An Efficient Approach for High Dimensional Data Clustering of Gene Expression using Dynamic Error Threshold Estimation Model

K. Arun Prabha[1], A. Amutha[2]

[1]Assistant Professor, PG Department of Computer Science
Vellalar College for Women (Autonomous)
Erode-12, Tamilnadu, India

[2]M. Phil Scholar, Department of Computer Science
Vellalar College for Women (Autonomous)
Erode-12, Tamilnadu, India

Abstract: ***Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition and bioinformatics. Gene expressions are one of the high dimensional data values and its motivating the development of clustering algorithm was used. The Existing system consists of popular algorithms like k-means and CAST. Implementing these algorithms for a large genome-scale gene expression data set is practically critical. A novel method for clustering large gene data set is introduced. In Existing work the TCLUST algorithm used, which introduce, Correlation Coefficient Graph (CCG) is constructed to maintain gene expression data values and Tanimoto Coefficient Graph (TCG) is used to measure the similarity value for the gene expression data. In proposed the enhanced TCLUST algorithm is used, it is called as E-TCLUST. Enhanced Tanimoto clustering method is implemented which feats the co-connectedness for efficiently clustering large, sparse expression data. Dynamic error threshold estimation model implements threshold values which filters data below the given threshold value. In the proposed work tree structure is constructed represent the input samples. Using graphs the variations are identified .Graph Re-arrangement mechanism is performed which effectively reduces the number of iterations. The process time is also reduced. Extensive evaluation of this method reveals an optimized performance which is depicted as a graph. This algorithm is applied to a genome-scale gene expression data set and used gene set enrichment analysis to obtain highly significant biological clusters. It have been implemented both TCLUST and E-TCLUST algorithms and tested their performance using three different data sets. The datasets are real gene expression data from yeast samples generated using micro-arrays technology.***

Keywords: Clustering, Gene Expression, Micro-array, Bio-informatics, Data mining

1. Introduction

With the introduction Clustering analysis has received significant attention in the area of gene expression. It allows the identification of the structure of a data set, i.e. the identification of groups of similar objects in multidimensional space. In bio-informatics, these techniques are used to analyze data expression generated using micro-array technologies [6]. Several clustering techniques have been applied to the analysis of gene expression data. However, current clustering algorithms commonly used in the domain of gene expression do not scale well for genome-scale expression data [5].so it is convenient to clustering on a graph in which the vertices correspond to genes and edge weights reflect the similarity/correlation between the expressions [7].

In existing the TCLUST algorithm allows Clustering of large gene expression data sets is challenging. It is used to generate a correlation coefficient graph (CCG) from gene expression data, which is a mathematical measure of the similar relation between two data sets. The tanimoto coefficient graph (TCG) is used to measure the similarity between real gene data set. It uses a fixed (Static) initial threshold value to start the clustering. This parameter directly affects the size and number of clusters produced [1].

2. Existing System

Several clustering techniques have been studied for analyzing expression data Clustering of large gene expression data sets is challenging. In existing work the TCLUST algorithm will used for Clustering of large gene expression data set.

A. Weighted Cluster Editing Problem

There are two NP-complete graph problems

Cluster Editing and Cluster Deletion [3].These are based on the notion of a similarity graph whose vertices correspond to data elements and in which there is an edge between two vertices is the similarity of their corresponding elements exceeds a predefined threshold. The goal is to obtain a cluster graph by as few edge modifications. Given a weighted undirected correlation graph G = (V, E), where E = {(u, v}|s{u, v)} > 0}.

B. Coconnectedness-Based Heuristic

Co-connectedness is defined as the fraction of neighbors shared by the pair, and has been used in many different graph-theoretical problems. The underlying clique structure of the input graph has two independent cliques formed by the vertices 1, 2, 3, 4, 5, 6 and 7, 8, 9, 10, 11. The observed graph has some noise as some of the edges

within the cliques are missing and there are some edges between the cliques.

So calculating vectors will use Tanimoto coefficient (TC) to describe Co-connectedness of two vertices u and v as follows:

$$TC(u, v) = \frac{\bar{u}.\bar{v}}{\bar{u}.\bar{u} + \bar{v}.\bar{v} - \bar{u}.\bar{v}}$$

$\bar{u}.\bar{u}$ = Number of elements in set $\bar{u}.\bar{u}$

$\bar{v}.\bar{v}$ = Number of elements in set $\bar{v}.\bar{v}$.

$\bar{u}.\bar{v}$ = Number of elements in intersecting set $\bar{u}.\bar{v}$.

It measures not only if two vertices are connected to the similar set of vertices but also if they are connected with similar edge weights.

C. T-Clustering Process

Figure 1 provides an overview of TCLUST. A Correlation coefficient graph (CCG) is constructed from gene expression data. This is followed by iterative computation of TCG until connected components of the TCG are clique like. Once a connected component is dense enough, it will only get denser in subsequent iterations and no two disjoint connected components will be merged. Thus, it is possible to output a connected component as a cluster at any iteration as its edge density, i.e., number of edges/number of pairs. Clustering problem is to be considered since one has to divide a complete set of elements into homogenous and a well separated subset.

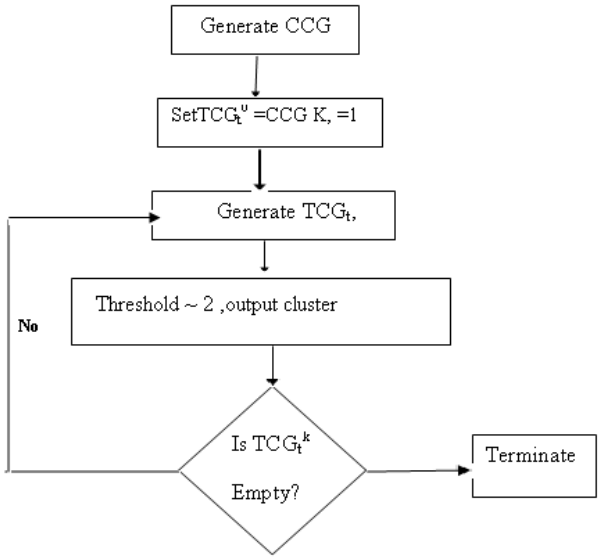

Figure 1: Overview of TCLUST algorithm

1. Generating Correlation Coefficient Graph

In CCG the Pearson's correlation coefficient (PCG) as a measure of co expression of genes/probe sets. It is a mathematical measure of the relation between two data sets.

Let x_{ik} denote the expression level of pi in the Kth sample. For a pair of probes p_i, p_j

$$s(p_i, p_j) = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2 (x_{jk} - \bar{x}_j)^2}} \text{ W}$$

here,

x_{ik} = Number of values in X_{ik}.

x_{jk} = Number of values in X_{jk}.X_{ik} and x_{jk} values are x_{ik}BAR and x_{jk}BAR respectively.

2. Generating Tanimoto Coefficient Graph (TCG)

TCG (Tanimoto coefficient graph) is used to measure the similarity value for the gene expression data during graph construction and the tanimoto clustering technique is used to cluster high dimensional data values [2].The TC is a measure of similarity between two real-valued vectors, define the TC between a pair of vertices (p_i, p_j) as follows:

$$TC\ (p_i, p_j) = TC\ (\overline{W_i}, \overline{W_j})$$

Where,

TC = Tanimoto Coefficient

P_i, Pj = Number of Vertices

W_i, Wj = Number of Weights

Threshold values are used to select cluster intervals. It uses a fixed (Static) initial threshold value to start the clustering, and controlled by setting a threshold value, depending upon the value will directly affects the size and number of clustering iterations produced, also less cluster accuracy and take more process time.

3. Proposed System

The gene expression data values are multi dimensional data values. The proposed system have implemented with enhanced TCLUST algorithm, it is called as E-TCLUST. In The correlation coefficient graph and tanimoto correlation coefficient graph are used for the similarity estimation process.

A. Enhanced T-CLUST

The Enhanced TCLUST (E-TCLUST) is followed by TCLUST method, In E-TCLUST the gene expression values are partitioned in graph tree as vertices and edges. In the proposed work, to increase the speed of clustering ETCLUST is used. Dynamic Error Threshold Estimation Model is applied to group the gene expression data set more fast and accurate. The data interval analysis is performed with error threshold values. The error threshold estimation is optimized. The graph update iterations are given a large genome scale gene expression dataset, first the missing and irrelevant data are removed. Initial iterations are performed to group the similar data values. Finally using the ETCLUST method the threshold value is defined. This simplifies the entire process by neglecting the values below the threshold and graph rearrangement is performed to speed up the process. Hence the end result would be very limited number of iterations for clustering very large dataset. And thereby the high process time is reduced with more cluster accuracy.

4. Results and Discussion

The TCLUST algorithm allows working with large scale expression of gene data, which produces the result by consuming more process time and iteration process leads to fewer accuracy. Compare to previous work the proposed enhanced TCLUST algorithm (E-TCLUST), along with dynamic threshold mechanism, provides more efficiency to work on large gene expression, by reduced process time and iteration and thus results with increased accuracy.

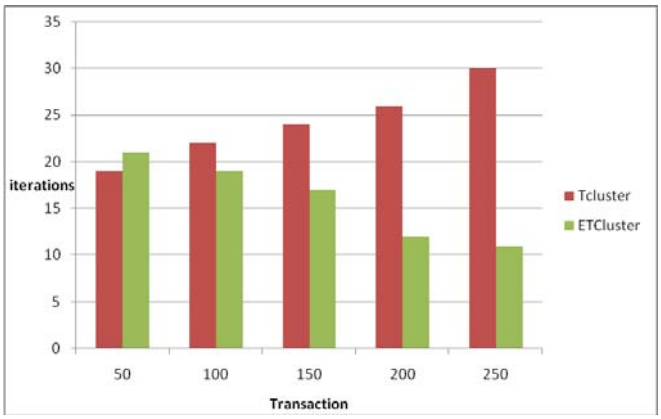

Figure 2: Iteration Analysis between TCluster and ETCluster

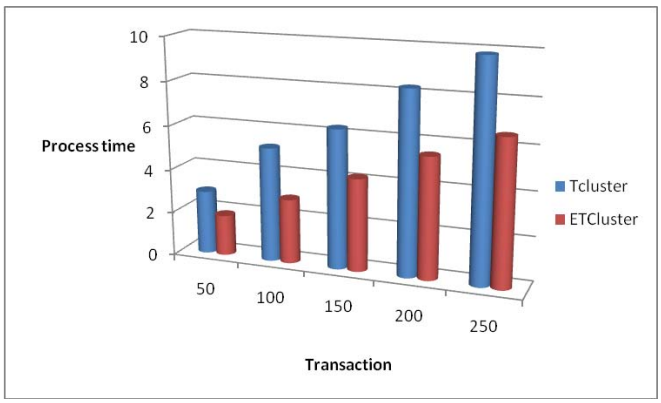

Figure 3: Process time Analysis between TCluster and ETCluster

5. Conclusion and Future Work

With the conclusion introduce here a new method, E-TCLUST for clustering large, genome-scale data sets. The algorithm is based on measures of co-connectedness to identify graphs present in the data. It have applied this method to a large reference gene expression data set, and showed that the resulting clusters show strong enrichment. E-TCLUST should be applicable in a variety of biological settings, and offers a new approach, complementing existing methods.

References

[1] Abdelghani Bellaachia and David Portnoy, Yidong Chen and Abdel. G. Elkahloun "E-CAST: A Data Mining Algorithm For Gene Expression Data" by National Institute of Health (NIH).

[2] Banu Dost, Chunlei Wu, Andrew Su, and Vineet Bafna "TCLUST: A Fast Method for Clustering Genome-Scale Expression Data" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 3, MAY/JUNE 2011.

[3] C. Huttenhower, A.I. Flamholz, J.N. Landis, S. Sahi, C.L. Myers, K.L. Olszewski, M.A. Hibbs, N.O. Siemers, O.G. Troyanskaya, and H.A. Coller, "Nearest Neighbor Networks: Clustering Expression Data Based on Gene Neighborhoods," BMC Bioinformatics, vol. 8, pp. 250-262, 2007.

[4] D. Gibson, R. Kumar, and A. Tomkins, "Discovering Large Dense

[5] Sub graphs in Massive Graphs," Proc. 31st Int'l Conf. Very Large Data Bases (VLDB '05), pp. 721-732, 2005.

[6] M. Gerstein and R. Jansen, "The Current Excitement in Bioinformatics—Analysis of Whole Genome Expression Data: How Does It Related to Protein Structure and Function,"citeseer.ist.psu.edu/gerstein00current.html, 2000.

[7] J. Quackenbush, "Computational Analysis of Microarray Data,"

[8] Nature Rev. Genetics, vol. 2, no. 6, pp. 418-427, http://dx.doi.org/

[9] 10.1038/35076576, June 2001.

[10] R. Shamir, R. Sharan, and D. Tsur, "Cluster Graph Modification

[11] Problems," Discrete Applied Math., vol. 144, nos. 1/2, pp. 173-182,

[12] http://dx.doi.org/10.1016/j.dam.2004.01.007, 2004.

Author Profile

Mrs. K. Arun Prabha received the B. Sc (Physics), MCA and M.Phil (Computer Science) degrees. She is an Assistant Professor in the PG Department of Computer Application Vellalar College for women (Autonomous) Erode. She has sixteen years of teaching experience and four years research field experiences. She published eight papers in international journals and national journals. Also she has presented six papers in the international, national and state level Conference and seminars. Her research interests are Data Mining and Soft Computing.

A. Amutha received the B.C.A degree from Bharathiyar University and M.C.A degree in 2008 and 2011 respectively from Anna University, Coimbatore. She is currently doing M. Phil (Computer Science) degree from Bharathiar University, Coimbatore.