An Analysis on Clustering Based Gene Selection and Classification for Gene Expression Data

K. Sathishkumar[1], Dr. V. Thiagarasu[2], Dr. E. Balamurugan[3]
[1] Lecturer of IT/IS, Bluecrest College, Accra North, Ghana, West Africa
[2]Associate Professor of Computer Science, Gobi Arts & Science College(Autonomous), Gobi, TN, India
[3]Associate Professor of IT/IS, Bluecrest College, Accra North, Ghana, West Africa

Abstract - Due to recent advances in DNA microarray technology, it is now feasible to obtain gene expression profiles of tissue samples at relatively low costs. Many scientists around the world use the advantages of this gene profiling to exemplify complex biological circumstances and diseases. Microarray techniques that are used in genome-wide gene expression and genome mutation analysis help scientists and physicians in understanding of the patho-physiological mechanisms, in diagnoses and prognoses, and choosing treatment plans. DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during the important biological processes and across collections of related samples. Elucidating the patterns hidden in gene expression data offers a tremendous prospect for an enhanced understanding of functional genomics. 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 the data mining process to reveal natural structures and identify the most interesting patterns in the original data. This work presents an analysis of several clustering algorithms proposed to deals with the gene expression data effectively. The existing clustering algorithms like Support Vector Machine (SVM), K-means algorithm and Evolutionary algorithm etc are analyzed thoroughly to identify the advantages and limitations. The performance evaluation of the existing algorithms is carried out to determine the best approach. In order to improve the categorization performance of the best approach in terms of Accuracy, Convergence Behavior and processing time, a hybrid clustering based optimization approach has been proposed.

Keywords - Microarray technology, Gene expression data, clustering.

1. INTRODUCTION

High-throughput techniques have become a primary approach to gathering the biological data. These data can be used to the actual clinical application of gene expression data analysis and guide development of drugs and other research [1] [2]. Generally, the data mining tasks comprise Classification, Regression, Clustering, Rule generation, Discovering association rules, Summarization, Dependency modeling and Sequence analysis. DNA microarray technology is also the field, which uses the data mining methods [3]. Also, the molecular biologists face the challenges in discovering the essential knowledge from this kind of enormous volume of data [4]. However, the deluge of data contains an overwhelming amount of unknown information about the organism under study. Therefore, clustering is a common first step in the exploratory analysis of high-throughput biological data. Although a number of clustering methods have been proposed, they incur problems such as Quality and Efficiency [5]. In the aspect of efficiency, most clustering algorithms aim to produce the best clustering result based on the input parameters. Clustering has been used in a number of applications such as engineering, biology, medicine and data mining.

Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes. DNA microarrays are emerged as the leading technology to measure gene expression levels primarily, because of their high throughput. Results from these experiments are usually presented in the form of a data matrix in which rows represent genes and columns represent conditions [6]. Each entry in the matrix is a measure of the expression level of a particular gene under a specific condition. Analysis of these data sets reveals genes of unknown functions and the discovery of functional relationships between genes [7].

The goal of clustering is to determine a natural combination in a group of patterns, points, or objects, without knowledge of any class labels. Clustering is widespread in any discipline that involves analysis of multivariate data. It is, of course, impractical to exhaustively list the numerous uses of clustering techniques. In the background of the human genome development, new technologies are emerged, and it facilitates the parallel execution of experiments on a large number of genes at the same time. Hence it is called as DNA microarrays, or DNA chips, which constitute a prominent example. This technology aims at the measurement of mRNA levels in particular cells or tissues for many genes at once. To this end, single strands of balancing DNA for the genes of interest can be immobilized on spots arranged in a grid on a support which will typically be a glass slide, a quartz wafer, or a nylon membrane. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample [8].

Normally, microarray experiments create a large number of datasets with expression values for thousands of genes but still not more than a few dozens of samples, thus very accurate arrangement of tissue samples in such high dimensional problems is a complicated task [9]. Also, there is a high redundancy in microarray data as well as several genes have irrelevant information for exact clustering of diseases or phenotypes [10]. Therefore, a robust clustering method is indispensable to retrieve the gene information from the microarray experimental data.

2. RELATED WORK

1.1 BACKGROUND STUDY

Classification and clustering are two major tasks in gene expression data analysis. Classification is concerned with assigning memberships to sample based expression patterns, and clustering aims at finding new biological classes and refining existing ones [11]. To cluster and/or recognize patterns in gene expression datasets, dimension problems are encountered. Typically, gene expression datasets consist of a large number of genes (attributes) but a small number of samples (tuples).

K-means is a form of partition-based clustering technique mainly utilized in clustering gene expression data [13]. K-means is well known for its simplicity and speed. It performs quite well on large datasets. However, it may not provide the identical result with each run of the algorithm. It is observed that, K-means is very good at handling outliers but its performance is not satisfactory in detecting clusters of random shapes.

In presents an attribute clustering method is able to group genes based on their interdependence so as to meaningful patterns from the gene expression data [14][23]. It can be used for gene grouping, selection, and classification. The partitioning of a relational table into attribute subgroups allows a small number of attributes within or across the groups to be selected for analysis. By clustering attributes, the search dimension of a data mining algorithm is reduced. The reduction of search dimension is especially important to data mining in gene expression data because such data typically consist of a huge number of genes (attributes) and a small number of gene expression profiles (tuples). Most data mining algorithms are typically developed and optimized to scale the number of tuples instead of the number of attributes. By applying the Attribute Clustering Algorithm (ACA) to gene expression data, meaningful clusters of genes are discovered.

The grouping of genes based on attribute interdependence within the group helps to capture different aspects of gene association patterns in each group. Significant genes are selected from each group then convert the useful information for gene expression classification and identification. To evaluate the performance. It can apply to two well-known gene expression data sets and compared the results with those obtained by other methods. Haiying et al., addresses the two algorithms self-organizing map and hierarchical clustering for SAGE (Serial Analysis of Gene Expression) data analysis. SOM performs the pattern discovery and visualization for SAGE data in a more effective way. It estimates the optimal number of clusters hidden in SAGE data. Subsets of genes are interdependent with each other. It identifies a set of related genes with similar samples yields tree like structure, which makes difficult to identify the functional groups [15].

A clustering technique (GenClus) for gene expression data which can also handle incremental data. It is designed based on density based approach. It retains the regulation information which is also the main advantage of the clustering. It uses no proximity measures and is therefore free of the restrictions offered by them. GenClus is capable of handling datasets which are already updated incrementally. Experimental results show the efficiency of GenClus in detecting quality clusters over gene expression data. In this approach improves the cluster quality by identifying sub-clusters within big clusters. It is compared with some well-known clustering algorithms and found to perform well in terms of the z-score cluster validity measure [16].

The feature selection and classification approach is a significant problem in microarray gene expression analysis due to huge number of variables and small number of experiential conditions. For diagnosing the disease, classifier performance has through impact on final results. The author has introduced a new-fangled method of gene selection and classification through nonlinear kernel support vector machine (SVM) based on process recursive performance removal (RFE). It is recognized experimentally that this method has better wide-ranging performance than other linear classification approaches like Linear Kernel SVM and Fisher Linear Discriminant Analysis (FLDA), they perform good than when compared to various non-linear classification approaches like Least Square SVM (LS-SVM) using non-linear kernel. In the testing, leave-one-out algorithm is also used to test the classifiers overview performance [17].

Mohr et al., has described an efficient approach for gene selection and classification of microarray data depending on the Potential Support Vector Machine (P-SVM) for feature selection and a nu-SVM for classification. P-SVM approach extends the decision function through sparse group of support features. A fully automated technique for gene selection based on hyper-parameter optimization and microarray classification has been presented [18].

Banka et al., has presented an evolutionary rough feature selection algorithm which is presented for classifying gene expression patterns. As the data classically comprises of a huge number of redundant features, an initial redundancy reduction of the features is also done to facilitate faster convergence. Rough set theory is carried out to produce the distinction table that facilitates PSO to identify reducts, which denote the minimal sets of non-redundant attributes capable of discerning between all objects. The significance of the approach is illustrated on datasets such as Colon, Lymphoma and Leukemia using MOGA [19].

Dash and Patra have presented and Supervised CFS-Quick Reduct algorithm by integrating Correlation based Feature Selection (CFS) and Rough Sets attribute minimization for gene selection from gene expression data. Correlation based Feature Selection is employed as a filter to remove the redundant features, then the minimal reduct of the filtered feature set is minimized by rough sets. Three different classification approaches are used to assess the performance of this approach. This approach improves the significance and minimizes the complexity of the classical algorithm. The results are carried out on two public multi-class gene expression datasets and it is shown that this approach is very efficient in selecting high discriminative genes for classification task [20].

Saras Saraswathi et al., has proposed a novel combination of Integer-Coded Genetic Algorithm (ICGA) and Particle Swarm Optimization (PSO), coupled with the neural-network-based Extreme Learning Machine (ELM) is used for gene selection and cancer classification [21]. ICGA is used with PSOELM to choose an optimal set of genes, which is then utilized to generate a classifier that handles sparse data and sample imbalance.

In a simple modified ant colony optimization (ACO), algorithm is proposed to select tumor-related marker genes, and support vector machine (SVM) is used as classifier to evaluate the performance of the extracted gene subset. Experimental results on several benchmark tumor microarray datasets have shown that the proposed approach produces better recognition with fewer marker genes than many other methods [22]. It has been demonstrated that the modified ACO is a useful tool for selecting marker genes and mining high dimension data. Lee et al., has introduced an effective gene selection method named Adaptive Genetic Algorithm (AGA).

The authors evaluate the classification accuracy for this proposal method in Colon cancer microarray dataset by using K Nearest Neighbor KNN algorithm [23]. The results of this study indicate that AGA/KN can be provided as an effective tool with excellent performance for dimension reduction and gene selection on microarray dataset. Recently, Lee & Leu (2011) have applied Genetic algorithm with dynamic parameter setting (GADP) to select more predictive genes. After selecting the set of significant genes, they use an SVM to test the prediction accuracy using the set of the selected genes. When compared with other methods in literature, the proposed GADP method selects fewer genes with higher prediction accuracy for GCM dataset. Additionally, the GADP method is faster than a traditional GA method in execution time.

1.2 INFERENCE FROM THE EXISTING WORK

Many researchers are presented an investigation to deal with these issues of gene expression data through a swarm intelligence approach that will make possible in diagnosing the diseases and it may helpful in medical field for the various purposes. Optimization has been a preferred analysis topic for many years. However, these powerful optimization ways have their inherent shortcomings and limitations. Know that, fusion of the computational intelligence approaches will typically offer superior performances over using them one by one. Therefore, a hybrid clustering based optimization algorithm will be proposed to cope of with complex problems in gene expression data.

1.3 COMPARITIVE ANALYSIS OF THE EXISTING WORK

The comparative analysis of the gene selection and classification on gene expression data by various researchers are discussed below

Table-1: Comparative Analysis

Author’s Name	Existing Methods	Results
Juanying Xie &	Global k-means	89 % of

Shuai Jiang (2010)	clustering approach	accurate result
Sauravjyoti Sarmah & Dhruba Bhattacharyya,(2010)	Clustering technique Incremental (GenClus)	90 % results
Banka & Dara (2012)	Evolutionary rough feature selection algorithm	Results in Faster convergence
Saras Saraswathi et al., (2011)	ICGA-PSO-ELM	91% accuracy
Lee et al., (2011)	AGA/KNN	93% accuracy

From the Table 1 it shows that the clustering based optimization approaches perform better when compared with other existing algorithms.

3. PERFORMANCE ANALYSIS OF THE EXISTING WORK

The performance of the optimization algorithm is evaluated based on the parameters like

- Convergence behavior
- Processing Time

Convergence Behavior

Table 2. shows the performance and comparison of the optimization techniques such as Clustering Techniques, Optimization Algorithms and Swarm Intelligence Approaches.

The swarm intelligence based on algorithms like PSO, ABC etc performs the Clustering Techniques and Optimization Algorithms in attaining the reliability in terms of convergence behavior.

Figure 1 shows the comparison of the convergence behavior of the Clustering Techniques, Optimization Algorithms and the hybrid clustering with Optimization Approaches. It is observed from the figure that the hybrid clustering with Optimization Approaches converges in lesser iterations when compared with the other clustering and optimization techniques. For instance, the Clustering Techniques approach takes 90 iterations for convergence and optimization Algorithms approach takes 70 iterations whereas the hybrid clustering with Optimization Approaches takes 40 iterations for convergence. Thus the hybrid clustering with Optimization Approaches is very significant when compared with the other optimization approaches are taken for consideration.

Table -2: Performance comparison of the various techniques using convergence behavior

Iterations	Clustering Techniques	Optimization Algorithms	Clustering with optimization approaches
0	4	3.5	3
10	2.95	2.75	2.2
20	2.8	2.6	2.1
30	2.75	2.55	1.8
40	2.65	2.4	1.65
50	2.55	2.35	1.4
60	2.5	2.25	1.4
70	2.1	1.65	1.4
80	1.9	1.65	1.4
90	1.8	1.65	1.4
100	1.8	1.65	1.4
110	1.8	1.65	1.4
120	1.8	1.65	1.4

In presents an attribute clustering

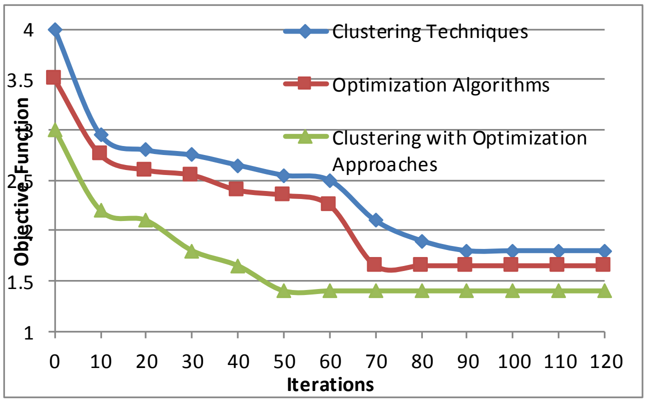

Fig.1. Comparison of Convergence Behavior.

Processing Time

Table 3.shows the performance comparison of the techniques such as Clustering Techniques, Optimization Algorithms and the hybrid clustering with Optimization approach.

The hybrid clustering with Optimization Approaches outperforms the Clustering Techniques and Optimization Algorithms in attaining the reliability in terms of the processing time.

Table-1: Performance Comparison of the Optimization Techniques using Processing Time

Optimization Algorithms	Processing Time (sec)
Clustering Techniques	21
Optimization Algorithms	18
Hybrid clustering with Optimization	10

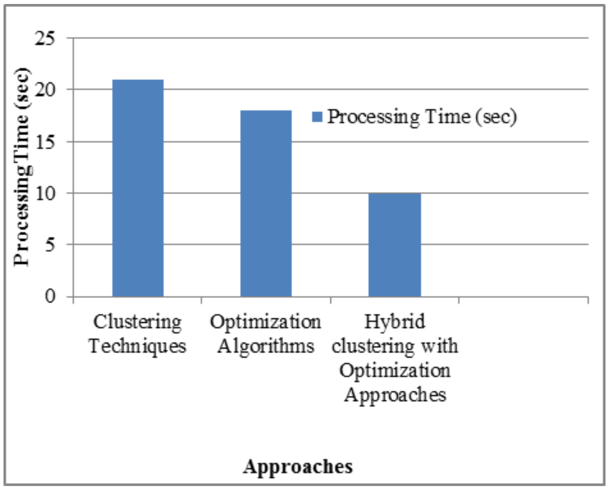

Fig.2. Comparison of Processing time.

It is observed from the Fig. 2. that the hybrid clustering with Optimization Approaches takes processing time of 10 seconds, whereas the other optimization techniques such as Clustering Techniques and Optimization Algorithms takes longer processing time such as 21 and 18 seconds respectively.

4. CONCLUSION

Clustering is a method of grouping similar types of data. This is very useful method applied in various applications. As an essential tool for data exploration, cluster analysis investigates unlabeled data, by either framing a hierarchical structure, or forming a set of groups based on a prespecified number. This paper focuses on the clustering algorithms and reviews a wide variety of approaches available in the literature. The performance of every clustering algorithm may vary significantly with diverse data sets, and there is no absolute finest algorithm among the clustering algorithms.

The significant disadvantage of clustering algorithms are the fact that time variation is not considered in its calculations, variations of densities in the data space resulting in overlapping clusters, cluster validation, presence of irrelevant attributes, high level of background noise, no prior knowledge and the dimensionality curse.

To overcome the problems with cluster analysis, in addition to these long-established traditional algorithms new developed approaches can be used to depict the underlying structure of the genetic network. So far Effectiveness of a hybrid cluster based optimization technique is highly influenced by the proximity measure used by the technique. Finally, review of various techniques will be helpful for better study and inventing new ideas for even better optimization techniques.

REFERENCES

[1] Satchidananda Dehuri and Sung-Bae Cho, "Multiobjective Classification Rule mining Using Gene Expression Programming," in proceedings of Third International Conference on convergence and Hybrid Information Technology, Vol. 2, pp. 754-760, 11-13 November, Busan, 2008.

[2] Andrew K. Rider, Geoffrey Siwo, Scott J. Emrich, Michael T. Ferdig, Nitesh V, "A Supervised Learning Approach to the Ensemble Clustering of Genes", International Journal of Data Mining and Bioinformatics, Vol. 3, No. 3, pp.229-259, 2009.

[3] Sushmita Mitra, Sankar K. Pal and Pabitra Mitra, "Data Mining in Soft Computing Framework: A Survey," IEEE Transactions On Neural Networks, Vol. 13, No. 1, 2002.

[4] Slavkov, I., Dzeroski, S., Struyf, J., Loskovska, S. "Constrained Clustering Of Gene Expression Profiles" in Proceedings of the Conference on Data Mining and Data Warehouses at the 7th International Multi-conference on Information Society, pp. 212-215, October 10-17, Slovenia, 2005.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] K.R De and A. Bhattacharya , "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles," bioinformatics, Vol. 24, pp.1359- 1366, 2008.

[7] Sunnyvale, Schena M. " Microarray biochip technology ". CA: Eaton Publishing; 2000.

[8] Wolfgang Huber, Anja von Heydebreck and Martin Vingron, "Analysis of microarray gene expression data", Max-Planck-Institute for Molecular Genetics 14195 Berlin April 2, 2003.

[9] Jian J. Dai, Linh Lieu, and David Rocke, "Dimension reduction for classification with gene expression microarray data," Statistical Applications in Genetics and Molecular Biology, Vol. 5, No. 1, pp. 1–21, 2006.

[10] D.Napoleon, S.Pavalakodi, "A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set", International Journal of Computer Applications Volume 13– No.7,pp. 41-46 January 2011

[11] G. Piatetsky-Shapiro, T. Khabaza, and S. Ramaswamy, "Capturing Best Practice for Microarray Gene Expression Data Analysis," in Proc. of the 9th ACM SIGKDD Int'l Conf. on

Knowledge Discovery and Data Mining, Washington, DC, 2003, pp. 407–415.

[12] A. K. C. Wong and Y. Wang, “Pattern Discovery: A Data Driven Approach to Decision Support,” IEEE Trans. on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 33, no. 1, pp. 114–124, 2003.

[13] Juanying Xie, Shuai Jiang, “A simple and fast algorithm for global k-means clustering”, 2 nd International Workshop on Education Technology and Computer Science (IEEE), pp.36-40, 2010.

[14] Wai-Ho Au* , Keith C. C. Chan, Andrew K. C. Wong, and Yang Wang, “Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data”, IEEE, 2005

[15] W. Haiying, Z. Huiru, and A. Francisco, “Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data,” IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 2,pp.163-175,Apr.-June 2007

[16] Sauravjyoti Sarmah and Dhruba K. Bhattacharyya, “An Effective Technique for Clustering Incremental Gene Expression data”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May 2010

[17] Zhang Qizhong, “Gene Selection and Classification Using Non-linear Kernel Support Vector Machines Based on Gene Expression Data”, IEEE, 2007.

[18] Mohr, J. ; Sambu Seo ; Obermayer, K., “Automated Microarray Classification Based on P-SVM Gene Selection”, IEEE, 2008.

[19] Banka, H. ; Dara, S., “Feature Selection and Classification for Gene Expression Data Using Evolutionary Computation”, IEEE, 2012.

[20] Dash, S. ; Patra, B., “Rough set aided gene selection for cancer classification”, IEEE, 2012.

[21] Saras Saraswathi, Suresh Sundaram, Narasimhan Sundararajan, Michael Zimmermann, and Marit Nilsen-Hamilton, “ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets in Which Genes Encoding Secreted Proteins Are Highly Represented”, IEEE/ACM Transactions on computational biology and bioinformatics, vol. 8, no. 2, march/april 2011

[22] Yu H, Gu G, Liu H, Shen J, Zhao J., “A modified ant colony optimization algorithm for tumor marker gene selection”, Genomics Proteomics Bioinformatics, 2009 Dec;7(4),2008.

[23] C.-P. Lee, W.-S. Lin, Y.-M. Chen, and B.-J. Kuo, “Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method,” Expert Systems with Applications, vol. 38, no. 5, pp. 4661 – 4667, 2011.