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.2, Special Issue 5, October 2014

A Modified Rough Fuzzy "Clustering - Classification" Model For Gene Expression Data

Lt.Thomas Scaria[1], Dr.T Christopher[2], Gifty Stephen[3]

Research Scholar, Periyar University, Salem, Tamil Nadu, India[1]

Assistant Professor and Head, Department of CS, Govt Arts College, Udumalpet,Tamilnadu, India[2]

Assistant Professor, Department of CS, Sir Sayd Institute of Technical Studies, Thaliparamba, Kerala, India [3]

ABSTRACT: Microarray technology is one of the important biotechnological means that has made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and a cross collections of related samples . An important application of microarray data is to elucidate the patterns hidden in gene expression data for an enhanced understanding of functional genomics. A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample or time point, and each entry of the matrix is the measured expression level of a particular gene in a sample or time point, respectively. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques, which is essential in pattern recognition process to reveal natural structures. Recent decades, more and more researchers study on gene expression profile analysis which provides a more precise and reliable way for disease diagnosis and treatment when compared with traditional cancer diagnosis approaches based on the morphological appearance of cells.Through this research we mainly aim toStudy and analyse different clustering and classification model regarding gene expression data, Design and develop an efficient method for gene expression data clustering and classification finally Conduct experimental analysis to evaluate the proposed methodology to prove the significance of the method

KEYWORDS: Micro Array, Expression Table, Marker genes, IBSA, Genomics, Proteomics, mRNA

I. INTRODUCTION

A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively [1], [2]. However, for most gene expression data, the number of training samples is still very small compared to the large number of genes involved in the experiments. When the number of genes is significantly greater than the number of samples, it is possible to find biologically relevant correlations of gene behaviour with the sample categories or response variables [3].

However, among the large amount of genes, only a small fraction is effective for performing a certain task. Also a small subset of genes is desirable in developing gene expression-based diagnostic tools for delivering precise, reliable, and interpretable results [4]. With the gene selection results, the cost of biological experiment and decision can be greatly reduced by analysing only the marker genes. Hence, identifying a reduced set of most relevant genes is the goal of gene selection. Cluster analysis is a technique for finding natural groups present in the gene set. It divides a given gene set into a set of clusters in such a way that two genes from the same cluster are as similar as possible and the genes from different clusters are as dissimilar as possible [7], [8]. To understand gene function, gene regulation, cellular processes, and subtypes of cells, clustering techniques have proven to be helpful. The co-expressed genes, that is, genes with similar expression patterns, can be clustered together with similar cellular

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.2, Special Issue 5, October 2014

functions. This approach may further understanding of the functions of many genes for which information has not been previously available [9], [10]. 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. The inference of regulation through gene expression data clustering also gives rise to hypotheses regarding the mechanism of transcriptional regulatory network [12].

Classification has been an important statistical data analysis tool in many fields. Particularly in computational biology and bioinformatics, clustering methods have been developed and applied extensively. In high throughput biological data sets such as those obtained from transcriptomics analysis, the mRNA levels of tens of thousands of genes are sampled simultaneously under particular experimental conditions. The success of co-expression networks identifying modules of co-regulated genes indicates that genes which show particular response profiles may well share a common function, or be regulated by the same transcription factors. It is therefore of interest to cluster genes on the basis of their response profiles. This gives an overview of the general patterns of gene expression, without getting lost in the sheer number of genes. The purpose of gene clustering is to group together co-expressed genes which indicate co-function and co-regulation. Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene clustering presents several new challenges and is still an open problem. The 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.

Different clustering techniques such as hierarchical clustering [13], k-means algorithm [14], self-organizing map[15], graph theoretical approaches [16], [17], [18], [19], model based clustering [20], and density-based approach [20] have been widely applied to find groups of co-expressed genes from microarray data. A comprehensive survey on various gene clustering algorithms can be found in[4] and [7].In this background ,some supervised attribute clustering algorithms such as supervised gene clustering , gene shaving ,tree harvesting , and partial least square procedure have been proposed to reveal groups of co-regulated genes with strong association to the sample categories. The supervised attribute clustering is defined as the grouping of genes or attributes, controlled by the information of sample categories or response variables.

II. RELATED WORK

PradiptaMaji and Sushmita Paul [1] have proposed judiciously integrating the merits of rough sets and fuzzy sets based gene clustering algorithm, termed as robust rough-fuzzy c-means. While the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in cluster definition, the integration of probabilistic and possibilistic memberships of fuzzy sets enables efficient handling of overlapping partitions in noisy environment. The concept of possibility lower bound and probabilistic boundary of a cluster, introduced in robust rough-fuzzy c-means, enables efficient selection of gene clusters. An efficient method is proposed to select initial prototypes of different gene clusters, which enables the proposed c-means algorithm to converge to an optimum or near optimum solutions and helps to discover co-expressed gene clusters. The effectiveness of the algorithm, along with a comparison with other algorithms, is demonstrated both qualitatively and quantitatively on 14 year microarray data sets.

Jianyong Sun et al. [2] have proposed a robust Bayesian mixture model for clustering data sets with replicated measurements. The model aims not only to accurately cluster the data points taking the replicated measurements into consideration, but also to find the outliers (i.e., scattered objects) which are possibly required to be studied further. A tree-structured variational Bayes (VB) algorithm is developed to carry out model fitting. Experimental studies showed that our model compares favourably with the infinite Gaussian mixture model, while maintaining computational simplicity. We demonstrate the benefits of including the replicated measurements in the model, in terms of improved outlier detection rates in varying measurement uncertainty conditions. Finally, we apply the approach to clustering biological transcriptomics mRNA expression data sets with replicated measurements.

Zhiwen Yu et al. [3] have proposed a method named as triple spectral clustering-based consensus clustering (SC3) and double spectral clustering-based consensus clustering (SC2Ncut) for cancer discovery from gene expression profiles. SC3 integrates the spectral clustering (SC) algorithm multiple times into the ensemble framework to process gene

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.2, Special Issue 5, October 2014

expression profiles. Specifically, spectral clustering is applied to perform clustering on the gene dimension and the cancer sample dimension, and also used as the consensus function to partition the consensus matrix constructed from multiple clustering solutions. Compared with SC3, SC2Ncut adopts the normalized cut algorithm, instead of spectral clustering, as the consensus function. Experiments on both synthetic data sets and real cancer gene expression profiles illustrate that the proposed approaches not only achieve good performance on gene expression profiles, but also outperforms most of the existing approaches in the process of class discovery from these profiles.

PradiptaMaji [4] have proposed a supervised attribute clustering algorithm is proposed to find such groups of genes. It directly incorporates the information of sample categories into the attribute clustering process. A new quantitative measure, based on mutual information, is introduced that incorporates the information of sample categories to measure the similarity between attributes. The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive bayes classifier, K nearest neighbour rule, and support vector machine on three cancer and two arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

Jaskowiak, P.A. et al. [5] have investigated the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. Our results support that measures rarely employed in the gene expression literature can provide better results than commonly employed ones, such as Pearson, Spearman, and Euclidean distance. Given that different measures stood out for time course and cancer data evaluations, their choice should be specific to each scenario. To evaluate measures on time-course data, we pre-processed and compiled 17 data sets from the microarray literature in a benchmark along with a new methodology, called Intrinsic Biological Separation Ability (IBSA). Both can be employed in future research to assess the effectiveness of new measures for gene time-course data.

III. PROPOSED METHOD

Gene expression data clustering and classification is one of the important tasks of functional genomics as it provides a powerful tool for studying functional relationships of genes in a biological process. Identifying co-expressed groups of genes represents the basic challenge in gene clustering problem. In this regard, a gene clustering algorithm, termed as robust rough-fuzzy c-means, is proposed judiciously integrating the merits of rough sets and fuzzy sets. The purpose of gene clustering is to group together co-expressed genes which indicate co-function and co-regulation. Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene clustering presents several new challenges and is still an open problem. One of the main problems in gene expression data analysis is uncertainty. Some of the sources of this uncertainty include incompleteness and vagueness in cluster definitions. The pattern identification of gene data is also referred to as a major problem in recent years. Recent studies indicated that effective clustering and classification of gene data can be utilized to tackle most of the difficulties exist in the gene data clustering domain.

In reference to the studies and analysis conducted, we have intended to proposed gene data clustering and classification algorithm. In most of the recent method, either clustering is concentrated or classification is concentrated. Our intension is that, in order to efficiently process the gene data, a clustering and classification algorithm is important .PradiptaMaji and Sushmita Paul [1] have proposed method to cluster gene expression data through a rough fuzzy clustering method. The method concentrates on rough set and fuzzy logic to implement clustering. Inspired from their research, we have intended to propose a method for gene data clustering by modifying the rough fuzzy clustering. In the proposing method, rand initialization of data will be used and the modification on fuzzy objective is supposed to improve the clustering process. The fuzzy objective function is modified in favour of the gene data. In order to avoid the difficulties with rand initialization, one of the recent classification algorithms are incorporated to the proposed

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.2, Special Issue 5, October 2014

methodology. Experimental analysis will be conducted over the proposed method in order to evaluate the efficiency of the method.

IV. CONCLUSION AND FUTURE WORK

Storing and analyzing historical information about the person to improve our Organs system in three ways Firstly, some gene or DNA patterns could be discovered in a long term, such as changes according to the seasons. By using historical data some anomalies could be determined, such as frequent periods of genomic sequence changes. Finally, another promising use of the historical information is the possibility of making the system capable of adapting to the user. With such an extension the users could send information to the system to indicate whether the level of body is proper or not. The result of the system is used to adapt the set of rules and handle some changes in the user's habits in the lifelong. In reference to the studies and analysis conducted, we have intended to propose gene data clustering and classification algorithm. In most of the recent method, either clustering is concentrated or classification is concentrated. Our intension is that, in order to efficiently process the gene data, a clustering and classification algorithm is important. The method concentrates on rough set and fuzzy logic to implement clustering.

REFERENCES

1. PradiptaMaji and Sushmita Paul, "Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 2, pp. 286- 299, 2013
2. Jianyong Sun, Jonathan M. Garibaldi, and Kim Kenobi, "Robust Bayesian Clustering for Replicated Gene Expression Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 5, pp. 1504 - 1514, 2012
3. Zhiwen Yu, Le Li, Jane You, Hau-San Wong, and Guoqiang Han, " SC3: Triple Spectral Clustering-Based Consensus Clustering Frame work for Class Discovery from Cancer Gene Expression Profiles", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 6, 1751 - 1765, 2012
4. PradiptaMaji, "Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1,pp. 127 -140, 2012
5. Jaskowiak, P.A, Campello, R.J.G.B. ; Costa, I.G., " Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 10 ,No 4, pp: 845 - 857, 2013.
6. H. Causton, J. Quackenbush, and A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide. Wiley-Blackwell,
7. 2003.
8. E. Domany, "Cluster Analysis of Gene Expression Data,"J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.
9. P. Maji and S.K. Pal, Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging. John Wiley & Sons, Inc.,2012
10. M.B. Eisen, P.T. Spellman, O. Patrick, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc.
11. Nat'l Academy of Sciences USA, vol. 95, no. 25, pp. 14-863-14-868, 1998.
12. S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, and G.M.Church, "Systematic Determination of Genetic Network Architecture,"
13. Nature Genetics, vol. 22, no. 3, pp. 281-285, 1999.
14. A. Brazma and J. Vilo, "Minireview: Gene Expression DataAnalysis," Federation of European Biochemical Societies Letters,
15. vol. 480, no. 1, pp. 17-24, 2000.
16. P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data," Proc. Second Int'l Workshop InformationProcessing in Cells and Tissues, pp. 203-212, 1998.J. Herrero, A. Valencia, and J. Dopazo,
17. "A Hierarchical Un supervised Growing Neural Network for Clustering Gene Expression
18. Patterns," Bioinformatics, vol. 17, no. 2, pp. 126-136, 2001.L.J. Heyer, S. Kruglyak.
19. L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring ExpressionData: Identification and Analysis of Coexpressed Genes," Genome Research, vol. 9, no. 11, pp. 1106-1115, 1999.
20. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E.Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression ith Self-Organizing Maps: Methods andApplication to Hematopoietic Differentiation,"
21. A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering
22. Gene Expression Patterns," J. Computational Biology, vol. 6, nos. 3/4, pp. 281-297, 1999.
23. E. Hartuv and R. Shamir, "A Clustering Algorithm
24. Based on Graph Connectivity," Information Processing Letters, vol. 76, nos. 4-6, pp. 175-181, 2000.
25. R. Shamir and R. Sharan, "CLICK: A Clustering
26. Algorithm forGene Expression Analysis," Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology, 2000.
27. E.P. Xing and R.M. Karp, "CLIFF: Clustering of
28. High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts," Bioinformatics, vol. 17, no. 1, pp. 306-315, 2001.
29. C. Fraley and A.E. Raftery, "How Many Clusters?
30. Which Clustering Method? Answers via Model-Based Cluster Analysis,"The Computer J., vol. 41, no. 8, pp. 578-588, 1998.