Enhanced Correlation Search Technique For Clustering Cancer Gene Expression Data

B. SATHIYABHAMA[1], N.P. GOPALAN[2]

[1] Department of CSE, Sona College of Technology, Salem, INDIA.

[2] Department of CSE, National Institute of Technology, Tiruchirappalli, INDIA.

Abstract: - The advent of DNA Microarray technologies has revolutionized the experimental studies of gene expressions. In the post-genomics era, clustering analysis has become a valuable tool for in-silico analysis of gene expression profiles. Although a number of clustering methods have been proposed, they are confronted with difficulties in meeting the requirements of high quality, large memory, performance and automation. In this paper, a novel-clustering algorithm namely Heuristic based Enhanced Correlation Search Technique (HECST) has been proposed. The distinct characteristic of HECST is that it integrates the validation techniques into the clustering process so that it produces high quality clusters dynamically. This algorithm is implemented using memory efficient data structure namely sparse matrices to store the gene expression profile. Sparse matrices tremendously reduce the size of the memory, hence provides computational efficiency. The performance of the algorithm is evaluated against number of reasonable benchmarks (e.g Direct application of raw data) for cancer gene expression data sets. The empirical results proved that this new algorithm automatically produces the optimal clusters in a much faster way than the traditional clustering methods like K-means, CAST and E-CAST. Analysis of data produced by HECST tenders potential insight into gene function, molecular biological processes and regulatory mechanisms.

Keywords: - Clustering, Gene Expression Data, validation Techniques, Sparse matrices, Correlation search technique and algorithm.

1 Introduction

Microarray experiments for simultaneously measuring expression levels of thousands of genes are becoming widely used in genomic research. They have enormous promise in such areas as revealing function of genes in various cell populations, tumor classification, drug target identification, understanding cellular pathways, and prediction of outcome to therapy [11], [14]. A major application of microarray technology is gene expression profiling to predict outcome in multiple tumor types [17]. Various data-mining methods can be applied to cancer datasets in order to identify class distinction genes and to classify tumors. A partial list of methods includes data preprocessing, visualization methods and clustering. The focus here is on clustering methods. Clustering techniques are useful in identifying (yet unknown) subclasses of tumors, or identifying clusters of genes that are coregulated or share the same function [1].

The elements or objects within clusters have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [1], [14]. Clustering can be used to categorize genes with similar functionalities and gain insight into structures inherent in population. These methods have been successful in separating certain types of tumors associated with different types of leukemia and lymphoma [17]. The groupings of biologically relevant clusters containing genes are having similar expression patterns. Thus clustering reveals co expression of genes, which were uncharacterized previously.

Clustering technique has become an efficient and mandatory tool for in-silico analysis of gene expression data [4], [5], [6], [9], [12]. A variant of hierarchical clustering algorithm is used by Eisen et al. [6] to identify groups of co expressed yeast genes. Two-way clustering technique [4] is used to detect clusters of correlated genes and tissues. To identify clusters in the yeast cell cycle data set, human hematopoietic differentiation data set Self-organizing maps [12] were used. Biologically meaningful clusters of yeast chodata have been determined by using genetic enhanced K-Means method [2]. These techniques have the drawbacks of computational adequacies, lack clustering quality and destabilization of clusters. Vincent S. Tseng et al. [3] proposed a new clustering algorithm that incorporates validation technique and produces high quality clusters. Number of clustering methods have been proposed [1], [7], [8], [10], they are confronted with the following difficulties:

- Most clustering algorithms request the users to specify some input parameters like number of clusters, structures and conditions.
- Clustering algorithms are incapable of producing optimal results for large data sets.
- Traditional clustering methods may not perform well, when non-optimal clustering result is enforced.

Input parameters are playing vital role in determining the efficiency of the clustering results. In biological applications it is very difficult to obtain certain parameters manually. Thus an automatic clustering technique is required to identify the suitable input parameters. Further validation indexes are used to improve and evaluate the quality of the clustering, the suitability of parameters and the reliability of clustering algorithms. The biclustering gene expression data using Numerous validation indexes are used in practice like Jaccard coefficient, Simple matching coefficient and Hubert's Γ (gamma) statistic [13], [15]. Correlation based clustering algorithm [3] uses validation index Hubert's Γ statistic [15]. This method consumes more memory and execution time hence it becomes computationally inefficient. Vincent Tseng et al. [3] used the same constraint for both addition and removal of elements to the clusters. The validation index used by them is more complex and time consuming. These constraints make the algorithm very unstable while forming the final clusters and the outliers also not properly been filtered out. In the proposed work, correlation based clustering method is integrated with validation technique. The validation index Hubert's Γ statistic is simplified and enhanced to cluster gene expression data set. The significant characteristics of the proposed approach are as follows: First the algorithm uses memory efficient data structure called sparse matrix, which is used to store the gene expression similarity matrix, which reduces the amount of memory required. Unlike the traditional clustering algorithms the proposed algorithm uses the constraint based addition procedure to add the elements to the clusters. This algorithm

never removes any element from the clusters once added and also outliers are filtered out during the initial stage itself. Hence the stability and quality of the clustering process is improved.

2 Enhanced Hubert'S Γ Statistic

A similarity matrix S is generated based on the given Microarray data set. The matrix S stores the similarity between each pair of genes in the data set, with the degrees in range of [0,1]. To obtain the similarity, Pearson's correlation coefficient [15] similarity measurement has been used. The sparse matrix T can be generated from the matrix S. Sparse matrix is the three columns matrix that stores only non-zero entries of the original matrix. In the first row, number of rows, number of columns and total number of non-zero elements are stored. From the second column onwards the row value, column value (i.e., the position of the non-zero element in the original matrix) and the value of the non-zero element are stored successively.

0	0	0	1	0
1	0	0	0	1
0	0	0	0	0
0	1	0	0	0
0	0	1	0	0

Table 1 Original Matrix

5	5	6
1	4	1
2	1	1
2	5	1
4	2	1
5	3	1

Table 2 Sparse Matrix

The representation of the sparse matrix for the given input matrix and the amount of memory reduction is clearly understood from the Tables 1 and 2. Then the clustering process automatically clusters the genes according to the similarity matrix with out any user-input parameters. To cluster the genes quickly and automatically validation technique is integrated with the clustering process. Let X = [X(i,j)], Y=[Y(i,j)] be the proximity or closeness matrices for the same n ge nes. X(i,j) indicates the observed correlation coefficient of genes i and j, Y(i,j) is defined as

$$Y(i,j)=\begin{cases}1 & if\ genes\ are\ clustered\ in\ the\ same\ cluste\\ 0 & otherwise\end{cases}$$

(1)

The Hubert's Γ statistic represents the point serial correlation between the similarity matrices X and Y. Given a gene expression data clustering results the more genes that fall in the same clusters of higher similarity and different clusters of lower similarity are considered best quality clusters. Therefore, the point serial correlation between the matrices X and Y can be used to measure the quality and reliability of clustering results. The enhanced Hubert's Γ statistic is a very simple and rapid technique that also improves the performance of the algorithm.

$$\Gamma=\frac{1}{\mathrm{M}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{X(i,j)-\overline{X}}{\sigma X}\right)\left(\frac{Y(i,j)-\overline{Y}}{\sigma Y}\right) \tag{2}$$

Where $M=\frac{n(n-1)}{2}$ is the number of entries the double sum, and σX and σY are the standard deviations, while X and Y denote means of the entries of matrices X and Y. The enhanced statistic is derived from expanding and substituting the necessary fast heuristic from the basic Hubert's Γ statistic in (2).

$$\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left((x(i,j)(y(i,j))-\frac{2}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}X(i,j)Y(i,j)+\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x(i,j)}{M}\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Y(i,j)}{M}\right)}{\sqrt{\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(X(i,j)-\overline{X^2}\right)\right)\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Y(i,j)-\overline{Y^2}\right)}} \tag{3}$$

The value of σX in (2) is invariable, and an expansion of σY is as follows:

$$\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(Y(i,j)-\overline{Y}\right)^2} \tag{4}$$

The similarity matrix values are computed in terms of degrees in range (0,1) so the square of Y(i,j) is also equal to Y(i,j). Hence the denominator of (3) is much shorter term than the numerator. So the equation in (3) is referred as fast and enhanced Hubert's Γ statistic.

3 Pseudo Code

The input for this algorithm is a sparse matrix that is constructed from the symmetric similarity matrix of the given gene expression data set. This constructs clusters one at a time. The current cluster is denoted by C_{open}. Each cluster is started by a seed value and constructed incrementally by adding items to C_{open}. The addition of data items is computed using enhanced fast Hubert's Γ statistic (3) and is defined as $\Gamma_{add}(k)$. The current maximum is represented as Γ_{max}. An element k is added if it has high positive correlation i.e high similarity. Also it clusters low similarity gene data items in different clusters according to the value. The value of Γ is between (-1,1) and a higher value of Γ represents the best clustering quality. A data item is added to the cluster if it satisfies the maximum neighbors criteria and a threshold value. In general, the threshold value depends on the number of patterns and the number of features in the data set. The C_{open} procedure is stabilized by consecutive addition operations. To inaugurate a new cluster, a data item with maximum number of neighbors or closest data items is used. Also a threshold value is used while adding an element, it automatically filters out the outlier data items and appropriately insert in to the respective clusters. These are the principal heuristics that is been attached to the algorithm and is also responsible for assigning clusters to all the valid items only once. The pseudo code is as follows:

PROCEDURE HECST;

// Algorithm for correlation based enhanced fast Hubert's Γ statistic clustering;

INPUT - An n X n sparse(Symmetric similarity)matrix T;

OUTPUT – Clusters having high intra cluster similarity and low inter cluster similarity;

BEGIN

M = n(n –1) /2;

$S_{tx} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} T(i,j)$;

S_{ty} = 0;

S_{txy} = 0;

C = Φ ; // the set of clusters

U = { 1,2,3, …., n};

Γ_{max} = 0;

While (U != 0) ***do***

begin

C_{open} = 0;

for i = 1 to d do // Expected number of clusters d

a[i] = 0;

// Assigning seed values, select the data item u from the U with maximum neighbors

for i = 1 to n do

U = U – {u};

for i = 1 to n do

a[i] = T(u,i);

$C_{open} = \{u\};$

// Addition of elements to the clusters based on constraints and threshold value

while maxvalid() and ctthresh

begin

for i = 1 to n do

begin

if (u has maximum neighbors (mn) in a[i])

select u as to start;

end;

$U = U - \{u\};$

$S_{ty} = S_{ty} + |C_{open}|;$

$S_{txy} = S_{txy} + a[u];$

for i = 1 to n do

begin

if (i Є U and i Є C_{open})

a[i] = a[i] + T(u,i);

end;

$C_{open} = C_{open} \cup \{u\};$

$\Gamma_{max} =$ maxvalid();

end;

end;

$C = C \cup C_{open};$

END.

// maxvalid() is a subroutine ie calculated as follows:

PROCEDURE maxvalid()

BEGIN

$$res = \frac{\left(M * \left(S_{txy} + \max\{a(u) \mid u \in U\}\right) - S_{tx} * \left(S_{ty} + |C_{open}|\right)\right)}{\sqrt{M}\left(S_{ty} + |C_{open}|\right) - \left(S_{ty} + |C_{open}|\right)}$$

return (res);

END.

4 Evaluation

To evaluate the performance of the proposed approach cancer gene expression data set is used. Datasets [18] from breast cell lines transfected with the CSF1R oncogene creating a phenotype that invades and metastasizes. The benign cell line was then transfected with the two mutated oncogenes, creating one phenotype that invades and another one that metastasizes. Gene expression levels were measured eight times for each phenotype. Transfection with a single oncogene is expected to generate similar expression profiles, presumably because only a few genes are biologically influenced. Therefore, it was desirable to see whether profiles of the different phenotypes can be partitioned. Due to noise in the data and similarity between the different samples, common clustering techniques such as hierarchical, K-Means, and E-CAST did not succeed in cleanly partitioning the data. Expression levels of the four cell lines were measured in two separate sets of four measurements. To measure the ratio of three of the cell lines: benign, invasive, and metastasizes with respect to the cell line that invades in the first batch, and the corresponding ratios were similarly derived for the second batch.

These data sets' cluster structures are determined in advance. From the given data set, the users can set up some parameters for generating various kinds of gene expression data sets with variation in terms of the number of clusters and number of

genes in each cluster. The program also generates the seed genes. The seed genes must have the same number of constraints for all the clusters. If the seed genes and the threshold values are appropriately incorporated and tested in the algorithm, then all the genes in the same clusters will have very high similarity and they will have dissimilarity with genes in the other clusters. It is also filtering the outliers or noise data. To test the algorithm's performance two data sets of gene expression profiles are generated.

Data sets	Cluster details	Patterns	Proposed approach (HECST) & C. time	E-CAST & C. time	K-means & C. time
1	Number of clusters = 4 1000,950,900,850	3000	0.9 O(n log n)	0.7 $O(n^2)$	0.45 $O(n^2)$
2	Number of clusters = 6 550,500,450,300,400,350	3800	0.85 O(n log n)	0.6 $O(n^2)$	0.33 $O(n^2)$

Table 3 Experimental results for the synthetic data sets

The proposed algorithm is compared with the other clustering algorithms for gene expression data sets. The table 3 provides the complete details about the cluster structure, clustering patterns for HECST, E-CAST, K-Means and their computational time (C. time in Table 1). The newly designed algorithm HECST outperforms quantitatively and qualitatively in computational time and the memory utilization. In addition, the results show that the quality of clustering will be better in the proposed algorithm. This can provide more accurate clustering results and insight into molecular process, morphological characteristics and gene control functions. Figures 3 and 4 show that large contiguous group of genes share the similar expression patterns over set of conditions. This type of clustering structure elaborates the biological significance of the underlying genes. The result of this clustering analysis may be a group of co-regulated genes (i.e. genes that exhibit similar experimental behavior) that are placed in the same cluster. They express the relationships between the clusters and the functional categories in biological activities.

The data sets presented here demonstrates a feature of gene expression that makes this method particularly useful, namely tendency of expression data to organize genes in to functional categories. It is known that genes expressed together share common functions. Gene expression patterns suffice to separate genes into functional categories across a relatively small and redundant collection of conditions. It seems likely that the addition of more and diverse conditions can only enhance these observations.

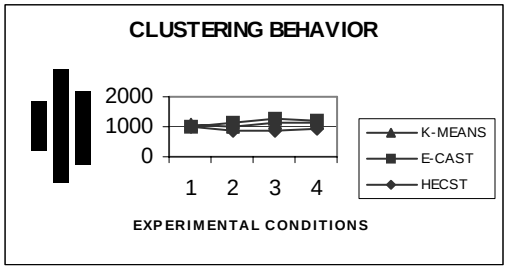

Fig. 1 Cluster Profile for Data set 1

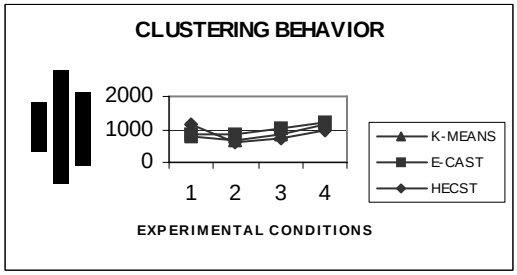

Fig. 2 Cluster Profile for Data set 2

The proposed algorithm is superior to the existing approaches in quality and efficiency, stability and memory utilization. HECST algorithm is compared with E-CAST [16] and traditional k-means algorithm [1]. It is understood from the Fig. 1 and 2 that HECST algorithm emphasizes its supremacy of capturing sharp coherent tendency among gene expression data. In addition, the result of functional enrichment of HECST clusters highlight the fact that these clusters carry significant biological meaning.

5 Conclusion

Clustering analysis is a valuable and useful technique for in-silico analysis of Microarray data. Most of the clustering algorithms used in practice are having certain inherent difficulties in the aspects of automation, economic memory usage, quality, efficiency, and stability. The proposed work integrates enhanced, simplified and heuristic based validation index to the clustering process. This algorithm clusters the gene expression data sets dynamically and produces optimal results. The dynamic clustering process signifies great promise for using this technique to glean information from gene expression profile. To evaluate the performance of this novel algorithm cancer gene expression data sets have been used and it is compared with the E-CAST and K-Means clustering algorithms. HECST is outperforming in terms efficiency, clustering quality, stability and performance. Future work includes the application of HECST on more real data sets and the theoretical analysis of the determination of the threshold parameter.

Reference:

[1] M.S. Chen, J. Han, and P.S. Yu, " *Data Mining: An overview from a Database Perspective*", IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, pp. 866 –883, Dec. 1996.

[2] N.P. Gopalan, B.Sathiyabhama, "*Scalable biclustering gene expression data using genetic enhanced K-Means algorithm*", Proc. National conference on High Performance Computing – VISION'06, pp. 494-498.

[3] Vinvent S. Tseng and Ching-Pin Kao, " *Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method*", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 2, No. 4, pp. 355 – 365, Dec. 2005.

[4] U. Alon, N. Barkai, D.A. Nottleman, k. Gish, S. Ybarra, D. Mack, and A.J. Levine, "*Broad patterns of Gene Espression revealed by Clustering analysis of Tumor and Normal Colon Tissues Probed by Clustering Oligonucleotide arrays*", Proc. Nat'l Academy of Sciences, vol. 96, pp. 6745-6750, 1999.

[5] A.Ben-Dor and Z. Yakhini , " *Clustering Gene Expression Patterns*", J.Computational Biology, vol. 6 pp. 281-297, 1998.

[6] M.B. Eisen , P.T. Spellman, P.O.Brown, and D. Botstein, "*Clustering Analysis and Display of Genome Wide Expression Patterns*", Proc. Nat'l Academy of Sciences, vol. 95, pp. 14863-14868, 1998.

[7] S. Guha, R. Rastogi, and K. Shim, "*CURE: An Efficient Clustering Algorithm for Large Databases*", Proc. ACM Int'l Conf. Management of Data, pp. 73-84, 1998.

[8] S. Guha, R. Rastogi, and K. Shim, "*ROCK: A Robust Clustering Algorithm for Categorical Attributes*", Proc. 15th Int'l Conf. Data Eng., pp. 512-521, 1999.

[9] M.K. Kerr and G.A. Churchill, "*Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments*", Proc. Nat'l Academy of Science, vol. 98, no. 16, pp. 8961-8965, 2001.

[10] T. Kohonen, "*The Self-Organizing Map*", Proc.IEEE, vol. 78, no. 9, pp. 1464–1479, 1990.

[11] P.T. Spellman , G. Sherlock, M.Q. Zhang, V.R. Iyer, K. anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Fucher, " *Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridi zation*", Molecular Biology of the Cell, vol. 9, no. 12, pp. 3273-3297, 1998.

[12] P. Tamayo, D. Slonim, J.Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "*Interpreting Patterns of Gene Expression With Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation*", Proc. Nat'l Academy of Sciences, vol. 96, no. 6, pp. 2907- 2912, 1999.

[13] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, "*Validating Clustering for Gene Expression Data*", Bioinformatics, vol. 17, no. 4, pp. 309-318, 2001.

[14] T. Zhang, R. Ramakrishnan, and M. Livny, "*Birch: an Efficient Data Clustering Methosd for very Large Databases*", Proc. 1996 ACM SIGMOD Int'l Conf. Management of Data, pp. 103-114, 1996.

[15] A.K. Jain and R.C. Dubes, "*Algorithms for Clustering Data. Englewood Cliffs*", N.J.: Prentice Hall, 1988.

[16] Abdelghani Bellaachia et. al. "*E-CAST: A Data Mining Algorithm for Gene Expression Data*", Proc. BIOKDD02: Workshop on Data Mining in Bioinformatics (With SIGKDD02 conference) pp. 49-54.

[17] Golub. T. R. Slonim, D.K. Slonim, D.K. Tamayo, P. Huard, C. Gaasenbeek, M. Mesiroy, J.P. Coller, H. Loh, M. Downing, J.R., Caligiuri, M. et al. "*Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*". Science 286: 531-537, 1999.

[18] Kluger, H. Kacinski, B., Kluger, Y., Mironenko, Gilmore Hebert, M., Chang, J., Perkins, A.S., and Sapi, E. "*Microarray analysis of invasive and metastatic in a breast cancer model*". In Poster presented at the Gordon Conference on Cancer, Newport, RI, 2001.