engineering and technology, Thanjavur, Tamilnadu, India. She has attended 1 national conferences and 1 international Journal. She is interested in DataMining.