ISSN (ONLINE): 2455-9261

SOUTH ASIAN JOURNAL OF RESEARCH IN ENGINEERING SCIENCE AND TECHNOLOGY (SAJREST)

VOLUME: 01 ISSUE: 06 WWW.SAJREST.COM

PATTERN SIMILARITY SEARCH USING EXPECTATION MAXIMIZATION (EM) ALGORITHM

M. Anandakumar[1], S. Aiswarya[2], N. Bakyalakshmi[3], S. Brindha[4],
[1] Assistant Professor, [2,3,4] UG Scholar,
Arasu Engineering College,
Kumbakonam, Tamilnadu, India.
anandlogo@gmail.com ,Aishuiyer176@gmail.com , bakyanatarajan28@gmail.com , brindha0196@gmail.com

Abstract*— *The DNA microarray technology has modernized the approach of biology research in such a way that scientists can now measure the expression levels of thousands of genes simultaneously in a single experiment. Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. Diseases classification with gene expression data is known to include the keys for addressing the fundamental harms relating to diagnosis and discovery. The recent introduction of DNA microarray technique has complete simultaneous monitoring large number of gene expressions possible. With this large quantity of gene expression data, experts have started to discover the possibilities of disease classification using gene expression data. Quite a large number of methods have been planned in recent years with hopeful results. But there are still a set of is sues which need to be address and understood. In order to gain insight into the disease classification difficulty, it is necessary to get a closer look at the problem, the proposed solutions and the associated issues all together. In this project, we present a comprehensive clustering method and classification method such as Spatial Expectation Maximization, K-NN classification algorithm and estimate them based on their evaluation time, classification accuracy and ability to reveal biologically meaningful gene information. Based on our multiclass classification method to diagnosis the diseases and also find severity levels of diseases. Our experimental results show that classifier performance through graphs with improved accuracy.

Keywords*— *Bio-medical research, DNA microarray, Gene sequence, Clustering, Classification

ISSN (ONLINE): 2455-9261

SOUTH ASIAN JOURNAL OF RESEARCH IN ENGINEERING SCIENCE AND TECHNOLOGY (SAJREST)

VOLUME: 01 ISSUE: 06 WWW.SAJREST.COM

I. INTRODUCTION

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesized by the process of photolithography. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available [66, 20]. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

1.1 Challenges in gene clustering:

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem. First, cluster analysis is typically the first step in data mining and knowledge discovery. 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. For example, a clustering algorithm which can accurately estimate the "true" number of clusters in the data set would be more favored than one requiring the pre-determined number of clusters. Second, due to the complex procedures of microarray experiments, gene expression data often contain 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. Third, our empirical study has demonstrated that gene expression data are often "highly connected", and clusters may be highly intersected with each other or even embedded one in another. Therefore, algorithms for gene-based clustering should be able to effectively handle this situation. Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters (e.g., which clusters are more close to each other, and which clusters are remote from each other), and the relationship between the genes within the same cluster (e.g., which gene can be considered as the representative of the cluster and which genes are at the boundary area of the cluster). A clustering algorithm, which can not only partition the data set but also provide some graphical representation of the cluster structure, would be more favored by the biologists.

II. RELATED WORK

Booma,et.al,…[1] identified normal or abnormal genes is important for clinical analysis and diagnosis. In this work, a novel framework for analyzing gene data was designed and developed. For this, initially, Bio-information from gene expression data was evaluated with the establishment of analyzing biological process using heuristic search. This method extracted the biological process on gene expression data. The proposed method used heuristic search algorithm for identifying the biological process and processed based on two phases. The first phase was initialization phase and another was iterative adjustment phase. With respect to these two phases, the biological process of each gene and gene selection for a dataset is identified in terms of physiological data on gene expression datasets.

Balasubramanian,et.al,…[2] proposed fuzzy logic based preprocessing technique to reduce the redundant information and grouping the similar genes from large amount of microarray data. The propose Parallel Island Model GA is implemented for gene feature selection process. Our propose feature selection algorithm is implemented based on multi objective genetic algorithm. This uses a different operator called multi objective operator. Multi objective aspect is defined to find the pareto optimal solutions for ranking. The best features are selected in short time. The best identified gene subsets are evaluated by parallel version of SVM Classifier. Our method has given good classification accuracy than other methods. This method uses the island model for generating the best population. The multiple islands are implemented in parallel, which has significantly reduced the execution time in the process of best feature selection.

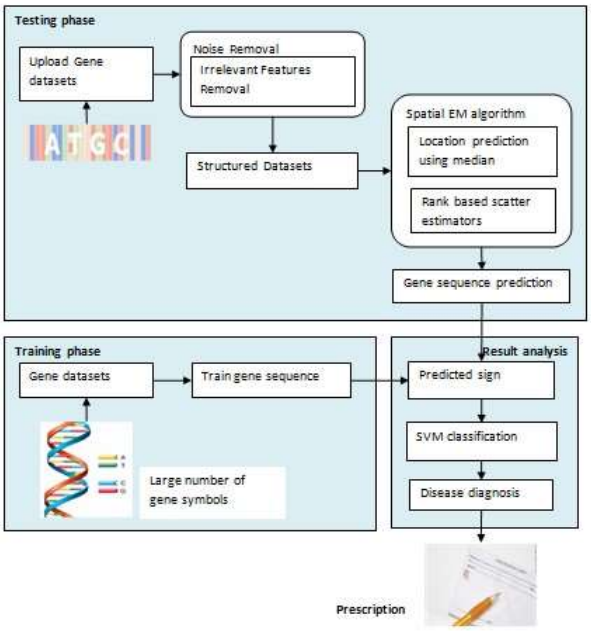
Bennet,et,al,… [3] conducted in the space of genes, evaluating the goodness of each gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. It is claimed that this approach obtains better predictive accuracy estimates

ISSN (ONLINE): 2455-9261

SOUTH ASIAN JOURNAL OF RESEARCH IN ENGINEERING SCIENCE AND TECHNOLOGY (SAJREST)

VOLUME: 01 ISSUE: 06 WWW.SAJREST.COM

machine learning techniques, in particular a KNN method. The basic tool used is a modified version of KNN classifier which employs a set of mapping functions to map the input data into the reproducing kernel Hilbert space, where the mapping function is implicitly defined by the kernel function :$k(x_i, x_j) = \Phi(x_i). \Phi(x_j)$. The efficiency of classification depends on the type of kernel function that is used. So here we will analyze the performance of various kernel functions used for classification purpose. Finally predict the diseases with severity levels and predict various types of diseases. The proposed architecture is shown in fig 1.



VI. ADVANTAGES OF THE PROPOSED SYSTEM

The key advantage of supervised learning methods over unsupervised methods like clustering is that by having an explicit knowledge of the classes the different objects belong to, these algorithms can perform an effective feature selection if that leads to better prediction accuracy. Severity levels are predicted automatically.

VII. CONCLUSION

Microarray is an important tool for cancer classification at the molecular level. It monitors the expression levels of large number of genes in parallel. With large amount of expression data obtained through microarray experiments, suitable statistical and machine learning methods are needed to search for genes that are relevant to the identification of different types of disease tissues. In this paper, we have proposed a hybrid gene selection method, which combines a spatial EM methods and KNN classification to achieve high classification performance. The method was designed to address the importance of gene ranking and selection prior to classification, which improves the prediction strength of the classifier. The project focused on promising accuracy results with very few number of gene subsets enabling the doctors to predict the type of cancer. The results on various disease datasets shows the importance of the same classifier used for both the gene selection and classification can improve the strength of the model. Then provide severity level for each classified diseases.

REFERENCES

[1] Booma, P. M., and S. Prabhakaran. "CLASSIFICATION OF GENES FOR DISEASE IDENTIFICATION USING DATA MINING TECHNIQUES." Journal of Theoretical and Applied Information Technology 83.3 (2016): 399.

[2] Natarajan, A., and R. Balasubramanian. "A Fuzzy Parallel Island Model Multi Objective Genetic Algorithm Gene Feature Selection For Microarray Classification." International Journal of Applied Engineering Research 11.4 (2016): 2761-2770.

[3] Bennet, Jaison, Chilambuchelvan Ganaprakasam, and Nirmal Kumar. "A hybrid approach for gene selection and classification using support vector machine." Int. Arab J. Inf. Technol. 12.6A (2015): 695-700.

[4] Nagpal, Rashmi, and Rashmi Shrivas. "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data." Journal of Scientific and Technical Advancements 1.4 (2015): 19-23.

[5] Thangaraju, Mr P., and R. Mehala. "Novel Classification based approaches over Cancer Diseases." system 4.3 (2015). [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, PatternClassification and Scene Analysis. John Wiley and Sons,1999.

ISSN (ONLINE): 2455-9261

SOUTH ASIAN JOURNAL OF RESEARCH IN ENGINEERING SCIENCE AND TECHNOLOGY (SAJREST)

VOLUME: 01 ISSUE: 06 WWW.SAJREST.COM