Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Discovering Phenotype Structures in Mining Discriminative Signature

Sudha.C[1], Sathiya.A[2], Divya.K[3], Jones Merlin.E[4]

[1]Department of ME-CSE, Srinivasan Engineering College
[2]Assistant Professor/CSE, Srinivasan Engineering College
[3,4]Department of ME-CSE, Srinivasan Engineering College

Abstract— Data Mining is the process of extracting the information from a dataset and transforms it into an understandable structure. An essential problem in microarray data analysis is to discover phenotype structures. The existing techniques for phenotype structure discovery are singleton discriminability based approach and combination discriminability based approach.The goal is to discovery groups of samples equivalent to different phenotypes (such as disease or normal). Novel sequence dissimilarity is to be proposed for systematic expression values among genes. This is important for the subsequent analysis by the biologists. The sequence model is that only a small number of genes are needed to achieve high phenotype discriminability.A g* sequence model to characterize the phenotype structure.This property helps to improve the robustness of the proposed model and enables to identify the highly discriminative signatures with only a small number of genes.

Keywords— *Data mining, bioinformatics, microarray data*

I. INTRODUCTION

Data Mining has great potential for exploring the meaningful and hidden patterns in the data sets at the medical domain. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques is a remedy to this situation. Data mining functions include clustering, classification, prediction, and associations. One of the most important data mining applications is that of mining association rules. Association rules, introduced in 1993, are used to identify relationships among a set of items in databases. These relationships are not based on inherit properties of the data themselves, but rather based on co-occurrence of the data items. Emphasis in this research work is analysis of medical data. Medical profiles such as patient name, age, sex, disease name, address, time, date, etc., can be used to mining the frequent disease of patients in different geographical area at given time period.

Focus on the topic of learning phenotype structure discovery. The introduction of DNA microarray technologies has reformed the experimental learning of gene expression. Thousands of genes are routinely explored in a parallel way, and the expression levels of their transcribed mRNA are stated. By repeating such tests under altered circumstances(e.g.,differentpatients,differenttissues ,or variation of the cells' environment), data from tens to hundred softest scan be collected. The investigation of the subsequent huge datasets poses frequent algorithmic tasks. Sofar, the chief method occupied for examining gene expression data is clustering(and variants there of). There is a very huge form of works on clustering in over all and on applying clustering methods to gene expression data in specific.Gene expression data are typically organized in a matrix, with each row equivalent to one gene, each column to one state, and every entry in the matrix signifying the expression level of a gene below an exact state. In the area of gene expression study, an important research problem is to discover sub matrix patterns in the gene expression matrix.

The existing methods for phenotype structure discovery can be classified into two types: singleton discriminability based approach and combination discriminability based approach. The existing singleton or combination discriminability based methods cannot separate the two phenotypes. The existing bi-clustering algorithms, the order-preserving sub matrix (OPSM) model also combines the order information of the gene expression values.

II. RELATED WORK

First process is the input selection. Select the input dataset with different data streams (attributes). Then input dataset has been loaded into the database. After the dataset has been loaded into the The datasets are preprocess including eliminate null symbols. The values are calculated into standard deviation. Then compute the relevance values. These values are formed into sequence model. The datasets are classified into cluster1 and cluster2 values for to predict the malignant and benign values. Finaly the performance are evaluated.

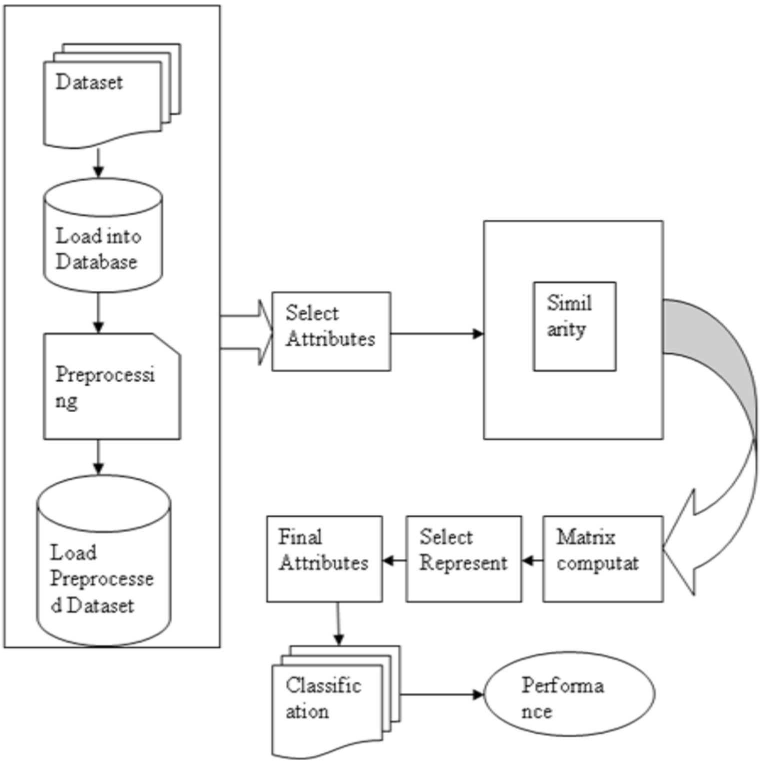

Figure 1 system architecture

An important task in microarray data analysis is phenotype structure discovery . Given a microarray dataset of m samples and n genes, a phenotype structure refers to a group of “blocks” (or submatrices), each of which consists of a subset of samples and a subset of genes such that the samples from all the blocks make up a partition of m samples, and the samples in a block correspond to a phenotype (such as a disease subtype); and the gene expression pattern within a block can be used as the signature to distinguish this group of samples from others. The genes in a signature may suggest the potential biomarkers related the disease. In particular, phenotype structure discovery is an unsupervised learning problem. It is more challenging than the problem of biomarker selection with known class labels

A microarray data set D is an m _ n matrix, with m samples S={s1, s2, . . . , sm} and n genes G={ g1, g2, . . . , gn }. A real value dij in D represents the expression value of gene gj on sample si. A microarray data set D is an mn matrix, with m samples S ¼f s1, s2, ..., sm and n genes G ¼f g1, g2, ..., gn. A real value dij in D represents the expression value of gene gj on sample si. An example microarray data set with nine genes and four samples. Microarray data are often noisy. We introduce the concept of equivalent dimension group which represents a set of genes with similar expression values. An EDG encloses a group of genes with the similar expression values together. The sequences of genes in which any pair of genes are not contained by the same EDG is robust to noise the group threshold. Moreover, only considering such genes makes the maximum size of the sequences is far less than that of the original ones. Thus, the time taken by sequence mining is greatly reduced while keeping the significant results. For a sample si, a sliding window approach can be applied to find all EDGs. First, all genes are sorted by their expression values in ascending order. Second, we slide a window from left to right.

A block (or submatrix) is the basic element of a phenotype structure, which consists of a subset of samples and the corresponding p-signature. Thus, phenotype structure discovery can be naturally divided into the following three components: candidate p-signatures generation, block derivation from candidate p-signatures, and quality test of block combinations.

In biology community, discriminative sequential patterns involving the ordered gene expression values have been shown effective in distinguishing phenotypes. Such patterns have an intuitive biological interpretation. Complex diseases often involve the cooperation of multiples genes. These genes work together as a system to keep the cell in a specific state, for example, disease or normal. In such a state, some special interrelationship among genes will exhibit. Once such relationship is disrupted, the state may change, for example, from normal to disease. Another advantage of the sequence model is that only a small number of genes are needed to achieve high phenotype discriminability.

III. SCOPE OF THE PAPER

A g*-sequence model to characterize the phenotype structure. It introduces the concept of significant chain to ensure that there is a significant difference between the expression values of any pair of genes. This property helps to improve the robustness of the proposed model, and enables to identify highly discriminative signatures with only a small number of genes. To measure the quality of a candidate phenotype structure, we propose a novel sequence dissimilarity metric, namely projection divergence. Based on this metric, the difference between a pair of blocks. (submatrices) can be quantified based on the discriminative power of the signatures within the blocks.

The problem of phenotype structure discovery is NP-complete. Given n genes, the total number of subsequences (candidate signatures).We prove that the prohibitively large search space can be reduced to a much smaller scale. An efficient algorithm, FINDER, is developed to find the optimal phenotype structure. By incorporating the cross projection into a progressive exploring framework, candidate phenotype structures are searched in a quality-guaranteed way. We conduct extensive experiments on both real and synthetic data sets. The results show that FINDER dramatically improves the efficiency of the mining process. With very few genes, the discovered signatures are able to unravel phenotype structures that are both statistically and biologically significant.

Phenotype structure discovery involves two key elements ,i.e., the partition of samples and the selection of genes. Correspondingly, we conducted two sets of experiments to compare FINDER with ESPD and HARP. In the first set of experiments, precision, recall, and accuracy are used to evaluate the correctness of the partition of samples, the computations of which follow a common evaluation framework proposed in]. In the second set of experiments ,we use gene selection rate (GSR, the ratio between the number of the selected genes and the total number of genes) to evaluate the succinctness of the selected genes.

The proposed algorithm is evaluated in several experiments to simulate scenarios, involving different types of changes. In the following sections, we describe all of the used datasets. In the top subfigure, 4 genes are expressed over 25 samples. Samples 16 are cancerous (labeled as “C”) and samples 25 are normal (labeled as “N”). In the bottom subfigure, another set of three genes are expressed over the same set of samples. The existing singleton or combination discriminability based methods cannot distinguish the two phenotypes. Since most genes are of similar average expression values in the two phenotypes, they will not be selected by the singleton approach. Moreover, all genes are expressed in both phenotypes. Thus, the combination approach based on the co occurrence of genes will not select them either. Both of the methods ignore the hidden inter- relation among genes.

The gene order over the samples of cancerous phenotype “C” is always gene4, gene3, gene2, gene1. Such order is disturbed in normal phenotype “N”, the gene order in normal phenotype “N” is gene5, gene6, gene7, while in cancerous phenotype “C” such order does not exist. Based on the ordered expression values, a perfect phenotype structure (consisting of the two shadowed “blocks”) is identified. In biology community, discriminative sequential patterns involving the ordered gene expression values have been shown effective in distinguishing phenotypes. Such patterns have an intuitive biological interpretation. Complex diseases often involve the cooperation of multiples genes. These genes work together as a system to keep the cell in a specific state, for example, disease or normal. In such a state, some special interrelationship among genes will exhibit. Once such relationship is disrupted, the state may change, for example, from normal to disease. Another advantage of the sequence model is that only a small number of genes are needed to achieve high phenotype discriminability. Intuitively, this is because it exploits more information ignored by other models, i.e., the interrelation among the genes beyond the cooccurrence. Finding fewer but more powerful discriminative genes is crucial for interpretation and validation in the subsequent wet-lab experiments . Biclustering algorithms have been studied to analyze gene expression data. Among the existing biclustering algorithms, the order-preserving submatrix (OPSM) model also incorporates the order information of the gene expression values. An OPSM consists of a subset of genes and a subset of experimental conditions such that the expression profiles of the genes.

IV. CONCLUSION

A g* sequence model to find extremely accurate phenotype structure with a small number of genes. The problem of phenotype structure discovery is NP-complete and develop a progressive exploring strategy to tackle the computational challenge. In the FINDER algorithm, a novel sequence dissimilarity measurement and a cross projection approach allow to try discovering candidate phenotype structures in a quality-guaranteed technique. Various effective techniques are developed to additional improve the efficiency.General experimental outcomes on real and synthetic datasets show that our technique intensely improves the accuracy of the discovered phenotype structure even though using much less genes compared to the existing methods.

REFERENCES

[1] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church,(1999),"Systematic Determination of Genetic Network Architecture," Nature Genetics, vol. 22, pp. 281-85,.

[2] M. Eisen, P. Spellman, P. Brown, and D. Botstein,(1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Nat'l Academy of Sciences USA, vol. 95, pp. 14 863-68.

[3] A. Alizadeh, (2000), "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," Nature, vol. 403, pp. 503-11.

[4] C. Tang, A. Zhang, and M. Ramanathan,(2004), "ESPD: A Pattern Detection Model Underlying Gene Expression Profiles," Bioinfor-matics, vol. 20, no. 6, pp. 829-838.

[5] J.R. Nevins and A. Potti,(2013), "An Interactive approach to Gene Expression Data," Nature Rev.Genetics, vol. 8, no. 8, pp. 601-609.

[6] K.Y.Yi p , D.W.Cheung , and M.K . Ng ,(1999),"Harp: A Practical Projected Clustering Algorithm," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1387-1397[7] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,"Science, vol. 286, pp. 531-537.

[8] J. Luo et el.,(2001), "Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling."Cancer Research, vol. 61, no. 12, pp. 4683-8.

[9] U. Alon et al.,(1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proc. Nat'l Academy of Sciences USA, vol. 96, no. 12, pp. 6745-6750.

[10] M. Xiong, X. Fang, and J. Zhao,(2001),"Biomarker Identification by Feature Wrappers," Genome Research, vol. 11, no. 11, pp. 1878-1887.

[11] J. Liu and W. Wang,(2003), "Op-Cluster: Clustering by Tendency in High Dimensional Space," Proc. IEEE Third Int'l Conf. Data Mining (ICDM), pp. 187-194.

[12] Y. Cheng and G.M. Church,(2000), "Biclustering of Expression Data,"Proc. Int'l Conf. Intelligent System Moleculer Biology, pp. 93-103.

[13] X. Xu, Y. Lu, and A. Tung,(2006), "Mining Shifting-and-Scaling Co-Regulation Patterns on Gene Expression Profiles," Proc. 22nd Int'l Conf. Data Eng. (ICDE '06), pp. 89-100.

[14] A. Ben-Dor, B. Chor, R.M. Karp, and Z. Yakhini,(2002), "Discovering Local Structure in Gene Expression Data: the Order-Preserving Submatrix Problem," Proc. Sixth Ann. Int'l Conf. Computational Biology (RECOMB), pp. 49-57.

[15] Q. Fang, W. Ng, and J. Feng, "Discovering Significant Relaxed Order-Preserving Submatrices,(2010)," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '10), pp. 433-442.

[16] S. Dong et al.,(2005),"Histology-Based Expression Profiling Yields NovelPrognostic Markers in Human Glioblastoma," J. Neuropathology and Experimental Neurology, vol. 64, no. 11, pp. 948-955.

[17] H. Liu and H. Motoda,(2007),"Computational Methods of Feature Selection). CRC Press.

[18] D. Zuckerman, "On Unapproximable Versions of Np-Complete Problems,(1996)," SIAM J. Computing, vol. 25, no. 6, pp. 1293-1304.