Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-2, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Mining of Microarray Databases for the Classification of Genes Expression for Various Diseases Using Cluster Analysis and Predictive Regression

M.sumalatha[1], Dr. Latha Parthiban[2] & Rane zab anish kumar[3]
[1]Research Scholar, Periyar University Salem, Tamilnadu
[2]Asst. Professor, Pondicherry University Community College, Pondicherry, India
[3]Asst. Professor, PRIST University, Pondicherry, India

Abstract*: These days' researchers are providing great awareness about 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 used to predicate the various diseases which is spread through gene. This paper mainly focused on predicating heart disease, obesity and diabetes*

1. Introduction

Every moment we faced a life with a progress of new diseases. The quick and efficient identification of such diseases are vital to save human life ultimately. In present era many disease are caused by gene transformation. Hence we predicate the disease from root level that means form the gene transformation level. Computer science and Bio informatics serves a lot in the research related to disease and treatment by providing efficient tools and techniques in this research.

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 large amount of gene data generation and analyzing these large amounts of data. The large amount of gene data generation is obtained using microarray technique in which expression of thousands of genes are concurrently measured and we are in need of an efficient data mining technique for these large amounts of data

In bioinformatics, mining micro-array gene expression data is an imperative technique in diagnosis of disease, drug development, genetic functional interpretation and gene metamorphisms etc. Recently biological information mining plays an important role in the disease predication. There are different 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 huge 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 change and 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 [9]. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group. Classification is supervised learning and also known as class prediction or discriminate analysis [5]. 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.

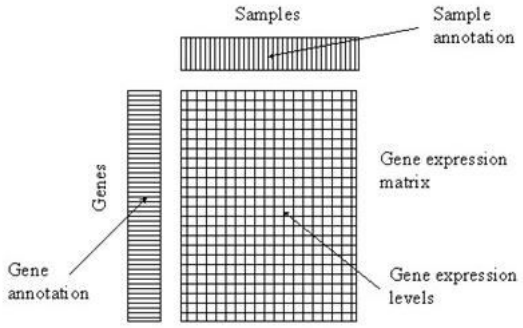

Figure 1 A microarray

2. 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, rheology of blood flow, smooth muscle proliferation, thrombogenesis, 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.

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 [7]

2.1. Related Work for the Prediction Of CVD

Three different supervised machine learning algorithm are proposed in [3] analyses for CVD gene expression dataset [9]. They are Naïve Bayes, K-NN, and Decision List algorithm. In this paper, the author concluded that Naïve Bayes algorithm performs well for analyzing the gene expression dataset when compared to other algorithm.

Table 1.Results Obtained from IHDPS

Technique	Accuracy
Navie Bayes	86.55%
Decision Tree	89%
KNN	85.53%

For predicating CVD association rule data mining technique is proposed in for gene expression dataset. In this, unfortunately they have produced a large number of rules when association rules are applied to gene expression dataset [8]. Most of the rules are medically irrelevant to the gene data. The authors proposed four constraints to reduce the number of rules i.e., item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering [11]. The important issue is without validation, the association rules are mined on the entire gene dataset. To solve these limitations, the author introduced an algorithm that uses search constraints to decrease the number of rules. The training set searches association rules and test gene dataset to check the validation. Here, a new parameter lift is used instead of support and confidence. Lift has been used as the metrics to evaluate the reliability and medical significance of association rules. To validate the results the two basic statistics sensitivity and specificity are used by researchers the chance of correctly identifying sick patients are defined by sensitivity and chance of correctly identifying healthy individuals is defined by specificity. To find predictive association rules in CVD gene expression dataset the algorithm has three steps: [11]

(i) In gene dataset both the categorical and numeric attribute are transformed into transaction dataset.
(ii) have to find the predictive association rules with medically relevant attributes the search process
(iii) And validate the association rules the train and test approach should be used.

Genetic algorithms have been used in to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is one of the supervised learning methods to extract models describing important classes of data. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the Presence of heart disease in patients [11].

3. Classification via clustering: Clustering is the process of grouping same elements. This technique

may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool [5]. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering is not performing well when compared to other two methods.

In the survey of [10] Naïve bayes have been used to predict attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. The clinical dataset is having been collected from one of the leading diabetic research institute in Chennai. The records of 500 patients are taken. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. It consists of data pre-processing, classification, regression, association rules, clustering and visualization tools. We have used Naïve bayes method to perform the mining and classification process. They used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process. From the experiment the result of bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% in average, recall of 74% in average, and F-measure of 71.2% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

4. 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

4.1. Related Work for Predection 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 consist 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 they are 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 n a separate c luster. These clusters correspond to the tips of t he

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 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

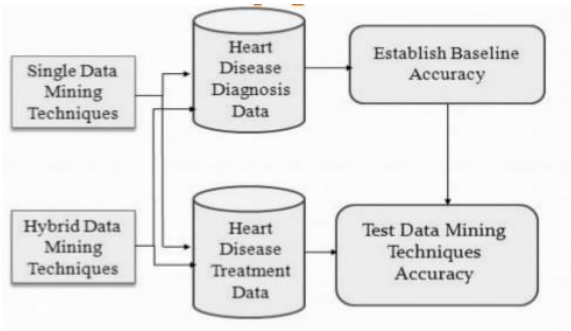

Figure 2 Data Mining

5. Obesity

Obesity is a typical representative of disorders affecting particularly people in developed countries. Obesity is base of so many major chronic diseases such as Heart attack, Diabetes type II, In India is rising trend in obesity, progress of urbanization, lifestyle and behavioral changes are cause of Obesity. Patients having complex metabolic disorders or serious diseases may suffer from obesity as well. Lack of awareness about complications of obesity its one of main cause of increasing the prevalence of obesity in India. It is important to detect whether the cause is an excessive income of food associated with a limited physical activity or an effect of metabolic problems

5.1 Related Work for Predicting Of Obesity

Within nutrigenomics, clustering using data generated by microarray gene expression profiles can be used to identify sub-populations of subjects that respond differently to a given diet intervention. The use of clustering analyses is promising in obesity-related research as personalized nutrition is gaining popularity. In thisfocuses on clustering a human subcutaneous adipose tissue gene expression data set obtained during a low-calorie diet intervention to aid in the prediction of 6-month weight loss maintenance [8]. The aims of the study were to identify the best performing clustering method for clustering samples,to identify differential responders to the low-calorie diet, and to identify the biological pathways affected during the low-calorie diet by weight maintainers and weight regainers. MCLUST performed the best when clustering samples using relative weight change and either fasting insulin or insulin resistance change. Furthermore, it identified differences in the regulationof pathways between weight maintainers and regainers

In this research used data from the DiO Genes population to illustrate advances in predicting successful weight loss maintenance. Clustering of biological samples indicated that the combination of insulin resistance change or fasting insulin change and relative weight change defined successful weight loss maintenance better than combinations including other parameters or relative weight change alone. Clustering of genes indicated that gene expression is regulated differently between weight maintainers and weight regainers. The study exemplifies the usefulness of gene expression and clustering analyses on assessing weight loss maintenance. Successful weight maintenance continues to be a priority in the battle against obesity, and nutrigenomic approaches and results such as the ones presented here will aid in the endeavor

In this survey paper the problem of summarizing the different algorithm of data mining are used in the field of medical prediction are discussed. The main focus is on using different algorithm and combination of several targets attributes for different types of disease prediction using data mining. First we discuss about the heart disease prediction, in that machine learning algorithms namely naïve bayes, K-NN, Decision List. Of these the classification accuracy of the naïve bayes algorithm is better when compared to other algorithm. In Weighted Associative Rule Classifier, the GUI has been designed to enter the patient record and the presence of Heart Disease for a patient is predicted by using the rules stored in the rule base. Next we discuss the feature subset selection using genetic algorithm. In this attributes are reduced using genetic search. Here

the accuracy is compared to the three classifiers namely Decision Tree, Naïve bayes and classification via clustering. Association rule discovery is mainly based on four constraints namely item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering. To find predictive rules in medical data set the three important steps are generated in this algorithm [8,6]. The heart disease is diagnosed for diabetic patients using naïve bayes technique [2]. Of these the author concluded that naïve bayes classify 74% of input instances correctly. Next we discuss about the breast cancer prediction. It is performed by using various data mining techniques namely C4.5, ANN and fuzzy decision trees. By using C4.5 the author discussed and resolved the issues and algorithms of the problem. Using ANN the author concluded that the consistent accuracy over time and good performance of the network is trained. The fuzzy decision tree survives by using 10 fold cross validation method. Finally we discuss about diabetes prediction, by using homogeneity based algorithm the author find over fitting and overgeneralization behavior of classification. By using genetic algorithm the author predicts accuracy of the class. In future the work can be expanded and enhanced for the automation of various types of disease prediction.

6. Conclusion

Both data mining and bioinformatics are fast expanding research frontiers. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective bio-data analysis. We believe that the active interactions and Collaborations between these two fields have just started and a lot of exciting results will appear in the near future.

7. References

[1] **Asha Rajkumar, G.Sophia Reena**, "*Diagnosis Of Heart Disease Using Datamining Algorithm",* Global Journal of Computer Science and Technology 38 Vol.10 Issue 10 Ver. 1.0 September 2010.

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

[3] **Barile, M. (2011),** `Taxicab metric - MathWorld: A Wolfram.

[4]**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), 289{300.

[5]. **SC Dinger;MA Van Wyk; S Carmona; DM Rubin**, BioMedical Engineering OnLine,2012, 11(1), 85.

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

[7]. **Heng, C.K**., "*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 21,33-37,1999.

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

[9] **Mutch DM** *et al. Genome Biol*. 2001(12): Preprint0009 [PMID:11790248].

[10]. **G. Parthiban, A. Rajesh, S.K.Srivatsa** "*Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method",* International Journal of Computer App lica tions (0975 –8887) Volume 24–No.3, June 2011.

[11]. **Shantakumar, B.Patil, Y.S.Kumaraswamy,** "*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 (2009), pp.642-656 .