Shalom David, 2021, 9:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Learning Phenotype Structure Using Maximal Phenotype Alteration

PG Student Reny Sam, Asst. Prof. Mr. Shalom David
Department of Computer Science
Christ Nagar College
Maranalloor, Trivandrum

Abstract- Data mining with microarray technologies can be used for the extracting hidden predictive information and are enabled to continuously monitor the expression rates of all genes. Discovering and studying phenotype structures have become an important problem in the field of microarray data analysis. There are mainly two goals 1) various samples related to various phenotypes such as normal or disease are to be found. 2) For various groups of samples, find the signature or representative expression pattern that will differentiate one group from others. There are several methods proposed, however, some common drawbacks could be identified such as the signatures selected can have a large amount of genes with very low discriminative power. In this paper, the g*-sequence model is improved and updated to address the limitations thereby expression values which are ordered among genes can be utilized profitably, the proposed sequence model could be seen more robust to noise and allows to find the signatures with higher discriminative power using very less genes. An efficient algorithm, Finder with markovian policies is developed, which contains three steps: 1) trivial g* sequences identifying, 2) discovering phenotype structure, and 3) refinement. To further improve efficiency effective pruning is carried out. Real genes and synthetic gene sets could be used to evaluate the performance of finder with markovian policies. The results proves that the approach could be used for phenotype structure discovery with high accuracy and detects signatures with high discriminative power, the orders of magnitude is faster than other alternatives

Keywords: - Data mining, bioinformatics, microarray data.

I. INTRODUCTION

DNA microarray can be explained as microscopic DNA spots contained to particular solid areas. Large amount of genes expression levels or rates could be measured in this way.

Phenotype is an organism's observable character istics/traits such as morphology; biochemical, physiological properties, and a phenotype mostly get influences of environmental factors or from expression of an organism's genes. The variation in phenotypes could be noted as a prerequisite for evolution by natural selection, genetic structure of an organism could be affected by the contribution of phenotype. Phenomena could be referred to collection of traits and phenomics as the study for such collection. The phenotype structure could be studied with many efficient algorithms. Although there exit many methods, these methods lack in scalability, robustness, time and computational complexity.

Maximal phenotype alteration, an efficient method with suboptimal policies is introduced to overcome these problems. The G* sequence model with the combination of the phenotype alteration technique and policies makes the method more efficient and effective. The method will be able to work faster by

several orders of magnitude compared to other algorithms. Datasets such as real and synthetic could be used for the study of discovering phenotype structures.

RELATED WORKS

Mohammad Mahdi R describes the gene interactions modelled using gene regulatory networks and thus the effects of intervention and its strategies are studied and derived, The phenotype are characterized by long run behaviour (steady state distribution) of the network, two control approaches which are considered to be external are used to shift the steady-state mass of a GRN: user-defined cost function for which desirable shift of the steady-state mass is a by-product, heuristics to design a greedy algorithm, but neither of these approach provides an optimal control policy. Gene regulatory network are important part of translational medicine, whose ultimate aim is to develop solutions on disruption or mitigation of aberrant gene function contributing to Pathology of a disease, two basic intervention approaches are used in context of probability Boolean networks, external control and structural intervention, according to the value flipping of a specific or possibly (more than one) called control gene an external defined, finite – state markov chain defines dynamic behaviour of PBN.

Synthetic networks with seven genes. For shifting steady – state mass from undesirable to desirable states a linear programming approach is used, for unconstrained and constrained optimization basic linear programming structure is used, the amount of mass that may be shifted to 'ambiguous' states depend on constraints on the optimization limit, and clinically significant subtypes of cancer. There were various steps taken to perform necessary operations .The first method and step taken was to construct a specialized DNA array, in this method a specialized lymph chip is designed by selecting genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology or cancer, after this the analysis of gene expression in lymphoid malignancies is done, the microarrays is used to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL,FL and CLL.

A hierarchical clustering algorithm is used to group genes on the basis of similarity in the pattern with which their expression varied over all samples. The same clustering method is used to group tumour and cell samples on the basis of similarities in their expression of these genes. The data are in matrix format, with each row representing all the hybridization results for a single cDNA element of the array, and each column representing the measured expression levels for all genes in a single sample.

To visualize the results, the expression level of each gene (relative to its median expression level across all samples) is represented by a colour, with red representing expression greater than the mean, green representing expression less than the mean, and the colour intensity representing the magnitude of the deviation from the mean.

The next step is to identify the tumour phenotypes with gene expression patterns, once this process is completed the discovery of DLBCL subtype is carried, the structure of the hierarchical dendrogram indicates the gene expression patterns in DLBCLs might be inhomogeneous. Three branches of the dendrogram captured most of the DLBCLs with only three outlying samples. The position of any given DLBCL sample in the dendrogram is determined in a complicated fashion by the influences of several distinct biological themes that are rejected in the expression pattern.

The search excluded genes that were readily assigned to the proliferation, T-cell and lymph-node signatures in order to focus attention on more subtle intrinsic molecular features of this group of tumours. Hierarchical clustering is used to reorder the set of 2,984 genes while maintaining the order of the DLBCL cases, as is evident a cluster of genes could be recognized on the basis of their elevated expression in the activated B-like DLBCLs, as compared with GC B-like DLBCLs.

It is important to note that considerable gene expression heterogeneity exists within each subgroup and that no single gene in either of these large clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy. Rather, patients assigned by this method to either DLBCL subgroup shared a large gene expression program that distinguished them from the other subgroup. The final step defines the prognostic categories by DLBCL gene expression subgroups. A clinical indicator of

prognosis, the International Prognostic Indicator (IPI), is used to define prognostic subgroups in DLBCL. The indicator takes into account the patient's age, performance status, and the extent and location of disease. As suspected, within the patient population a low IPI score (0±2) identified patients with better overall survival as compared with patients with a high IPI score (3±5).

After determining whether molecular definition of DLBCL subgroups could add to the prognostic value of this clinical indicator of prognosis. Considering only patients with low clinical risk, as judged by the IPI, patients in the activated B-like DLBCL group had a distinctly worse overall survival than patients in the GC B-like DLBCL group (P, 0.05). Thus, the molecular dissection of DLBCL by gene expression profiling and the IPI apparently identify different features of these patients that influence their survival. The genomic-scale view of gene expression in cancer provides a unique perspective on the development of new cancer therapeutics that could be based on a molecular understanding of the cancer phenotype.

The study shows that the two DLBCL subgroups differentially expressed entire transcriptional modules composed of hundreds of genes, many of which could be expected to contribute to the malignant behavior of the tumor. This observation suggests that successful new therapeutics might be aimed at the upstream signal-transuding molecules whose constitutive activity in these lymphomas leads to expression of pathological transcriptional programs.

Chun Tang [2] proposes a new model called empirical sample pattern detection (ESPD) to delineate pattern quality with informative genes. By integrating statistical metrics, data mining and machine learning techniques, this model dynamically measures and manipulates the relationship between samples and genes while conducting an iterative detection of informative space and the empirical pattern. The problem of unsupervised sample pattern detection by developing a novel analysis model called empirical pattern detection (ESPD) which includes a series of statistics-based metrics and iterative adjustment.

A formalized problem statement of ESPD of sparse high-dimensional datasets is proposed. Major differences from traditional clustering or recent subspace clustering problems are elaborated, a series of statistics-based metrics incorporated in unsupervised empirical pattern discovery are introduced. These metrics delineate local pattern qualities to coordinate between sample pattern discovery and informative genes selection, an iterative adjustment algorithm is presented to approach the optimal solution.

The method dynamically manipulates the relationship between samples and genes while conducting an iterative adjustment to approximate the informative space and the empirical pattern simultaneously, an extensive experimental evaluation over real datasets is presented. It shows that our method is both effective and efficient and outperforms the existing methods.

Given a data matrix and the number of samples' phenotypes, the goal is to find mutually exclusive groups of the samples matching their empirical phenotypes and to find the set of genes which manifests the meaningful pattern.

P. M. Booma, S. Prabhakaran, and R. Dhanalakshmi [6] proposed a model to monitor higher rate of expression levels between genes. The biological association between genes is measured simultaneously using proximity measure of improved Pearson's correlation (PCPHC). Experimental studies show that the PCPHC model outperforms all the current models, and, importantly, it leads to the discovery of more quality patterns. The experimental result of PCPHC model attains the improved gene expressional data, minimal execution time.

Shuzhong Zhang, Kun Wang, Bilian Chen, and Xiuzhen Huang [18] proposed a framework to study the co-clustering of gene expression data. This framework is based on a generic tensor optimization model and an optimization method termed Maximum Block Improvement (MBI). Not only can this framework be applied for co-clustering gene expression data with genes expressed at different conditions represented in 2D matrices, but it can also be readily applied for co-clustering more complex high-dimensional gene expression data with genes expressed at different tissues, different development stages, different time points, different stimulations, etc. It is flexible that it poses no difficulty at all to incorporate a variety of clustering quality measurements.

R. Das, D. K. Bhattacharyya, and J. K. Kalita [20] were proposed a system which presents two clustering methods: the first one uses a density-based approach (DGC) and the second one uses a frequent item set mining approach (FINN). The clusters obtained by DGC have been validated using several cluster validity measures over six microarray data sets. The regulation based cluster expansion also overcomes the problem of maintaining the pattern information usually linked with the different clustering approaches due to traditional similarity measures.

In FINN, the frequent item set generation step gives the innermost or the fine clusters from the gene expression data and the shared neighbour clustering approach gives the final clusters in the dataset. Both the methods use a novel dissimilarity measure discussed in the work.

Geetha. T, Michael Arock [17] their work presents enhanced hierarchical clustering algorithm for gene expression data sets. In the previous works, database scanning and distance matrix calculation are needed for all iterations. This method reads the database and finds distance matrix only once, which reduces the amount of time. Also, our method requires the minimum space, as the lower triangular distance matrix can be represented in single dimensional array, even when large databases are used. And, we represent the cluster results as a binary tree which gives clear grouping. Cut distance is used to find the number of clusters and clustered objects.

Kwon Moo Lee and Ju Han Kim [21] proposed the heuristic global optimization method, Deterministic Annealing (DA), to the same clustering method. In DA approach, the cost function is locally minimized subject to a constraint on a given randomness (Shannon entropy), controlled by 'temperature' that is gradually lowered. As the temperature goes slowly down to zero, obtain the best binary partitioning by the analogy of statistical physics in annealing process.

Even though global optimization approach produces the high quality clustering results, in general, it requires much computational cost. But, since the speed of DA algorithm depends on that of the local optimizer, if employ high performance local optimization technique, the computation cost can be substantially reduced. It was also demonstrated that error-prone objects can be identified by monitoring the annealing process, which can be applied to increase the quality of clustering analysis. Here used the standard multidimensional local optimization technique which incorporates one dimensional line search method.

III. MAXIMAL PHENOTYPE ALTERATION

To improve efficiency in phenotype structure discovery (Yuhai Zhao. 2014) and to reduce the computation cost of the system, a new scheme called "maximal phenotype alteration" is proposed (Edward R. 2013).

From this proposed system the performance of discovery of the phenotype structure is improved. An intervention policy that maximally shifts the long run probability mass of undesirable states to desirable states. This objective function essentially concerns the long-run behavior of the occupation measures marginalized over the actions. Thus, policy space to MS (Markov stationary policies) can be limited without loss of generality.

Let A = A (j) = {0, 1} for all j S. If policy MS, then the amount of shift in the aggregated probability of undesirable states for a probabilistic Boolean networks (PBNs) controlled under is defined as;

$$\Delta\pi_u(\mu) = \sum_{j \in u} \pi_j - \sum_{j \in u} \pi_j\ (\mu)$$

Where $\boldsymbol{\pi}$ and $\boldsymbol{\pi}(\boldsymbol{\mu})$ are the unique vectors of the invariant probability measure for the Markov chains governed under the (transition probability matrix) TPMs P and Q(μ), respectively. In general, -1 ≤ $\pi_u(\mu)$ ≤ 1 and the goal is to maximize it. A block (or sub matrix) 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. By using this maximal phenotype alteration process, we can obtain the higher efficiency in phenotype structure discovery and as well as we can reduce the computational overhead of the system.

International Journal of Science, Engineering and Technology

An Open Access Journal

IV. ARCHITECTURE

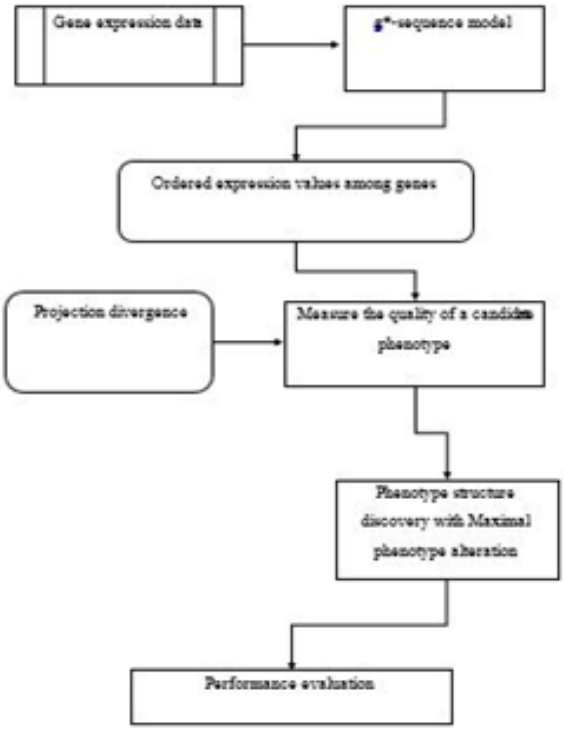

Fig 1. Architecture for learning phenotype structure discovery using maximal phenotype alteration.

V. PERFORMANCE METRICS AND RESULT ANALYSIS

The performance offered by existing system and the proposed system can be compared by evaluating the parameters such as accuracy, error rate, time complexity and number of selected genes, Based on the comparison and the results from the experiment show that the proposed approach works better than the existing system.

To evaluate robustness to noise, we artificially added noise to some data sets. Since we did not know a priori the amount of noise originally present in each data set, as a first step, we chose four data sets such as breast cancer dataset, yeast dataset, diabetes dataset and tumor dataset in which differences among the performances of the distances were minimal, i.e., data sets in which different distances provided the closest results without the presence of noise.

With this selection, we intended to provide a fair starting point and comparison among the distances as the noise is added.

1. Accuracy: Accuracy can be calculated from formula given as follows:

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Where, **TP (True Positive):** If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP).

2. TN (True Negative): A true negative (TN) has occurred when both the prediction outcome and the actual value are n in the number of input data.

3. FP (False Positive): If the outcome from a prediction is p and the actual value is n then it is said to be a false positive (FP).

4. FN (False Negative): False negative (FN) is when the prediction outcome is n while the actual value is p.

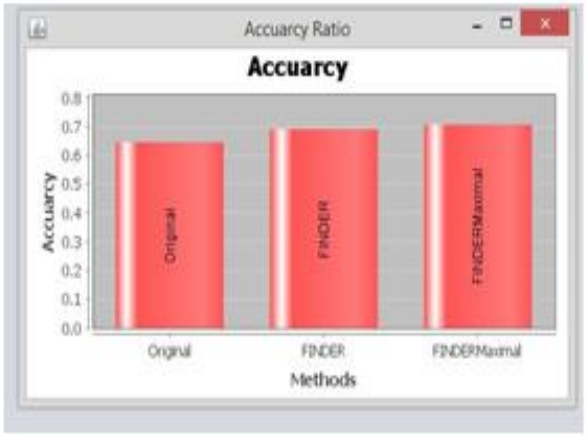

Fig 2. Accuracy performance comparison with noise in dataset.

In the above graph, we are comparing the accuracy rate of the proposed system (with noise) with the existing technique (with noise). For this experiment, we are adding the noise to the dataset such that breast cancer dataset, yeast dataset, diabetes dataset and tumor dataset. Accuracy rate is mathematically calculated by using formula. As usual in the graph X-axis will be number of proximity measures methods such as existing system i.e., few datasets and proposed system such as many datasets and Y-axis will be accuracy rate.

From the graph we can easily understand that the proposed system has higher accuracy rate which is taken the output result. From view of this accuracy comparison graph we obtain conclude as the

International Journal of Science, Engineering and Technology

An Open Access Journal

proposed algorithm has more effective in accuracy rate performance compare to existing algorithms.

5. Error Rate: Error rate can be calculated from formula given as follows

Error Rate = (FP+FN)/ (TP+TN+FP+FN)

Fig 3. Error rate performance comparison with noise in dataset.

In the above graph, we are comparing the error rate of the proposed system (with noise) with the existing technique (with noise). For this experiment, we are adding the noise to the dataset such that breast cancer dataset, yeast dataset, diabetes dataset and tumor dataset. Error rate is mathematically calculated by using formula. As usual in the graph X-axis will be number of methods such as existing system i.e., few datasets and proposed system such as many datasets and Y-axis will be error rate.

From the graph we can easily understand that the proposed system has very low error rate which is taken the output result. From view of this error comparison graph we obtain conclude as the proposed algorithm has more effective in error rate performance compare to existing algorithms.

VI. CONCLUSION

The proposed system is introducing the novel approach of Maximal phenotype alteration to the model of phenotype structure discovery. The proposed system improves the effectiveness of phenotype structure discovery process and it well reduced the cost of computation of the system. The proposed system has high effectiveness than other existing systems.

The proposed phenotype structure discovery approach is used to discover the statistical significant phenotype structures with higher accuracy and fewer genes. Extensive experimental results on real and synthetic data sets show that the method dramatically improves the accuracy of the discovered phenotype structure (in terms of statistical and biological significance) while using much less genes compared to the existing methods.

REFERENCES

[1] Alizadeh, (2000) "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling."

[2] C. Tang., A. Zhang., and M. Ramanathan., (2004) "ESPD: A Pattern Detection Model Underlying Gene Expression Profiles."

[3] Alexander Schliep, Ivan G. Costa, Christine Steinhoff, and Alexander Schonhuth (2005)," Analyzing Gene Expression Time-Courses".

[4] Anbupalam Thalamuthu (2006)," Evaluation and comparison of gene clustering methods in microarray analysis".

[5] Gad Getz, Erel Levine, and Eytan Domany (2000)," Coupled two-way clustering analysis of gene microarray data".

[6] P. M. Booma, S. Prabhakaran, and R. Dhanalakshmi (2010)," An Improved Pearson's Correlation Proximity– Based Hierarchical Clustering for Mining Biological Association between Genes".

[7] Janez Demsar (2006)," Statistical Comparisons of Classifiers over Multiple Data Sets". Ka Yee Yeung, Mario Medvedovic and Roger E Bumgarner (2003)," Clustering gene expression data with repeated measurements".

[8] Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir and Mehdi Pirooznia, Jack Y Yang, Mary Qu Yang and Youping Deng (2008)," A comparative study of different machine learning methods on microarray gene expression data".

[9] Nadia Bolshakova, Anton Zamolotskikh and Pádraig Cunningham (2006)," Comparison of the Data-based and Gene Ontology-based Approaches to Cluster Validation Methods for Gene Microarrays".

[10] Raja Loganantharaj, Satish Cheepala and John Clifford (2006)," Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression".

International Journal of Science, Engineering and Technology

An Open Access Journal

[11] Rosy Das, D. K. Bhattacharyya and Jugal K. Kalita (2009)," Clustering Gene Expression Data using a Regulation based Density Clustering".
[12] Muhammad Rukunuddin Ghalib, Rittwika Ghosh, Priti Sasmal Udisha Pande (2013), "Microarray Gene Expression Data Analysis Using Enhanced k-means Clustering Method".
[13] Wenming Cao, Shoujue Wang (2005)," Application of Geometrical Learning for Similarity Index in Clustering DNA Microarray Data".
[14] Young Sook Son, Jangsun Baek (2008)," A modified correlation coefficient based similarity measure for clustering time-course gene expression data".
[15] Travis J. Hestilow1 and Yufei Huang (2009)," Clustering of Gene Expression Data Based on Shape Similarity".
[16] Daxin Jiang, Chun Tang, and Aidong Zhang (2004)," Cluster Analysis for Gene Expression Data: A Survey".
[17] Geetha.T, Michael Arock (2010)," Enhanced Hierarchical Clustering for Gene Expression data".
[18] A New Framework for Co-clustering of Gene Expression Data(2006), by Shuzhong Zhang, Kun Wang, Bilian Chen, and Xiuzhen Huang
[19] Susmita Datta and Somnath Datta (2003)," Comparisons and validation of statistical clustering techniques for microarray gene expression data".
[20] R. Das, D.K. Bhattacharyya, and J.K. Kalita (2010)," Clustering gene expression data using an effective dissimilarity measure".
[21] Kwon Moo Lee, Tae Su Chung and Ju Han Kim(2003)," Global Optimization of Clusters in Gene Expression Data of DNA Microarrays by Deterministic Annealing".