Bioinformatics and Biology Insights

ORIGINAL RESEARCH

A Fast Quad-Tree Based Two Dimensional Hierarchical Clustering

Priscilla Rajadurai[1] and Swamynathan Sankaranarayanan[2]

[1]Department of Computer Science and Engineering, Anna University, Chennai, India. [2]Department of Information Science and Technology, Anna University, Chennai, India.
Corresponding author email: prisci.christa@gmail.com, swamyns@gmail.com

Abstract: Recently, microarray technologies have become a robust technique in the area of genomics. An important step in the analysis of gene expression data is the identification of groups of genes disclosing analogous expression patterns. Cluster analysis partitions a given dataset into groups based on specified features. Euclidean distance is a widely used similarity measure for gene expression data that considers the amount of changes in gene expression. However, the huge number of genes and the intricacy of biological networks have highly increased the challenges of comprehending and interpreting the resulting group of data, increasing processing time. The proposed technique focuses on a QT based fast 2-dimensional hierarchical clustering algorithm to perform clustering. The construction of the closest pair data structure is an each level is an important time factor, which determines the processing time of clustering. The proposed model reduces the processing time and improves analysis of gene expression data.

Keywords: clustering, euclidean distance, quad tree, hierarchical clustering

Bioinformatics and Biology Insights 2012:6 265–274

doi: 10.4137/BBI.S10383

This article is available from http://www.la-press.com.

Introduction

The DNA microarray has emerged as the latest breakthrough in molecular biology; it provides researchers with the opportunity to monitor genome-wide expression systematically.[1] The microarrays are used to study gene expression profiles in biological samples.[7] A microarray or microchip is a chip made with glass or other solid material, with an array of tiny DNA spots placed on it. Each of the spots contains fragments of DNA or RNA molecules whose sequence is predefined and corresponds to portions of a particular gene.[8] Microarrays do not provide full information about genes, but rather they denote genes indirectly through their expressions. Also, the expressions obtained can be inaccurate depending on the applied microarray technology.[30]

Microarrays have emerged as a standard for simultaneous evaluation of the expression level of thousands of genes.[2] Clustering techniques play a significant role by discovering sets of objects with identical functions from huge quantities of data.[3] A good clustering algorithm should be able to identify genes that have similar expression profiles, including time-shifted or inverted profiles, and provides phase information.[4] There are many popular techniques for clustering gene expression data by elucidating different functional roles in genes that play an important role in the biological process.[5] A major area of concern in the clustering analysis of gene expression data is the sensitivity and vulnerability of results to noise and overfitting, due to their excessive dependence on limited biological and medical information.[3] Most clustering algorithms are distance-based whereas some involve hierarchical clustering, K-means clustering, or a self-organizing map.[6,17]

Hierarchical clustering techniques are extensively used in microarray data analysis, which combines all data points into a single set which are placed adjacent to each other in the feature space.[11] At present, hierarchical clustering is the most often used method for grouping data.[13,14] The main objective of hierarchical clustering is to obtain a best cluster that will signify a set of patterns in the background of a given distance metric. This method permits biologists to visualize global expression patterns in DNA microarray data through graphic representation.[12] The hierarchical clustering method is classified into 2 distinct types: agglomerative (bottom-up) and divisive (top-down).[12,15] This method is based on a similarity or distance measure of the data, such as a correlation, Euclidean, squared Euclidean, or city-block (Manhattan) distance.[13] Normally, clusters are constructed using a hierarchical tree. This tree is created after calculating the distance between pairs of objects in the correlation matrix.[16] A quad tree (QT) is a tree data structure in which each internal node has up to 4 children. The quad tree and its different derivatives are considered to be the backbone for the storage, retrieval and analysis of spatial data. An efficient searching technique mainly depends on the height of the tree; an arbitrary insertion of the point features will make the tree unbalanced and will increase the time of searching.

From the review it is obvious that clustering algorithms used in previous research have partitioned the data, where each gene belongs to only one cluster. Some clustering algorithms that can allow each gene to only be in one cluster, including the k-means algorithm, hierarchical clustering algorithm, biclustering algorithm, fuzzy k-means algorithm and Self-Organizing Map (SOM). These methods have some disadvantages while working with microarray gene expression data because of the high complexity of biological processes. The nature of proteins and their interaction is a major reason for this. The genes that generate proteins are expected to express in more than one group of genes because proteins generally perform unique biological roles by interacting with other groups of proteins. This explains the inclusion of a gene in more than one cluster of microarray gene expression data. The QT has focused on the clustering of gene expression data with the expectation maximization (EM) algorithm, which estimates incomplete data samples. It also validates the cluster and measures the representation of objects in the clusters. The algorithm is compared with the k-means algorithm for performance validation.[28] Hierarchical biclustering suffers by selecting a depth of the tree, so the use of cross-validation measures and automatic depth selection capability will be needed. Each sample chooses a single path, which leads to a feasibility problem. The pathway enrichment among genes is activated through the same edges by introducing the activation model. Whenever the sample increases, the depth of the tree also increases, leading to computation complexity.[31] The large samples represented in the QT may reduce its complexity by storing a large amount of samples

at each level. A 2-dimensional hierarchical clustering approach was introduced by effectively representing the existence of genes in more than one cluster. The proposed approach was used to speed up the clustering process. It uses the QT-based data structure to find the closest pair, which also reduces the processing time.[27]

Quad Tree (QT)

A QT is a tree-like data structure, where each inner node has exactly 4 children. The QT and its different derivatives are mainly used for the purpose of storage, recovery and analysis of spatial data.[26] A QT is used to index data records in a spatial database in accordance with spatial location. Most businesses and government agencies may wish to query the database using spatial constraints; in such cases, the QT facilitates the storage and querying of spatial data. The QT can also be used to index data records in multiple spatial databases. The QT index values are used to search across the different databases by extracting components such as associations, patterns, clusters, outliers, or nearest neighbors.

The point QT is constructed successively by adding data points one by one. Initially, a point search is performed to add a point. In the QT, if there is no point related to target point (the point that has to be added), then the target point is inserted into the leaf node, where the search has to end.

The QTs is designed to be faster in various applications and the performance depends on the distribution of data in the domain. The height and shape of the point QT greatly relies on the insertion sequence. Until now, no significant effort has been taken to overcome this height balance problem, or to make the QT search more efficient.[17]

Review of Research

A handful of research studies have been presented for clustering microarray gene expression. Their data are illustrated below.

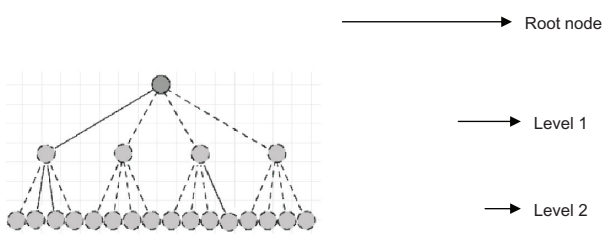

Figure 1. Sample structure of the QT.

Kim and colleagues[10] address a wide range of problems such as categorization of disease subtypes and tumors in biological and medical research. They describe the microarray, which has emerged as the most effective and broadly used tool for this categorization. The main objective of analyzing gene expression data has been to isolate data samples or genes. Identical expression patterns and statistical techniques exist to analyze and organize these complex data in a meaningful way. They have discovered that normalization, extent of noise and clarity in the datasets will change the clustering methods that are most commonly used in the analysis of microarray data.[10]

The microarray technique in concurrent measurement of the expression level in thousands of messenger RNA (mRNA)s has been enabled. This has been made possible by mining the data; it is feasible to recognize the dynamics of a gene expression time series in this manner. They have decreased the dimensionality of the data set by employing Principal Component Analysis (PCA). Examination of the components has provided an approach into the underlying factors calculated in the experiments. PCA has demonstrated that it is proved from their consequences that all rhythmic content of data can be decreased to three main components.[19]

Hereditary inclusion body myopathy (HIBM) of adult start steadily rising distal and proximal myopathy has also been discussed.[20] After examining the expression outline data sets by the overlap of three statistic methods (Student's *t*-test, TNoM and Info score), it has been found that the HIBM-specific transcriptome contains 374 differentially expressed genes. With the delicate contribution of mitochondrial processes exposed in HIBM, an unexpected feature of HIBM pathophysiology has been discovered. This could be expanded to provide reasons for the slow development of this disorder, and afford some understanding of its disease mechanism.

The main objective of gene expression analysis is to comprehend the processes of regulatory networks, as well as what pathways are restricted during inter-cellular and intra-cellular activities. Currently, microarray datasets are broadly used for this purpose. By employing their algorithm on a yeast speculation dataset, D'Souza and colleagues[21] have demonstrated that their algorithm can detect gene networks with reasonable ease.

A firm gene selection and efficient cancer prediction structure called SGS has been introduced. This structure first recognizes gene groups in which the genes have high correlation coefficient by means of a clustering algorithm. Finally, a prediction model is constructed based on shrinkage gene space, using a capable classification algorithm (such as *Support Vector regression* (SVM), 1-nearest neighbor (1NN), or regression). By means of the trial results obtained on real-world data, the structure has been shown to regularize highly available feature selection and prediction methods, such as Significant Analysis of Microarray (SAM), *Information gain* (IG) and the Lasso-type prediction model.[22]

A Fast Two-Dimensional Hierarchical Clustering

The recent arrival of microarray technologies has permitted biologists to simultaneously monitor the behavior of numerous genes, which produces large quantities of complex data. A large amount of gene expression data has been generated continuously using microarray experiments. It has been demonstrated that gene expression data containing vital information is very useful in medical diagnosis, therapy, and drug design. The aim of the proposed technique is to examine such data in order to obtain this essential information. Cluster analysis has played a key role in analyzing gene expression data. The main aim of clustering is to partition a set of objects into clusters, so that objects in a group are more analogous to one another than to objects in other clusters. Several clustering algorithms have been used to find co-expressed genes, but the processing time of these techniques are very high. Hence, our aim is to reduce the processing time of the clustering analysis. In order to solve the aforementioned problem, in this research, a fast, novel, semi-supervised, 2-dimensional QT-based hierarchical clustering technique is proposed. Here, the clustering elements are selected from the microarray gene expression database by means of the index matrix, and these elements are clustered using a QT-based 2-dimensional clustering technique. After the clustering process, the best 'k' clusters are identified using a fitness evaluation. Then, the closest index of all best 'k' clusters is calculated, and is used to obtain the next set of clustering elements from the database. This process is repeated 'r' times until the optimum cluster is found.

QT-based two dimensional clustering

In cluster analysis, groups of analogous (similar) objects are identified by maximizing inter-group similarity and minimizing intra-group similarity. Cluster analysis partitions data into significant or useful groups (clusters). It also plays a key role in evaluating the gene expression data. Clustering techniques are very useful for understanding basic biological processes. There are many popular and robust techniques for clustering gene expression data, with the aim of describing different functional roles of genes in biological processes. It also describes genes in the cluster with similar expression (co-expressed genes), which serve similar functional roles in a process. In this investigation, we used a QT-based technique to expedite the 2-dimensional hierarchical clustering. Figure 2 illustrates the overall process of the proposed system in a single iteration. The clustering elements are selected from the gene expression database using the random values of an index matrix. Each 'q' type gene representation of 'Q' elements is clustered into 4 clusters, by means of QT-based vertical dimensional clustering, and then clustered into 2 clusters using QT-based horizontal clustering.

Genes clustering: vertical dimension

The pseudo-code for QT-based vertical dimensional clustering techniques explains how the genes are clustered using QT-based vertical dimensional clustering. Initially, the value of the root node of the QT is set to 'Null', and 4 values are selected randomly from the first type gene expression elements and inserted as a child of the root node. In the function Rchild (n, Root, P), the 'n' is the randomly selected data to be inserted, 'Root' is the parent node and 'P' is the position of child in the root node. Then, the element to be inserted is selected from the first type gene expression data, and to find its insertion position in the QT, its minimum Euclidean distance to the first level nodes that is a child of the root node is calculated. Subsequently, the node is inserted as a child node of the closest node that has a minimum distance value. The validity of the parent node is evaluated while inserting a child node to check whether the weight of parent node is less than 4 or not. If the above condition is satisfied, then the element is inserted as a child node, or it will

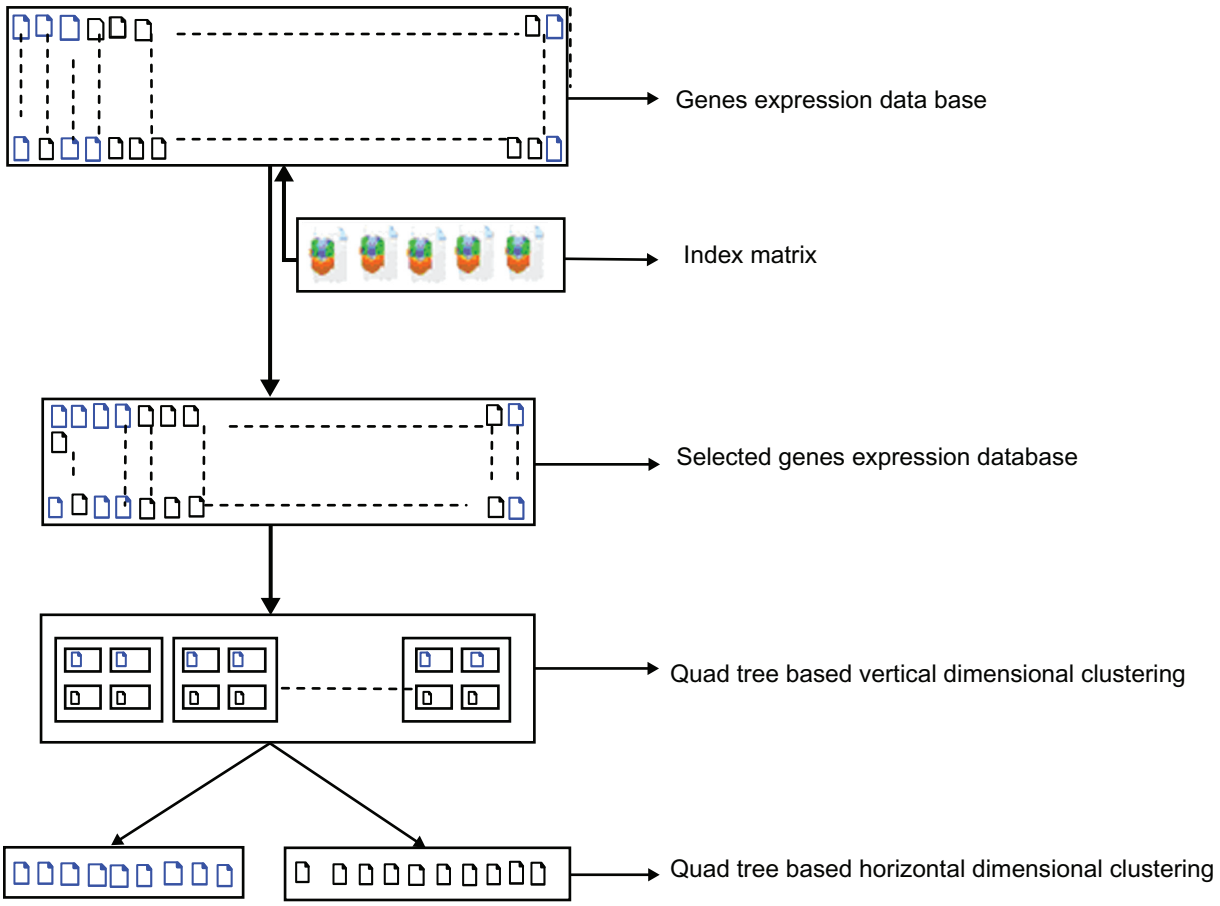

Figure 2. Process of QT based 2-dimensional clustering.
Notes: Let X_{ab} be a database that contains 'a' gene representation of 'b' clustering elements and $x = \{x_{ij} | x_{ij} \in X\}\ 1 < i \le q; 1 < j \le Q$ be the 'q' type gene representation of 'Q' elements, selected randomly from the database X_{ab} using the index $Y = \{Y_{ij} | Y < A \forall i, j\}\ 1 < i \le m; 1 < j \le n$. Each value in the 'Y', which represents the row index value of Database X_{ab}, must be unique and less than the maximum number of gene representations 'A' in the database X_{ab}.

find the closest node at the subsequent levels of the QT by calculating minimum distance. While inserting a node, if the parent node value is less than the child node to be inserted, then both the values are swapped. Thus, every element in the first type gene expression values are inserted into the QT and then clustered into 4 groups. Similarly, all gene type elements are clustered into 4 groups using the QT structure.

Input: Clustering elements
Output: The resultant solution clusters
Parameters:
Root → Root node
P → Position
L → Depth of QT
l → level
Nchild → Number of child node

Pseudo code

```
Set Root=NULL
c = 1;
For i = 1 to Q
For j = 1 to q
  For P = 1 to 4
    Set n = Random (x)
    Insert Rchild (n, Root, P)
  End For
  While c < L
    l = 1;
    For k = 1 to 4
      Set E_k = sqrt((x_ji − R_lk)^2)
      Set s = findmin (E_k)
  End For
  If Nchild ((s)) < 4
    Insert child (x_ji, R_lk (s), P, l);
    Exit While
  Else
    Movedown (l, R_lk (s));
  End If
```

End While
End for
End for

Genes clustering: horizontal dimension

Every gene value is clustered inner-wise into 4 clusters according to their distance, and then these clusters are again clustered using a horizontal dimension for analyzing the gene values. The distance is calculated as follows.

$$E_k = \sqrt{(x_{ji} - R_{lk})^2}$$

where, x_{ji} are the clustering elements in 'x' and R_{lk} is the clustered element in the kth level of the QT. The distance between 2 elements x_{ji} and R_{lk} is the quantity of the root of, sum of, square of, deviation among x_{ji} and R_{lk}. The first level of the horizontal-wise hierarchical clustering starts with the selection of 2 elements having greatest distance and inserted into the root node. Using Euclidian distance, the closest pair node of the next-gene representation element in the kth level is found and inserted into the corresponding node if the weight of the node is less than 4. Otherwise, it will find the closest pair of elements in the subsequent levels and insert the node here. Finally, every gene type is clustered based on QT.

Fitness evaluation

Let C be the resultant cluster and the fitness of C be calculated using the following equation.

$$\frac{1}{0.1 + \sum w(C)}$$

where $w = \begin{cases} 1 & if\ C[i] = R^{def} \\ 0 & otherwise \end{cases}$ is the weight of the each clustering element and R^{def} is the defined cluster used for the semi-supervised hierarchical clustering. If an element in the consequential cluster is in the defined cluster R^{def}, at that moment the weight of the element will be 1 and will be 0 otherwise. The 2-dimensional clustering progression and fitness evaluation are processed for every row of the index 'I' and $K = \{C_i \mid 1 < i \le l\}$ is the resultant cluster. From the resultant cluster set 'K', the best clusters having the highest fitness value are selected and the contiguous index of all clusters are computed, which are then used to get the next set of clustering elements from the database. The clustering is performed repeatedly until an optimum cluster is produced using this cluster and the gene is analyzed.

Experimental Results

The proposed technique is implemented in the working platform of MATLAB 7.11 with the system configuration Intel(R) Core(TM) i5 CPU, 650@3.20GHz, 3.19 GHz, 3.17 GB of RAM and it is evaluated using the microarray gene expression data of human acute leukemia. The 2 datasets of standard leukemia for training and testing is obtained from ALL/ALM datasets[23] and the performance of the proposed technique on clustering the ground truth data of the cancer classes, namely, acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL) are demonstrated. The 2 training leukemia datasets are again partitioned and become 4 sets, such as dataset_1, dataset_2, dataset_3 and dataset_4, each having N values (20, 18, 18, 17) respectively. This high-dimensional training dataset is subjected to clustering to analyze the occurrence of microarray genes in more than one cluster. In this clustering technique, an adaptive approach is used to represent the number of clusters that are dynamically generated from the microarray gene expression dataset.

The proposed technique is a multi-stage clustering technique, which performs clustering at diverse levels. Initially, the gene values in every gene type are clustered vertically into 4 groups and then these gene types are clustered horizontally for further analysis. In the inner gene clustering, initially the root node is created and then the 4 clustering elements that are selected randomly from the first type gene expression data. These are inserted into the root node. Then the insertion position of next element to be inserted is found by calculating its minimum Euclidian distance from the first level nodes. This element is then inserted into the corresponding node as a child whose weight is less than 4. If the weight of the node is greater than or equal to 4, then the closest node at the subsequent level of the QT is found by calculating the minimum distance. This process is repeated until all elements are clustered. Thus, every gene representation value is clustered into 4 clusters. The clustering output of dataset_1, dataset_2, dataset_3 and dataset_4 are shown in Figures 3–6.

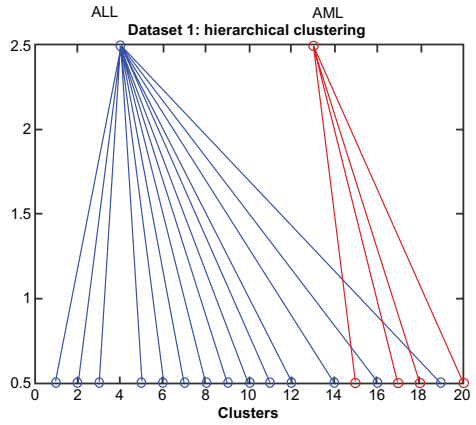

Figure 3. Dataset 1: hierarchical clustering.

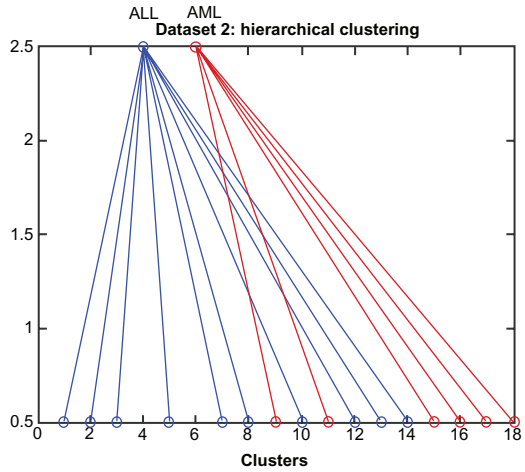

Figure 4. Dataset 2: hierarchical clustering.

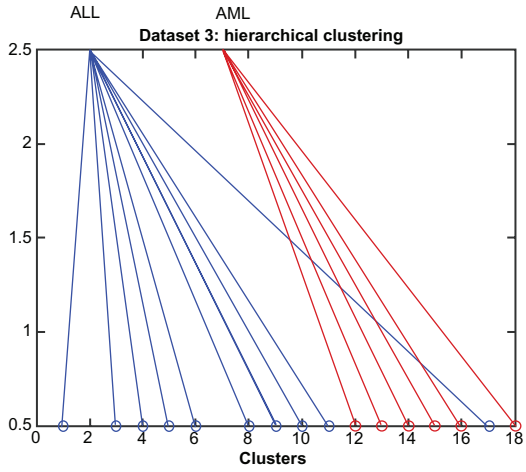

Figure 5. Dataset 3: hierarchical clustering.
Notes: The experimental results show hierarchical clustering with 4 data sets. The X-axis represents the number of clusters whereas the Y-axis represents the hierarchical level. The AML and ALL data in the graph are resided in more than one cluster so the proposed technique shows that it is a more efficient algorithm than those present in the literature.

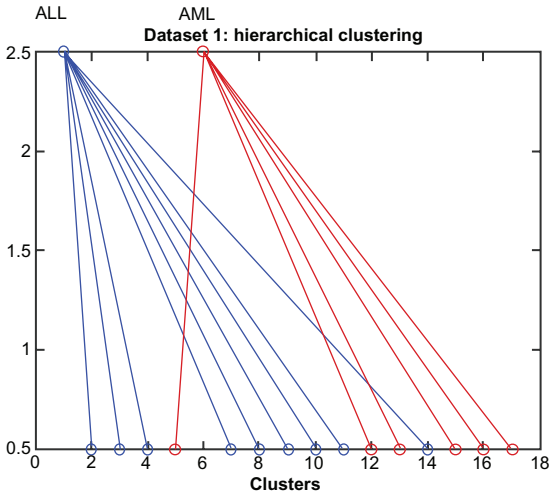

Figure 6. Dataset 4: hierarchical clustering.
Notes: This experiment is done using 2 dataset such as AML and ALL. The 2 training leukemia dataset are partitioned again and turned to 4 set (dataset_1, dataset_2, dataset_3 and dataset_4) each having N values (20, 18, 18, 17) respectively. The result of these 4 dataset indicates that both AML and ALL are crossing each other that means AML have ALL and vice versa.

Comparison of Existing Algorithms

The comparison of the algorithms is related to their performance, as shown in Figure 7. The performance of the SOM is less whenever the number of clusters is high. The hierarchical and SOM clustering algorithms provide better results for small data sets. The expectation of maximization and K-Means algorithms are very suitable for the large datasets. These algorithms may suffer to a degree due to noise in the datasets.[29] The above algorithms provide the clustered output but are lacking in several ways. In this paper, the 2-dimensional QT-based clustering technique is proposed for grouping gene expression data, with better performance than other clustering algorithms. It also finds genes in more than one cluster using a novel technique for achieving high accuracy. In Figure 7, the X-axis represents various clustering algorithms. The Y-axis represents the perfor-

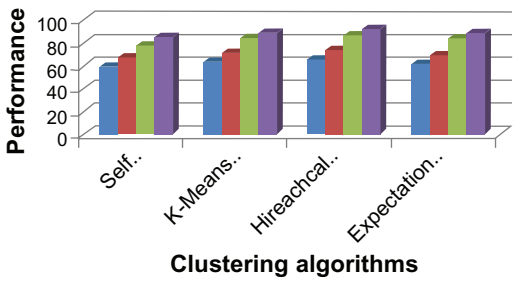

Figure 7. Performance comparison of existing algorithms.[29]

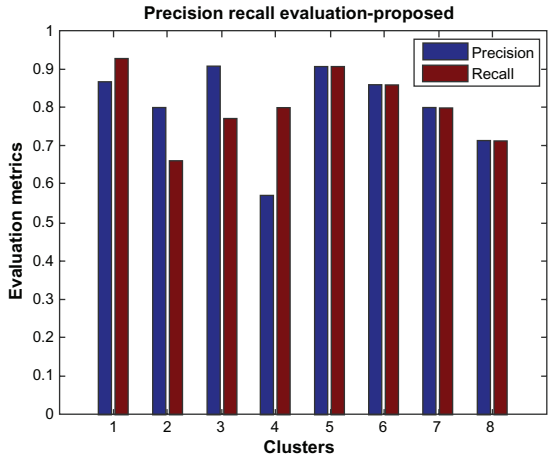

Figure 8. Precision, recall evaluation of proposed technique.
Notes: Figure 8 describes the precision and recall evaluation of the proposed technique. The X-axis represents the number of clusters whereas the Y-axis represents the evaluation metrics, ie, value of precision and recall. Figure 9 describes that the conventional method of hierarchical clustering. The precision and recall values are less when compared to the proposed technique. For example, in the conventional method, cluster 1 has the precision and recall values of 0.5 and 0.2, respectively. In the proposed technique, cluster 1 has the precision and recall values of 0.8667 and 0.9286, respectively. All clusters have a higher value than the conventional method, indicating that the proposed method is faster.

mance of these algorithms with many parameters. Each algorithm was analyzed based on input data sets as well as number of clusters (ie, 8, 16, 32 and 64) taken for the observation. The hierarchical clustering algorithm has higher performance than other algorithms, but it suffers in large datasets, so the novel 2D QT based hierarchical clustering has been proposed.

Performance Evaluation

The performance of the proposed 2-dimensional hierarchical data clustering technique is evaluated by

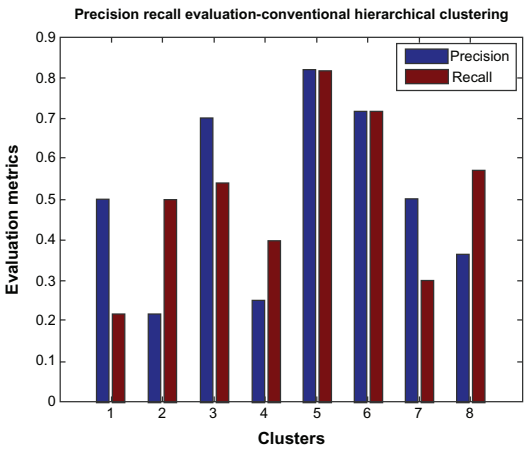

Figure 9. Precision, recall evaluation of the conventional technique.

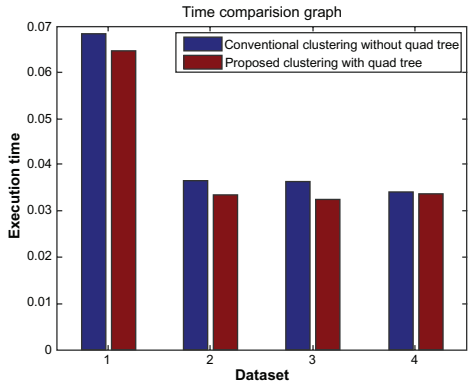

Figure 10. Time comparison graph.

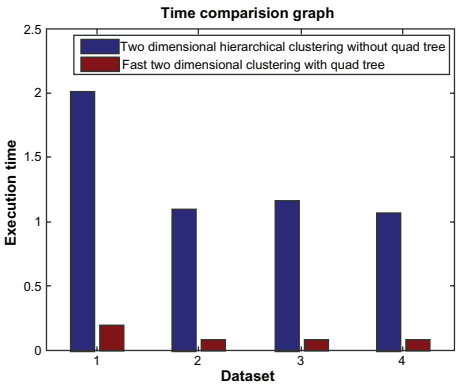

Figure 11. Time comparison graph for 2-dimensional hierarchical clustering without QT and with QT.

Table 1. Precision, recall and F-measure values of the clusters using the proposed technique.

Dataset	Cluster	Precision	Recall	F-measure
Dataset_1	C1	0.8667	0.9286	0.8966
	C2	0.8000	0.6667	0.7273
Dataset_2	C3	0.9091	0.7692	0.8333
	C4	0.5714	0.8000	0.6667
Dataset_3	C5	0.9091	0.9091	0.9091
	C6	0.8571	0.8571	0.8571
Dataset_4	C7	0.8000	0.8000	0.8000
	C8	0.7143	0.7143	0.7143

Table 2. Precision, recall and F-measure values of the clusters using conventional hierarchical clustering.

Dataset	Cluster	Precision	Recall	F-measure
Dataset_1	C1	0.5000	0.2143	0.3000
	C2	0.2143	0.5000	0.3000
Dataset_2	C3	0.7000	0.5385	0.6087
	C4	0.2500	0.4000	0.3077
Dataset_3	C5	0.8182	0.8182	0.8182
	C6	0.7143	0.7143	0.7143
Dataset_4	C7	0.5000	0.3000	0.3750
	C8	0.3636	0.5714	0.4444

clustering ground truth data of cancer classes, namely, acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL) using precision, recall and F-measures. Subsequently, these values are compared with the precision, recall and F-measure values of conventional hierarchical clustering. Precision and recall values of the clusters obtained by the proposed technique are given in the Table 1 and Figure 8 illustrate the corresponding graph. Precision and recall values of the clusters obtained by the Conventional Hierarchical Clustering are given in Table 2 and Figure 9 illustrates the same. We have used the precision, recall and F-measures described by Larsen and colleagues[24] and Steinbach and colleagues[25] to evaluate the performance of the proposed incremental text-clustering approach.

$$\text{Precision}\ (i, j) = M_{ij}/M_j$$

$$\text{Recall}\ (i, j) = M_{ij}/M_i$$

$$\text{F-Measure}\ (i, j) = \frac{2 * \text{Recall}\ (i, j) * \text{Precision}\ (i, j)}{\text{Precision}\ (i, j) + \text{Recall}\ (i, j)}$$

where

- M_{ij} is the number of members of gene i in cluster j,
- M_i is the number of members of cluster j,
- M_j is the number of members of gene i.

The performance of 2-dimensional hierarchical clustering without QT and with QT is also evaluated by comparing its processing time. In dataset_1 the processing time of 2-dimensional hierarchical clustering without QT and with QT is (2, 0.2). Likewise, dataset_2, dataset_3 and dataset_4 have processing times of (1.2, 0.1), (1.3, 0.1), (1.2, 0.1). Their processing time differences are 1.8, 1.1, 1.2 and 1.1 respectively. Figure 10 shows the comparison between conventional and proposed techniques based on QT. Figure 11 illustrates the comparison between 2-dimensional hierarchical clustering and fast 2-dimensional hierarchical clustering based on QT.

Conclusion

A novel and fast 2-dimensional hierarchical clustering technique has been proposed to deal with microarray genes that are present in more than one cluster. Initially, a set of clustering elements are selected randomly from the microarray gene expression data by using the index matrix, then they are clustered. The objective is to evaluate the fitness for selecting the best 'k' clusters from the various clusters. The next set of clustering elements is selected by finding the closest 2 cluster indices among the best 'k' clusters. These clusters are combined until the best clusters are found. The resultant genes are expressed in an efficient manner by eliminating biological complexities during the clustering process. In the existing techniques, each and every gene expression database was analyzed to find the closest pair, whereas the QT-based data structure uses some specific set of gene expression databases. The time comparison graph shows the average processing time of 2-dimensional hierarchical clustering without QT, which is 1.3 times larger than the proposed method. The proposed technique is faster when compared to existing clustering techniques in terms of performance. The experimental results based on real datasets have demonstrated that the proposed technique is truly more robust and efficient than traditional hierarchical clustering. In the future, the 2D QT algorithm will be implemented in various application areas in a more efficient manner.

Acknowledgements

The Authors express their sincere thanks to the Department of Information Science and Technology, Anna University, Chennai for providing necessary facilities to conduct their research.

Author Contributions

Conceived and designed the experiments: PR, SS. Analyzed the data: PR. Wrote the first draft of the manuscript: PR. Contributed to the writing of the manuscript: PR, SS. Agree with manuscript results and conclusions: PR, SS. Jointly developed the structure and arguments for the paper: PR, SS. Made critical revisions and approved final version: PR, SS. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Xue R, Li J, Streveler DJ. Microarray gene expression profile data mining model for clinical cancer research. In: *Proceedings of the IEEE 37th Hawaii International Conference on System Sciences.* Hawaii; 2004.
2. Cvek U, Trutschl M, Stone R II, et al. Multidimensional visualization tools for analysis of expression data. *World Academy of Science, Engineering and Technology*. 2009;54:281–9.
3. Kim SY, Choi TM. Fuzzy types clustering for microarray data. *World Academy of Science, Engineering and Technology.* 2005;4:12–5.
4. Wu X, Chen Y, Brooks BR, Su YA. The local maximum clustering method and its application in microarray gene expression data analysis. *Eurasip Journal on Advances in Signal Processing*. 2004;1:53–63.
5. Beal MJ, Krishnamurthy P. Gene expression time course clustering with countably infinite hidden markov models. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence.* 2006;16:13–6.
6. Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics.* 2006;22:1988–97.
7. Fadiel A, Naftolin F. Microarray applications and challenges: a vast array of possibilities. *Reproductive Sciences*. 2003;1:1111–21.
8. Liang J, Kachalo S. Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. *Chemometrics and Intelligent Laboratory Systems*. 2002;62:199–216.
9. Gruzdz A, Ihnatowicz A, Siddiqi J, Akhgar B. Mining genes relations in microarray data combined with ontology in colon cancer automated diagnosis system. *Proceedings of World Academy of Science, Engineering and Technology*. 2006;16:140–4.
10. Kim SY, Lee JW, Bae JS. Iterative clustering algorithm for analyzing temporal patterns of gene expression. *World Academy of Science, Engineering and Technology*. 2005;4:8–11.
11. Wang R, Scharenbroich L, Hart C, Wold B, Mjolsness E. Clustering analysis of microarray gene expression data by splitting algorithm. *Journal of Parallel and Distributed Computing*. 2003;63:692–706.
12. Do JH, Choi DK. Clustering approaches to identifying gene expression patterns from DNA microarray data. *Molecules and Cells*. 2008;25:279–88.
13. Kalocsai P, Shams S. Visualization and analysis of gene expression data. *Journal of the Society for Laboratory Automation and Screening.* 1999;5: 58–61.
14. van der Laan MJ, Pollard KS. A new algorithm for hybrid clustering of gene expression data with visualization and the bootstrap. *Journal of Statistical Planning and Inference.* 2003;117:275–303.
15. Trepalin SV, Yarkov AV. Hierarchical clustering of large databases and classification of antibiotics at high noise levels. *Algorithms*. 2008;1:183–200.
16. Tuncbag N, Haliloglu T, Keskin O. Correspondence between function and interaction in protein interaction network of Saccaromyces cerevisiae. *International Journal of Biological and Life Sciences*. 2006;1:167–74.
17. Lee M, Kim YM, Kim YJ, Lee YK, Yoon H. An ant-based clustering system for knowledge discovery in dna chip analysis data. *World Academy of Science, Engineering and Technology.* 2007;29:261–6.
18. Kim SY, Hamasaki T. Evaluation of clustering based on preprocessing in gene expression data. *International Journal of Biological and Life Sciences*. 2008;3:48–53.
19. Layana C, Diambra L. Dynamical analysis of circadian gene expression. *International Journal of Biological and Life Sciences*. 20078;3:101–5.
20. Eisenberg I, Novershtern N, Itzhaki Z, et al. Mitochondrial processes are impaired in hereditary inclusion body myopathy. *Human Molecular Genetics.* 2008;17:3663–74.
21. D'Souza RGL, Sekaran C, Kandasamy A. A phenomic algorithm for reconstruction of gene networks. *International Journal of Biological and Life Sciences.* 2009;4:76–81.
22. Jing L, Ng MK, Zeng T. Novel hybrid method for gene selection and cancer prediction. *World Academy of Science, Engineering and Technology*. 2010; 62:482–9.
23. Broad Institute. Gene Pattern. ALL/AML datasets. Available at: http://www.broadinstitute.org/cancer/software/genepattern/datasets/. Accessed Oct 15, 2012.
24. Larsen B, Aone C. Fast and effective text mining using linear-time document clustering. In: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Diego; 1999:16–22.
25. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. In: *Proceedings of the KDD*-2000 *Workshop on Text Mining*. Boston; 2000:109–11.
26. Chakraborty A, De SK, Dasgupta R. Balancing of quad tree using point pattern analysis. *World Academy of Science, Engineering and Technology*. 2011;52:118–21.
27. Priscilla R, Swamynathan S. *A Semi-Supervised Hierarchical Approach: Two Dimensional Clustering of Microarray Gene Expression Data*; 2011.
28. Leela Rani P, Rajalakshmi P. Clustering gene expression data using quad tree based expectation maximization approach. *International Journal of Applied Information Systems.* 2012;2:10–3.
29. Abbas OA. Comparison between data clustering algorithms. *International Arab Journal of Information Technology*. 2008;5:320–5.
30. Gruzdz A, Ihnatowicz A, Siddiqi A, Akhgar B. Mining genes relations in microarray data combined with ontology in a colon cancer automated diagnosis system. *Proceedings of World Academy of Science, Engineering and Technology*. 2006;16:140–4.
31. Caldas J, Kaski S. Hierarchical generative biclustering for microRNA expression analysis. *Research in Computational Molecular Biology.* 2010; 6044:65–79.