ANUSANDHAN- AISECT University Journal
Vol. 06, Issue No. 12, September-2017, P-ISSN 2278-4187

Mining of Microarray Databases for Categorizing the Genes for Various Diseases

Dr. Sitesh Kumar Sinha[1]* Dr. Vijay Kumar Singh[2] Megha Sinha[3]

[1]Prof., Dept. of CSE, AISCET University, Bhopal (M.P.) India

[2]Associate Prof., SIT, Aurangabad (Bihar) India

[3]Asst. Prof., RVSCET, Jamshedpur (Jharkhand) India

Abstract – Presently researchers show lot interest in microarray gene expression dataset. Recently huge library of biological information mining algorithm has been developed for the analytical evaluation of gene expression. Mining microarray gene expression is an imperative subject in bioinformatics in diagnosis of disease. This research paper analyzes how microarray data sets are used to predict the various diseases that spread through gene. The paper mainly focuses on prediction of heart disease, obesity and diabetes. These are the diseases that are deadly in nature.

Keywords: Data Mining, Microarray Gene, Bioinformatics

----------♦----------

I. INTRODUCTION

Data Mining is one of the most vital and motivating area of research with the objective of clinical diagnosis and prognosis requires efficient and fast classification techniques, which in turn requires a large amount of genetic data generation and analyzing these huge data. The large amount of genetic data generated is obtained using the microarray technique in which expression of thousands of genes is concurrently measured and we are in the need of an efficient data mining technique for these huge data.

In Bioinformatics, mining micro-array gene expression data is an imperative technique in the diagnosis of disease, drug development, genetic functional interpretation and gene metamorphisms etc. Recently biological information mining plays an

Key role in the disease predication. There are diverse types of disease predicated by microarray database mining using clustering techniques, namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Obesity Diabetes etc.

A microarray is a massive collection of spots that contain massive amounts of compressed data. Researchers in the bioinformatics use microarray because DNA contains so much information on a micro-scale. Each spot of a microarray thus could contain a unique DNA sequence. So, it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased). Such analyses produce a list of genes whose expression is considered to transform and such genes are known as differentially expressed genes. Identification of differential gene expression is the first task of an in-depth microarray analysis. There are two common methods for in depth microarray data analysis, i.e. clustering and classification (Mutch et. al. 2001). Clustering is a unsupervised approach that classifies data into groups of genes or samples with similar patterns that are characteristic to the group. Classification is supervised learning and known as class prediction or discriminate analysis (Dinger et. al., 2012). Generally, classification is a process of learning-from-examples.

Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes.

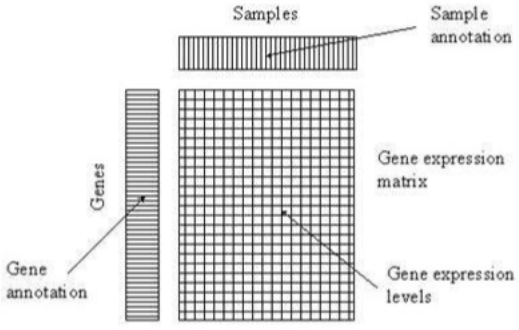

Fig. 1 a microarray

II. CHALLENGES FOR MICROARRAY DATA MINING

Analysis of microarrays presents many unique challenges for data mining. Typical data mining applications in domains like banking or web, have many records (thousands and sometimes millions), while the number of fields is much smaller (at most several hundred). In contrast, a typical microarray data analysis study may have only a small number of records (less than a hundred), while the number of fields, corresponding to the number of genes, is typically in thousands. Given the difficulty of collecting microarray samples, the number of samples is likely to remain small in many interesting cases.

However, having so many fields relative to so few samples creates a high likelihood of finding "false positives" that are due to chance – both in finding differentially expressed genes, and in building predictive models. We need especially robust methods to validate the models and assess their likelihood.

The main types of data analysis needed to for biomedical applications include:

(a) **Gene Selection** – in data mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class.

(b) **Classification** – classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature.

(c) **Clustering** – finding new biological classes or refining existing ones.

Attempts to find invariant or differential molecular behavior relevant to a given biological problem are also limited by the fact that in many cases little is known about the normal biological variation expected in a given tissue or biological state.

Their analysis applied to six mouse tissues resulted in several genes which showed significant biological variations even among identical mice and provides a valuable compendium of normal variation in gene expression for mouse models.

Another approach to determining variability in small samples is taken by S. Mukherjee, P. Sykacek, S. Roberts, and S. Gurr, who propose a gene-ranking algorithm using bootstrapped P-values. This approach is especially beneficial for considering small -sample variability in observed values of the test statistic. They show that this method outperforms widely used two-sample T-test on artificial data and apply the method to two real datasets.

Most of the current gene selection methods in use today evaluate each gene in isolation and ignore the gene to gene correlations. From a biological viewpoint, however, we are aware that groups of genes working together as pathway components and reflecting the states of the cell are the real atomic units, or features, by which we might be more likely to predict the character or type of a particular sample and its corresponding biological state. It is these patterns of coherent gene expression that must form the input data on which sophisticated computational methods should operate. In this context B. Hanczar, M. Courtine, A. Bennis, C. Hennegar, K. Clément, and J. Zucker suggest to increase the accuracy of microarray classification by selecting appropriate "prototype" genes that represent a group of genes that share a profile and better represent the phenotypic class of interest. They present interesting results of the advantages of using prototype-based feature selection to classify adenocarcinomas.

III. CARDIOVASCULAR DISEASE

Cardiovascular disease (CVD) is one of the leading causes of death in human life, and is influenced by both environmental and genetic factors. With the recent advances in micro array tools and technologies there is potential to predict and diagnose heart disease using micro array DNA data from analysis of blood cells. It is not a single disease but is a combination of many individual diseases as listed under the 9thRevision of the International Classification of Diseases (1975). It includes acute myocardial infarction and angina pectoris among others. It is a complex multi factorial process that involves lipid deposition on arteries of the heart, macrophages, blood pressure, and rheology of blood flow, smooth muscle proliferation, thermogenesis, platelet aggregation, insulin resistance and other factors. Every year, millions of deaths worldwide are attributed to cardiovascular diseases and more than half of them are found in developed countries.

Table 1

Results Obtained from IHDPS

Technique	Accuracy
NAVIE BAYES	86.55%
DECISION TREE	89%
KNN	85.53%

Chronic degenerative diseases such as cancer and cardiovascular disease have emerged as the major causes of death and hence, finding cost effective methods to control CVD is one of the challenges for public health in day today life the risk factors for CVD had been documented and among the more

established ones are: family history (genetic factors), plasma lipid, lipoprotein, plasma lipoprotein (a), diet, gender, elevated blood pressure, physical inactivity etc. (Heng, 1999).

IV. DIABETES MELLITUS

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease referred to diabetes. Diabetes is a chronic disease that is associated with considerable morbidity and mortality. Molecular Biology research involves in this area through the development of the technologies used for carrying them out. DNA Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non-traceable

(a) **Related Work for Prediction of Diabetes Mellitus** Microarray techniques using cDNAs are much high throughput approaches for large scale gene expression analysis and enable the investigation of mechanisms of fundamental processes and the molecular basis of disease on a genomic scale. Several clustering techniques have been used to analyze the microarray data. As gene chips become more routine in basic research, it is important for biologists to understand the biostatistical methods used to analyze these data so that they can better interpret the biological meaning of the results. Strategies for analyzing gene chip data can be broadly grouped into two categories: Discrimination and clustering. Discrimination requires that the data consist of two components. The first is the gene expression measurements from the chips run on a set of samples. The second component is data characterizing. For this method, the goal is to use a mathematical model to predict a sample characteristic, from the expression values. There are a large number of statistical and computational approaches for discrimination ranging from classical statistical linear discriminate analysis to modern machine learning approaches and Pattern recognition

In clustering, the data consists only of the gene expression values. The analytical goal is to find clusters of samples or clusters of genes such that observations within a cluster are more similar to each other than to observations in different clusters. Cluster analysis can be viewed as a data reduction method in that the observations in a cluster can be represented by an 'average' of the observations in that cluster. There are a large number of statistical and computational approaches available for clustering. These include hierarchical clustering and k-means clustering for the analyze the clusters of genes expression.

In hierarchical clustering, individuals are successively integrated based on the dissimilarity matrix computed by data, to obtain a dendrogram which contains inclusive clusters. In the context of microarray analysis, it is used to classify unknown genes or cases of disease. Several different algorithms will produce a hierarchical clustering from a pair-wise distance matrix. The algorithms begin with each gene by itself a separate c luster. These clusters correspond to the tips of the clustering tree (dendrogrm). The algorithms search the distance matrix for the pair of genes that have the smallest distance between them and merge these two genes into a cluster. Many algorithms follow this series of steps to produce hierarchical clustering of data. Average linkage is one of many hierarchical clustering algorithms that operate by iteratively merging the genes or gene clusters with the smallest distance between them followed by an updating of the distance matrix.

An overview of the literary review, hierarchical clustering of microarray data, emphasizing the relationship between a dendrogram and spatial representations of genes. Consideration of this relationship provides an intuitive understanding of how to analyze microarray data and can make it is easier to interpret the results of a cluster analysis in a **biological framework. The fact that the 'heat maps'** found in most of the microarray publications are based on hierarchical clustering indicates that an understanding of this general method is valuable to those who are just beginning to read the microarray literature and even to those who are using supervised methods

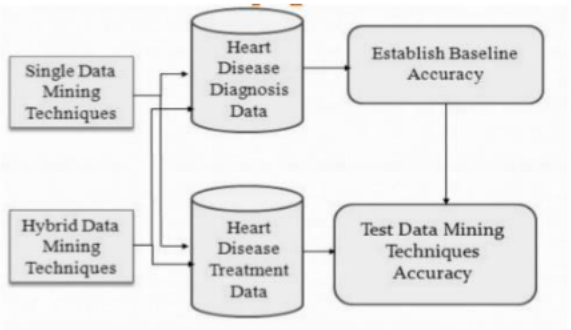

Fig. 2 Data Mining

V. CONCLUSION

Microarrays are a new technology revolutionary in nature. This technology has great potential to provide medical diagnostics with a great degree of accuracy.

Microarrays assist to find the right treatment and cure for several diseases and provide a detailed genome-wide molecular portrait of cellular states. This paper contains a description of second generation methodologies and techniques that are being used or are presently under the phase of development. As it can be seen from the results, they are very promising and extend the possibilities of applying computational analysis and data mining to aid research in biology and medicinal science. We have underlined to emphasize the large potential payoff of these analytical efforts and pointed out the huge challenges ahead as well.

REFERENCES

Asha Rajkumar, G. Sophia Reena (2010). "Diagnosis of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology 38 Vol.10 Issue 10 Ver. 1.0.

Barile M. (2011). `Taxicab metric - MathWorld: A Wolfram.

Benjamini, Y. & Hochberg, Y. (1995). "Controlling the false discovery rate: a practicaland powerful approach to multiple testing',Journal of the Royal Statistical Society 57(1), pp. 289-300.

G. Parthiban, A. Rajesh, S. K. Srivatsa (2011). "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications (0975 –8887) Volume 24–No.3.

H Hasan; K Raza, (2012). International Journal of Computer Sciences, World Academy of Science, Engineering & Technology, 6(5), pp. 1307-1310.

Heng C. K. (1999). "Candidate genes for Coronary Artery Disease", PhD Thesis, National University of Singapore, Department of Paediatrics, 1996. Brown, P.O., Botstein, D., "Exploring the new world of the genome with DNA microarrays", Nature Genetics Supplement, Volume 21, pp. 33-37.

https://searchenginereports.net/plagiarism-checker

https://smallseotools.com/plagiarism-checker

Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. No. 6.

Li H and F. Hong (2001). Cluster-Rasch models for microarray gene expression data. Genome Biology, 2(8)}:research0031.1-0031.13, Pl. check plagiarism on the Links below:

M. Anbarasi, E. Anupriya, CH. S. Iyenga (2010). "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm" , International Journal of Engineering Science and Technology Vol. 2 (10), pp. 5370- 5376.

Mutch D.M. et. al. (2001). Genome Biol. (12): Preprint0009 [PMID:11790248].

S.C. Dinger; M.A. Van Wyk; S. Carmona; D.M. Rubin (2012). Bio Medical Engineering OnLine, 11(1), p. 85.

Shantakumar, B. Patil, Y. S. Kumaraswamy (2009). "Intelligent and Effective Heart Attack Prediction System UsingData Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450 216X Vol.31 No.4, pp. 642-656.

Corresponding Author

Dr. Sitesh Kumar Sinha*

Prof., Dept. of CSE, AISCET University, Bhopal (M.P.) India

E-Mail – siteshkumarsinha@gmail.com