ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Mining Gene Expression Data in a Distributed Manner for Cancer Therapeutics

Cisy Soman, Abdul Ali

M Tech Scholar, Dept. of C.S, ICET, M.G University, Kottayam, India

Assistant Professor, Dept. of C.S., ICET, M.G University, Kottayam, India

ABSTRACT: Gene expression mining has been effectively used for classification and diagnosis of cancer. This paper highlights traditional approaches as well as current advancements in the analysis of the gene expression data from cancer perspective. Analysis of such data is important as it leads to knowledge discovery. However, mining data in a centralized system can cause a difficulty in execution. So we are proposing a distributed approach which can improve the performance by reducing the overhead. The distributed approach exploits parallel computation by splitting the entire process among a number of systems.

KEYWORDS: Data mining, gene expression data, clustering, association rules, cancer

I. INTRODUCTION

Cancer [1] is an abnormal and uncontrollable growth of cells in the body that turn malignant. Cancer is a major cause of the natural mortalities throughout the world. Cancer can develop in almost any organ or tissue, such as the lung, colon, breast, skin, bones, or nerve tissue.

Cancer classification is an important problem for both clinical treatment and biomedical research. Accurate diagnosis of cancer types can enhance efficacy and reduce toxicity of medical treatment for cancer patients. In the past, cancer classification has been relying on subjective judgment from professional pathologists. Currently, microarray experiments can be employed to screen gene expression levels from normal and cancer tissue samples. The comparisons of microarray results between normal and cancer cells can provide the important information of cancer diagnosis and treatment [7]. However, Microarray experiments provide enormous amount of data that require application of advanced computational methods such as data mining techniques, to discover the useful information and knowledge. The remarkable feature of gene expression microarray data for cancer classification is the number of variables (genes) far exceeding the number of samples in such a high dimensional space work is extremely difficult and traditional statistical methodologies in classification and prediction do not work well.

This paper proposes a distributed approach for mining the gene expression data. We emphasize that in data mining the goal is to discover knowledge that is not only accurate, but also comprehensible for the user. Comprehensibility is important whenever discovered knowledge will be used for supporting a decision made by a human user. After all, if discovered knowledge is not comprehensible for the user, he/she will not be able to interpret and validate it. In this case, probably the user will not have sufficient trust in the discovered knowledge to use it for decision making. This can lead to incorrect decisions. There are several data mining tasks, including classification, regression, clustering, dependence modeling, etc. Each of these tasks can be regarded as a kind of problem to be solved by a data mining algorithm.

For the discovery of the pattern we will be using association analysis. Association analysis, it came into prominence by the help of barcode technology which resulted construction of transactional databases in markets. Later it was thought that it would be beneficial to find frequently purchased items in the markets in order to increase sales. In [6], association rule mining was introduced as a new data mining technique which could be used for market basket analysis. Association rule mining, searches for items frequently purchased together when the market domain is considered. Mining frequent patterns is used in many applications, however, it is not that much applied to bioinformatics. Microarray is an effective technique used in molecular biology to analyze gene expression in different conditions such as control versus drug treatment or healthy versus disease conditions in many organisms. Huge amount of data has been generated already and it is difficult to handle such large amount of data for analysis. Mining frequent patterns in gene expression data seems applicable and may help bioinformatics researchers. In the gene expression analysis context, we are looking for which genes are frequently expressed together in different experiment conditions. This will lead us to knowledge of genes that are expressed association in defined conditions hereby this will facilitate the molecular biologists to define components

DOI: 10.15680/IJIRCCE.2015. 0310004

of biological pathways with further studies. Data mining will have set of data to express patterns in a way which can be used for intelligent decision making.

II. DNA MICROARRAY ANALYSIS

DNA microarrays [2] are one of the fastest growing technologies for genetic research. DNA microarrays are used to investigate cancer, for measuring changes in gene expression and learning how cells are respond to a disease or to a particular treatment. Even if microarrays represent a powerful source of biological information, using gene expression data to classify diseases on a molecular level for clinical diagnostic remains a challenging research problem.

Classifying microarray data poses several challenges to typical machine learning methods. The major aspects of the classifier design: the classification rule, the error estimation, and the feature selection. One of the main problems of traditional machine learning techniques concerns the ability of properly detecting false positives, i.e., samples erroneously assigned to a class even if they do not belong to the class library used to train the classifier. This misbehavior is clearly unacceptable since it would very likely lead to a misdiagnosis.

This micro array analysis model is very flexible, and it makes the implementation of classification, clustering. The classifier is not only able to correctly classify samples in thecorresponding classes, but it is also able to correctly detect out-of-class samples, thus drastically reducing the false positive rate. cDNA microarrays models provided very good results.

III DATA PROCESSING

Data processing is, broadly, the collection and manipulation of items of data to produce meaningful information In this sense it can be considered a subset of information processing. Data processing includes the following steps: pre-processing, post-processing, clustering,classification, association rule mining.

3.1. Pre-Processing Phase

Incomplete, noisy, and inconsistent data are commonplace properties of large real world datasets. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available. Other data may not be included simply because it was not considered important at the time of entry. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

Data pre-processing routines work to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users

believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. Therefore, a useful pre-processing step is to run your data through some data cleaning routines.

Genes, cDNA clones, or expressed sequence tags [ESTs] usually constitute the DNA sequences that are scanned by microarray experiments, conditions contingent. They may include time series data of a biological process, e.g., life cycle of a yeast cell, or a collection of varied tissue samples, e.g., normal versus cancerous tissues. For the same, a gene expression matrix is obtained, which for obvious reasons, contains gene data, notwithstanding a compendium of noise, missing values and irrelevant data. Data pre-processing is indispensable before any cluster analysis can be performed.

3.2 Post-Processing Phase

Since, the pre-processing phase aids in precipitating several groups, patterns, correlations of genes at the expression level basis, it becomes almost necessary to re-evaluate and formalize them in a phase called post-processing phase. During this phase, the domain experts analyze and match the extracted patterns to the business objectives and success criteria.

In [3] KEOPS methodology works on comparing extracted data with expert’s knowledge, as elaborated in Fig. 1

.

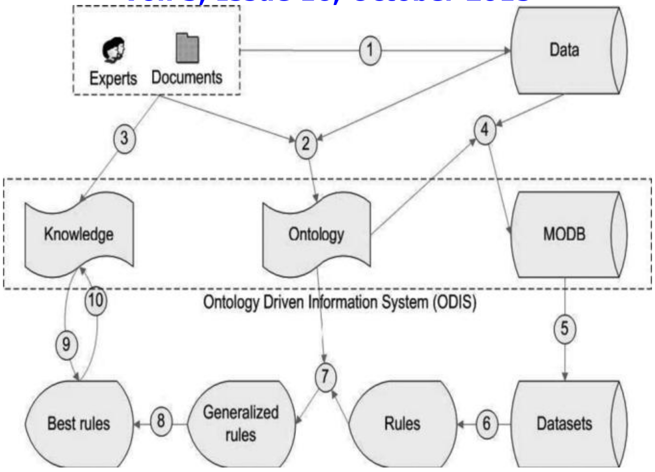


Figure 1 KEOPS methodology

3.3 Clustering

The purpose of gene-based clustering is to group togetherexpressed genes which indicate cofunction and coregulation. cluster analysis is typically the first step in data mining and knowledge discovery[10]. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. Due to the complex procedures of microarray experiments, gene expression data often contains a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.

One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples**.** Gene expression can be analyzed into two ways gene based clustering and sample based clustering.

A .Gene-Based Clustering

In gene-based clustering, the genes are treated as the objects, while the samples are the features. On the other hand, the samples can be partitioned into homogeneous groups.

B. Sample-Based Clustering

Sample-based clustering regards the samples as the objects and the genes as the features. The distinction of gene-based clustering and sample based clustering is based on different characteristics of clustering tasks for gene expression data. Some clustering algorithms, such as K-means and hierarchical approaches, can be used both to group genes and to partition samples.

3.3.1 K-Means

K-means clustering [4], [8] partitions objects into groups that have little variability within clusters and large variability across clusters. The user is required to specify the number k of clusters a priori. Estimation is iterative, starting with a random allocation of objects to clusters, re-allocating to minimize distance to the estimated “centroids” of the clusters, and stopping when no improvements can be made. The centroid is the point whose attributes take the mean expression level of the objects in the clusters. K-medoids clustering is similar, except that the center of the clusters is defined by “medoids”, similar to centroids, but based on medians.

Our empirical study has shown that the K-Means algorithm typically converges in a small number of iterations. However, it also has several drawbacks as a gene-based clustering algorithm.

First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of *k* and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine tuning process may not be practical.

Second, gene expression data typically contain a huge amount of noise; however, the K-Means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise.

3.3.2 Hierarchical Clustering

Hierarchical clustering [4] generates a hierarchical series of nested clusters which can be graphically represented by a tree, called dendrogram. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. By reordering the objects such that the branches of the corresponding dendrogram do not cross, the data set can be arranged with similar objects placed together. The hierarchical clustering scheme: Let S={*Si,j}* is the input similarity matrix, where *Si,j* indicates similarity between two data objects based on Euclidean distance.

3.4 Classification

Classification techniques can be used in microarray analysis to predict sample phenotypes based on gene expression patterns. While novel and microarray specific classification tools are constantly being developed, the existing body of pattern recognition and prediction algorithms provide effective tools. Dudoit and colleagues [5] offer a practical comparison of methods for the classification of tumors using gene expression data. Relevant tools from the statistical modeling tradition include: discriminant analysis, tree-based algorithms, such as classification and regression trees (CART) and variants.

3.4.1 Support Vector Machines

SVMs are supervised, machine learning algorithms that seek cuts of the data that separate classes effectively, that is by large gaps. Technically, SVMs operate by finding a hypersurface in the space of gene expression profiles, that will split the groups so that there is largest distance between thehyper surface and the nearest of the points in the groups.
More flexible implementations allow for imperfect filteringof groups and promiscuous analysis.

3.4.2 Discriminant Analysis

Discriminant analysis and its derivatives are approaches foroptimally partitioning a space of expression profiles into subsets that are highly predictive of the phenotype of interest, for example by maximizing the ratio between-classes variance to within-class variance.

3.4.3 Classification Trees

Classification trees recursively partition the space of expression profiles into subsets that are highly predictiveof the phenotype of interest. They are robust, easy-to-use,and can automatically sift large data sets, identifying important patterns and relationships. No pre - screening ofthe genes is required. The resulting predictive models canbe displayed using intuitive graphical representations.

3.5 Association Rule Mining

Gene expression data, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. Association rules [6] can reveal biologically relevant associations between different genes or between environmental effects and gene expression [9]. An association rule has the form *LHS*⇒*RHS*, where *LHS* and *RHS* are disjoint sets of items, the *RHS* set being likely to occur whenever the *LHS* set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes.

An example of an association rule mined from expression data might be {*cancer*}⇒{*gene A*↑, *gene B*↓, *gene C*↑}, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene *A* was measured as being up (i.e. highly expressed), gene *B* was down (i.e. highly repressed), and gene *C* was up, altogether. The first step in finding association rules is to look for frequent itemsets. A commonly used algorithm for doing this is the *Apriori* algorithm. The algorithm relies upon a simple yet fundamental property of frequent itemsets, called the *apriori* property: Every subset of a frequent itemset must also be a frequent itemset.

III. SIMULATION RESULTS

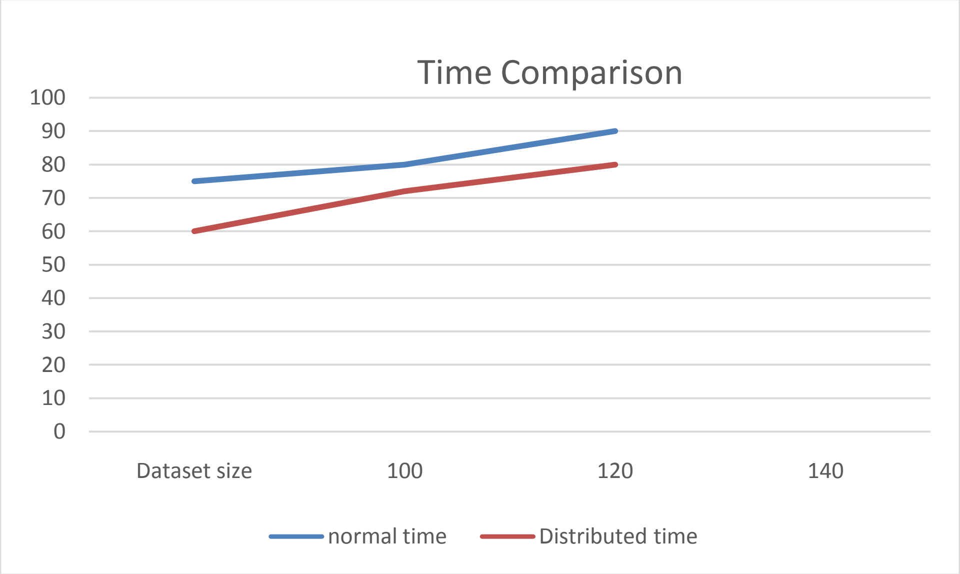
The simulation results shows the comparison between the execution time in the normal case and in the distributed manner. From the figure, we can understand that the distributed execution of the gene expression mining take comparatively less time than the normal approach. Thus a lot of time can be saved.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



IV. CONCLUSION

This paper deals with mining gene expression data in a distributed manner. Various steps included in the process have been described here. The advantage of using distributed approach is that it helps in reducing the overhead by splitting the entire work among a number of machines.

REFERENCES

1. Shaurya Jauhari And S.A.M Rizvi " Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest" , IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 11, No. 3 May/June 2014
2. E. Shay, (2003, Jan.). "Microarray Cluster Analysis And Applications" [Online]. Available: Ttp://Www.Science.Co.Il/Enuka/Essays/Microarray
3. B. Collard, "An Ontology Driven Data Mining Process" Inst. TELECOM, TELECOM Bretagne, CNRS FRE 3167 LAB-STICC, Technopole Brest-Iroise, France & Univ. Nice Sophia Antipolis, France, 2008.
4. D. Jiang, C. Tang, And A. Zhang, "Cluster Analysis For Gene Expression Data: A Survey," IEEE Trans. Knowl. Data Eng., Vol. 16, No. 11, Pp. 1370–1386, Nov. 2004.
5. Dudoit S, Fridlyand J, Speed TP. Comparison Of Discrimination Methods For The Classification Of Tumors Using Gene Expression Data. *JASA* 97; 77–87 2002
6. H. Creighton, "Mining Gene Expression Databases For Association Rules," Bioinformatics, Vol. 19, No. 1, Pp. 79–86, 2003
7. Pubmedhealth- U.S. Nat. Library Med., (2012). [Online]. Available: Http://Www.Ncbi.Nlm.Nih.Gov/Pubmedhealth/PMH0002267/
8. Daxin Jiang, Chun Tang, And Aidong Zhang,‖ Cluster Analysis For Gene Expression Data: A Survey‖ IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 11, November 2004
9. Doddi,S., Marathe,A., Ravi,S.S. And Torney,D.C. (2001) Discovery Of Association Rules In Medical Data. *Med. Inform. Internet. Med.*, **26**, 25–33.
10. A. Ben-Dor, R. Shamir, And Z. Yakhini, ―Clustering Gene Expression Patterns,‖ J. Computational Biology, Vol. 6, Nos. 3/4, Pp. 2 81-297,1999

BIOGRAPHY

Cisy Somanis a M Tech scholar in the Computer Science department, ICET, MG University. She received Bachelor of Technology (B Tech) degree in 2013 fromM.G University, Kottayam, Kerala. Her research interests are data mining and databases.

Abdul Ali is an Assistant professor in the Computer Science department, ICET,M.G University. He received B-Tech degree in 2007 from M. G. University, Kottayam, Kerala. He received M Tech degree in 2010 from M S university, Trinulveli.His research interests are image processing and networking.