INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A NOVEL PROXIMITY MEASURE BASED CLUSTERING ALGORITHM FOR GENE EXPRESSION MICROARRAY DATA

G.Kalaivani[1], G.Vijayabharathi[2]

[1]M.Phil Research Scholar, PG & Research Dept. of Computer Science
Kaamadhenu Arts & Science College, Sathayamangalam - 638 503
Charuvani.g@gmail.com
[2]Associate Professor of Comp. Science, PG & Research Dept. of Computer Science
Kaamadhenu Arts & Science College, Sathayamangalam - 638 503
Viji_vijee@yahoo.com

Abstract:

Microarrays enable biologists to study genome-wide patterns of gene expression in any given cell type at any given time and under any given set of conditions. Identifying group of genes that manifest similar expression pattern is important in the analysis of gene expression in time series data. In the existing work, investigate the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures from time course and cancer datasets experiments.

1. Introduction

Biomedical engineering (BME) is the application of engineering principles and design concepts to medicine and biology for healthcare purposes (e.g. diagnostic or therapeutic). This field seeks to close the gap between engineering and medicine. It combines the design and problem solving skills of engineering with medical and biological sciences to advance healthcare treatment, including diagnosis, monitoring, and therapy.

Biomedical engineering has only recently emerged as its own study, compared to many other engineering fields. Such an evolution is common as a new field transitions from being an interdisciplinary specialization among already-established fields, to being considered a field in itself. Much of the work in biomedical engineering consists of research and development, spanning a broad array of subfields (see below). Prominent biomedical engineering applications include the development of biocompatible prostheses, various diagnostic and therapeutic medical devices ranging from clinical equipment to micro-implants, common imaging equipment such as MRIs and EEGs, regenerative tissue growth, pharmaceutical drugs and therapeutic biological.

Notable sub-disciplines of biomedical engineering can be viewed in two angles, from the medical applications side and from the engineering side. A biomedical engineer must have some view of both sides. As with many medical specialties (e.g. cardiology, neurology), some BME sub-disciplines are identified by their associations with particular systems of the human body, such as:

- Cardiovascular technology - which includes all drugs, biologics, and devices related with diagnostics and therapeutics of cardiovascular systems
- Neural technology - which includes all drugs, biologics, and devices related with diagnostics and therapeutics of the brain and nervous systems
- Orthopedic technology - which includes all drugs, biologics, and devices related with diagnostics and therapeutics of skeletal systems
- These examples focus on particular aspects of anatomy or physiology. A variant on this approach is to identify types of technologies based on a kind of pathophysiology sought to remedy apart from any particular system of the body, for example:
- Cancer technology - which includes all drugs, biologics, and devices related with diagnostics and therapeutics of cancer

But more often, sub-disciplines within BME are classified by their association(s) with other more established engineering fields, which can include (at a broad level):

- Biochemical-BME, based on Chemical engineering - often associated with biochemical, cellular, molecular and tissue engineering, biomaterials, and bio transport.
- Bioelectrical-BME, based on Electrical engineering and Computer Science - often associated with bioelectrical and neural engineering, bioinstrumentation, biomedical imaging, and medical devices. This also tends to encompass optics and optical engineering - biomedical optics, bioinformatics, imaging and related medical devices.
- Biomechanical-BME, based on Mechanical engineering - often associated with biomechanics, bio-transport, medical devices, and modeling of biological systems, like soft tissue mechanics.

One more way to sub-classify the discipline is on the basis of the products created. The therapeutic and diagnostic products used in healthcare generally fall under the following categories:

- Biologics and Biopharmaceuticals often designed using the principles of synthetic biology (synthetic biology is an extension of genetic engineering). The design of biologic and biopharma products comes broadly under the BME-related (and overlapping) disciplines of biotechnology and bioengineering. Note that "biotechnology" can be a somewhat ambiguous term, sometimes loosely used interchangeably with BME in general; however, it more typically denotes specific products which use "biological systems, living organisms, or derivatives thereof." [2] Even some complex "medical devices" (see below) can reasonably be deemed "biotechnology" depending on the degree to which such elements are central to their principle of operation.
- Pharmaceutical Drugs (so-called "small-molecule" or non-biologic), which are commonly designed using the principles of synthetic chemistry and traditionally discovered using high-throughput screening methods at the beginning of the development process. Pharmaceuticals are related to biotechnology in two indirect ways: 1) certain major types (e.g. biologics) fall under both categories, and 2) together they essentially comprise the "non-medical-device" set of BME applications. (The "Device - Bio/Chemical" spectrum is an imperfect dichotomy, but one regulators often use, at least as a starting point.)
- Devices, which commonly employ mechanical and/or electrical aspects in conjunction with chemical and/or biological processing or analysis. They can range from microscopic or bench-top, and be either in vitro or in vivo. In the US, the FDA deems any medical product that is not a drug or a biologic to be a

"device" by default (see "Regulation" section). Software with a medical purpose is also regarded as a device, whether stand-alone or as part of another device.

- Combination Products (not to be confused with fixed-dose combination drug products or FDCs), which involve more than one of the above categories in an integrated product (for example, a microchip implant for targeted drug delivery).
- Bioinformatics, is a field committed to the interpretation and analysis of biological data using computational techniques, has evolved tremendously in recent years due to the explosive growth of biological information generated by the scientific community.
- Bioinformatics is the science of managing, mining, integrating, and interpreting information from biological data at the genomic, proteomic, phylogenetic, cellular, or whole organism levels.
- The need for bioinformatics tools and expertise has increased as genome sequencing projects have resulted in an exponential growth in complete and partial sequence databases.

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse.

2. Literature Review

Comparisons and validation of statistical clustering techniques for microarray gene expression data, 2003 by susmita data and somnath data.

This paper offers some guidelines in the choice of a clustering technique to be used in connection with a particular microarray data set. In this work, selected six clustering algorithms of various types and evaluated their performance on a well known publicly available microarray data set on sporulation of budding yeast, as well as on two simulated data sets which are introduced in the next section. Of course, one can extend and modify this list of competing clustering algorithms to include his/her favorite algorithm. At least five of these algorithms are chosen to represent different classes of methods. Thus, well known algorithms such as Pam and Clara, both of which fall under partition methods, are not included in favor of including the K-means algorithm.

A comparative study of different machine learning methods on microarray gene expression data, 2007 by mehdi pirooznia, jack y yang, mary qu yang and youping deng.

Several classification and feature selection methods have been studied for the identification of differentially expressed genes in microarray data. Classification methods such as SVM, RBF Neural Nets, MLP Neural Nets, Bayesian, Decision Tree and Random Forrest methods have been used in recent studies. The accuracy of these methods has been calculated with validation methods such as v-fold validation. However there is lack of comparison between these methods to find a better framework for classification, clustering and analysis of microarray gene expression results. In this work, comparing these Classification methods of microarray gene expression is presented.

Clustering cancer gene expression data: a comparative study, 2008 by marcilio cp de souto, ivan g costa, daniel sa de araujo, teresa b ludermir and alexander schliep.

The use of clustering methods for the discovery of cancer subtypes has drawn a great deal of attention in the scientific community. While bio informaticians have proposed new clustering methods that take advantage of characteristics of the gene expression data, the medical community has a preference for using "classic" clustering methods. There have been no studies thus far performing a large-scale evaluation of different clustering methods in this context. Here present the first large-scale analysis of seven different clustering methods and four proximity measures for the analysis of 35 cancer gene expression data sets.

Analyzing gene expression time-courses, 2005 by alexander schliep, ivan g. Costa, christine steinhoff, and alexander scho¨nhuth.

Here present a robust and efficient approach to analyze gene expression time-course data with a mixture of hidden Markov models. The method can easily make use of prior knowledge about genes due to a partially supervised training procedure, which greatly increases robustness and the quality of the local optima found. Availability of such labels is a realistic assumption for the analysis of gene expression time-courses. Simultaneous analysis of cyclic and noncyclic time-courses is possible and neither missing values nor high levels of noise pose a serious problem. Mixtures are, for reasons of the complexity of gene function and regulation, a more appropriate model of biological reality than clustering.

3. MODULE DESCRIPTION

List of modules

- Correlation Coefficients Proximity measures
- Classical Measures
- Time-Course Specific Measures
- Intrinsic Biological Separation Ability
- Minimum Spanning Tree (MST) based clustering
- Performance evaluation

1. Correlation Coefficients Proximity measures

Considering gene expression data, two objects (genes or samples) are usually regarded as similar if they exhibit similarity in shape (trend), rather than in absolute differences from their values. Therefore, correlation coefficients have been widely used, as they capture such a type of similarity.

Pearson

The Pearson correlation coefficient (PE) allows the identification of linear correlations between sequences. Pearson may be sensitive to the presence of outliers, thus producing false positives, i.e., sequence pairs that are not alike, but receive a high correlation value.

Goodman-Kruskal

Goodman-Kruskal (GK) takes into account only the ranks of a and b. It is defined according to the number of concordant (S+), discordant (S-), and neutral pairs of elements in the sequences. In a concordant pair, the same relative order applies to both sequences, i.e., ai < aj and bi < bj or ai > aj and bi > bj. For discordant pairs, the inverse relative order applies, i.e., ai < aj and bi > bj or ai > aj and bi < bj.

Kendall

The Kendall correlation coefficient (KE) is based on the same building blocks used by Goodman-Kruskal. Note that, differently from GK, extreme correlation values are obtained only in the absence of neutrals.

Spearman

The Spearman correlation (SP) can be seen as a particular case of Pearson, provided that values of both a and b are replaced with their ranks in the respective sequences. As only the ranks of the sequences are considered, SP is more robust to outliers than Pearson. SP has also been employed to gene expression data though less often than Pearson.

Rank Magnitude

The Rank-Magnitude correlation coefficient (RM) was proposed as an asymmetric measure, for cases in which one of the sequences is composed of ranks and the other is composed by real numbers.

Weighted Goodman-Kruskal

The Weighted Goodman-Kruskal correlation coefficient (WGK) and takes into consideration ranks and magnitudes of both sequences.

2. Classical Measures

We review in the sequel four “classical” proximity measures that are also considered in our analysis. We anticipate that these measures have OðnÞ time complexity.

Cosine Distance

The cosine similarity and can be regarded as the normalized inner product between a and b. Note that the cosine similarity is related to Pearson and is sometimes referred to as uncentered correlation or angular separation. The cosine measures the angle between two data points with respect to the origin, whereas Pearson correlation measures this angle considering the mean of the data

Minkowski Distance

One of the most popular proximity indices that measures dissimilarity between two data points is the Minkowski distance metric

3. Time-Course Specific Measures

We review proximity measures specifically proposed for the clustering of gene time-course experiments. For these measures, we define t = (t1; . . . ; tn) as the time instants at which each feature is measured for a gene.

Jackknife

The underlying idea behind the Jackknife (JK) correlation is to minimize the effect of single outliers on the final correlation value by removing one single element at a time from both sequences. If the sequences do not contain outliers, their correlation value remains stable; otherwise, their removal causes a decrease in their correlation, indicating that the sequences were correlated partly due to the presence of outliers.

Short Time-Series Dissimilarity

The Short Time-Series Dissimilarity (STS) was proposed and measures the distance between the n - 1 slopes that compound two gene time-series. For two genes a and b, STS is performed. The greater interval between the measurements, the lower its impact on the dissimilarity.

Local Shape-Based Similarity

Based on the observation that biological relationships between genes may be present in the form of local and possibly shifted similarity patterns, introduced the concept of Local Shape-based Similarity (LSS). LSS seeks the most similar subsequences of size k in sequences a and b. The minimum subsequence size is given by kmin, which is usually set to n - 2, allowing for two time instant shifts. Note that although subsequences must have the same sizes, they do not have to be aligned, thus allowing locally shifted similarity patterns.

YR1 and YS1 Dissimilarities

Based on the presumption that correlations may not capture all information contained in gene time series, previous work introduced two dissimilarities that combine different types of information along with correlation values.

4. Intrinsic Biological Separation Ability

The ISA can be computed only for data sets with a golden standard partition, i.e., data sets for which class labels are available. Note that for most gene clustering problems, as time-series data sets, no class labels are available. Therefore, we take advantage of the information provided by the GO to overcome the lack of labeled data and devise a new procedure to evaluate the ISA of a distance regarding the clustering of genes. This new procedure is called Intrinsic Biological Separation Ability (IBSA).

Instead of using class labels, our methodology employs external biological information (semantic similarities among genes) extracted from the GO. Note that since IBSA employs information from the GO to evaluate a particular proximity measure, it tends to favor proximity measures that are in agreement with GO external information. If the user is interested in finding a different type of structure in the data (not related with GO), another methodology should be selected and employed.

Given a data set with o objects (genes), we build a distance matrix D. Assuming that all the values of D are in the [0, 1] interval, all pairs of objects can be distinguished by the same binary classifier. In brief, object pairs are assigned to class 1 if the distance between them is smaller than or equal to a given threshold 1 in the [0, 1] interval and 0 otherwise. Applying this equation to all object pairs from a given data set (with a fixed threshold), we obtain a predicted solution based solely on the distances between object pairs. To build a desired solution for this classifier, the first step of our methodology consists in obtaining biological dissimilarities for all pairs of genes from the data set in hand, devising a biological dissimilarity matrix (B).

Considering the GO, several proximity measures can be employed to quantify the degree of concordance between the sets of terms that annotate any two genes. By combining dissimilarities that operate between sets of terms, it is possible to measure the degree of concordance between any two genes. Note that the methodology presented here is the same regardless of the biological similarity employed between genes. Therefore, we elaborate on the choice of the biological measure during the discussion of the Experimental Setup. Once a biological dissimilarity matrix is available, it can be interpreted as external information and fill the gap left by the lack of class labels. For a given biological dissimilarity matrix (B) with values in the [0, 1] interval, we proceed and build a desired biological solution, where 2 is a threshold in the [0, 1] interval. By applying to all pairs of objects from a given data set (with a fixed threshold), we obtain a desired biological solution, based on external information extracted from the GO

$$J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } B(i,j) \leq \phi_2 \\ 0 & \text{otherwise.} \end{cases}$$

5. Minimum Spanning Tree (MST) based clustering

The construction of sub tree or cluster begins with core edge in the MST *T (V, E)*, then the vertices of the sub tree will be removed from the data set (MST), so each cluster is constructed in the highest density region in the existing data set. In other words, a series of clusters are automatically generated from high density region to low density region. The clustering process will stop growing, if the distance of the new edge greater than the average edge weight of the MST *T (V, E).* This condition is used to guarantee the approximate evenly distribution of the vertices in each of the clusters or sub tree.

Algorithm:

Input: proximity measure values

Output: *n* Clusters

```
Let $n$ be the number of clusters
Let $Cn$ be the sub tree or cluster
1. Construct an MST $T(V, E)$ from proximity measure values
```

```
2. Compute the average edge weight (Ŵ) of the edges from MST T(V, E)
3. Find center C of the MST T (V, E) using eccentricity of points
4. n = 1, Cn = , Visited[V] = 0
5. Repeat
6. (a, b) = min (E) // finding core edge
7. While(distance (a ,b)<Ŵ) do
8. Cn = Cn { (a, b)} // new cluster formation begins
9. Visited[a] = n; Visited[b] = n
10. For each vertex Vi in Cn do
11. For each vertex Vj linked with Vi do
12. If distance (Vi , Vj ) < Ŵ and Visited[Vj] = 0 then
13. Cn = Cn { (Vi , Vj ) }; Visited[Vj] = n
// cluster or sub tree Cn growing
14. Remove all the clustered vertices and edges from T
15. n = n+1
16. Until all the vertices are visited in T (V, E)
17. Return n number of clusters
```

The proposed algorithm constructs MST T (V, E) from set of point S (line 1). Average edge weight of the edges in MST is computed (line 2). For the given MST T(V, E) Average edge weight (Ŵ) computed. Using the eccentricity of points, the central vertex is computed at line 3. Initially all the vertices in the MST are marked as unvisited, which are represented as Visited [V] = 0 at line 4. Next the core edge is identified (line 6) to form a new sub tree (line 8). Here the core edge is considered as a new initial cluster or sub tree. From this initial sub tree or cluster, cluster growth begins at line 8. The vertices linked with the core edge are marked as Visited vertex (line 9). For each vertex in the sub tree, add all the linked edges and vertices into the sub tree which satisfy the threshold value of average edge weight (line 13). The vertices or nodes which are included in the sub tree are marked as Visited vertex (line 13) and also removed from the MST (line 14). If the threshold condition is not satisfied then current cluster growth comes to an end, then the cluster number is incremented by 1 (line 15) and new sub tree or cluster will be created. The lines 6 through 15 in the algorithm are repeated until the entire vertex in the MST is marked as visited or the entire vertex in the MST are removed.

6. Performance evaluation

We are analyzing our proposed system approach with the existing system interms of performance effectiveness at different noise level. Measures are evaluated regarding their robustness to noise, considering data with different noise levels.

IBSA is the metric used to evaluate distances for the clustering of genes. ISA can only be computed for data sets with a standard partition, i.e., class labels. As class labels are usually unavailable for gene clustering (e.g., time-series data), the information provided by the Gene Ontology (GO) to overcome the absence of labeled data.

Predicted solution

Desired solution

$$J(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1 & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ belong} \\ & \text{to the same cluster} \\ 0 & \text{otherwise.} \end{cases}$$

$$TP = \sum_{\forall i,j,i \neq j} I_{\phi_1}(\mathbf{x_i}, \mathbf{x_j}) J_{\phi_2}(\mathbf{x_i}, \mathbf{x_j})$$

$$FP = \sum_{\forall i,j,i\neq j} I_{\phi_1}(\mathrm{x_i},\mathrm{x_j})(1 - J_{\phi_2}(\mathrm{x_i},\mathrm{x_j}))$$

$$TN = \sum_{\forall i,j,i\neq j} (1 - I_{\phi_1}(\mathrm{x_i},\mathrm{x_j}))(1 - J_{\phi_2}(\mathrm{x_i},\mathrm{x_j}))$$

$$FN = \sum_{\forall i,j,i\neq j} (1 - I_{\phi_1}(\mathrm{x_i},\mathrm{x_j}))J_{\phi_2}(\mathrm{x_i},\mathrm{x_j}).$$

4. Experimental Setup

Accuracy

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative+False positive+ False negative}}$$

- **TP (True positive)**

In a statistical hypothesis test, there are two types of incorrect conclusions that can be drawn. The hypothesis can be inappropriately. A positive test result that accurately reflects the tested-for activity of an analyzed. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP);

True positive rate (TPR) =TP/P
P= (TP+FN)
Where P is the positive. TP is the True Positive

- **TN (True negative)**

A result that appears negative when it should not. A true negative (TN) has occurred when both the prediction outcome and the actual value are n is the number of input data.

True negative rate (TNR) =TN/N
N= (TN+FN)
Where
N is the Negative value.
TN is the True Negative.

- **FP (False positive)**

A result that indicates that a given condition is present when it is not. However if the actual value is n then it is said to be a false positive (FP).

False positive rate (α) = FP / (FP + TN)

- **FN (False negative)**

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

False negative rate (β) =FN / (TP + FN)

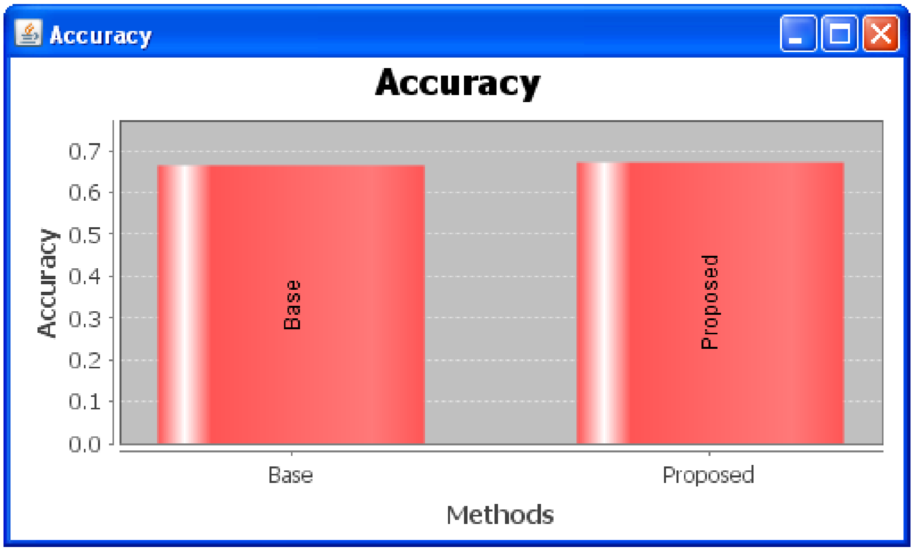

Fig.1. Accuracy rate comparison

This graph shows the accuracy rate of existing system such as Clustering technique for Gene Expression Microarray Data and proposed Minimum Spanning Tree (MST) based clustering algorithm based on two parameters of accuracy and methods such as existing and proposed system. In this graph, x axis will be the methods (existing and proposed system) and y axis will be accuracy in %. From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of proposed system is increased which will be the best one.

Error rate

Error rate can be calculated from formula given as follows

$$\text{Error rate} = \frac{\text{False positive} + \text{False negative}}{\text{True positive} + \text{True negative}+\text{False positive}+ \text{False negative}}$$

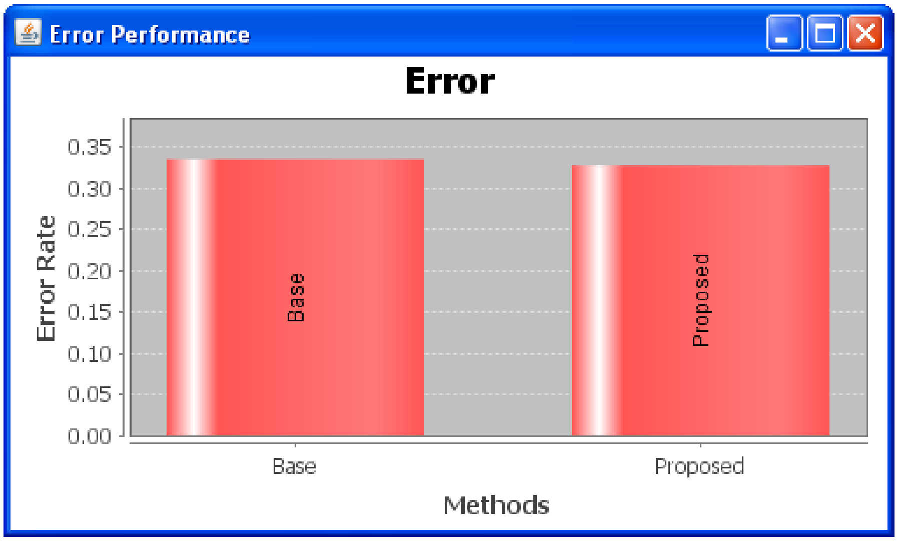

Fig.2. Error rate comparison

This graph shows the error rate of existing system such as Clustering technique for Gene Expression Microarray Data and proposed Minimum Spanning Tree (MST) based clustering algorithm based on two parameters of error rate and methods such as existing and proposed system. In this graph, x axis will be the methods (existing and proposed system) and y axis will be error rate. From the graph we can see that, error rate of the system is increased somewhat in existing system than the proposed system. From this graph we can say that the error rate of proposed system is reduced which will be the best one.

EXISTING SYSTEM

Here conducted a review and comparison of 16 proximity measures for the clustering of gene expression data. Here considered six correlation coefficients, four "classical" distances, and six proximity measures specifically proposed for the clustering of gene time-course data. Given their differences, we evaluated proximity measures separately for cancer and time-course experiments.

Apart from the comparison of proximity measures, here introduced a set of 17 time-course benchmark data along with a new methodology (IBSA) to evaluate distances for the clustering of genes. Both data sets and methodology can be used in future research to evaluate the effectiveness of new proximity measures in this particular scenario. IBSA can be employed to evaluate proximity measures regarding any gene clustering application, i.e., it is not restricted to gene time-course data, the scenario addressed here.

Here evaluate proximity measures independently of biases of clustering algorithms, because both ISA and IBSA allow the evaluation of proximity measures without any clustering algorithm. The main contributions can be summarized as follows: here compare proximity measures for the clustering of gene time-course data and cancer samples separately, as both scenarios have distinct characteristics.

Here evaluate a total of 16 proximity measures. For cancer data 10 measures are taken into account. For time-course data, all 16 measures are evaluated. For cancer samples, proximity measures are evaluated with respect to their ISA, as class labels are available. Here introduce a new methodology to evaluate proximity measures for the clustering of genes, called IBSA. IBSA employs external information extracted from the GO to overcome the lack of class labels in these data sets. Measures are evaluated regarding their robustness to noise, considering data with different noise levels.

PROPOSED SYSTEM

The proposed work, multidimensional gene expression data is represented using Minimum Spanning Tree (MST). A key property of this representation is that each cluster of the expression data corresponds to one sub tree of the Minimum Spanning Tree, which converts a multidimensional clustering problem to a tree partitioning problem. Each node represents one gene, and every edge is associated with a certain level of pheromone intensity, densities and the co-expression level between two genes. Minimum Spanning Tree based clustering algorithm aims to speed up the clustering process by using the alignment free similarity measures and is able to produce clustering result. We have applied Minimum Spanning Tree (MST) based clustering algorithm with proximity measure similarity methods.

Core Edge: Given an MST *T (V, E), core edge CE (E) is* defined as the edge which is having minimum length.

$$CE\ (E) = min\ (E)$$

Average Edge Weight: Given an MST *T (V, E), average edge weight* $\hat{W}$ is defined as the ratio between sum of the weight of the edges and total number of edges.

$$\hat{W}(E) = \sum_{i=1}^{n} \square \, |ei| \, / \, n$$

Based on the above definitions, the process of construction of sub trees (clusters) from the given data set represented in the form of an MST *T (V, E)* is described as follows:

1. Search for *core edge* CE1 (E) from the MST. It will be the base for first sub tree or cluster denoted by *C1*. For each vertex *Vi* in the sub tree, add the edges and the associated vertices *Vj* from the MST *T*(*V, E*) in to the sub tree, if the *distance* (*Vi* , *Vj*) is less than the *average edge weight*.

2. Do step 2 until no edges can be added into the sub tree.

3. Search for next core edge CE2 and then construct another sub tree *C2*. In this way series of sub trees (clusters) are generated as *C1, C2, C3…Cn.*

4. If any of the vertices not added to any of the clusters *C1, C2, C3 …..Cn* are considered as outliers, which can be determined based on distance from center of the MST *T*.

The construction of sub tree or cluster begins with *core edge* in the MST *T (V, E)*, then the vertices of the sub tree will be removed from the data set (MST), so each cluster is constructed in the highest density region in the existing data set. In other words, a series of clusters are automatically generated from high density region to low density region. The clustering process will stop growing, if the distance of the new edge greater than the average edge weight of the MST *T (V, E)*. This condition is used to guarantee the approximate evenly distribution of the vertices in each of the clusters or sub tree.

6. Conclusion

In the existing work, investigate the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures from time course and cancer datasets experiments. In order to improve the accuracy of clustering of gene expression data, we are proposing the novel clustering method. Our Minimum spanning tree based clustering algorithm does not require domain knowledge of the given problem. Our algorithm finds series of clusters C1, C2, C3…Cn. These clusters ensure guaranteed intra-cluster similarity. This algorithm does not require the users to select and try various parameters combinations in order to getthe desired output. The key feature of Minimum spanning tree based clustering algorithm is it fuses the advantages of both *density* and *graph based* clustering approaches.

7. Acknowledgement

As a future work we can use the weighted semantic features and cluster similarity is introduced to cluster meaningful topics from document set. The algorithm finds clusters and outliers with less computational time. The proposed algorithm gives better results than the existing methods. The running time of the algorithm is also less compared with the existing algorithm. The experiment results shows that the proposed system is well effective than the existing system in evaluating the effectiveness of the clustering of gene expression data.

REFERENCES

[1] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[2] A. Zhang, Advanced Analysis of Gene Expression Microarray Data, first ed. World Scientific, 2006.

[3] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Nat'l Academy Sciences USA, vol. 95, no. 25, pp. 14863-14868, 1998.

[4] L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," Genome Research, vol. 9, no. 11, pp. 1106-1115, 1999.

[5] P. D'haeseleer, "How Does Gene Expression Clustering Work?" Nature Biotechnology, vol. 23, no. 12, pp. 1499-1501, 2005.

[6] G. Kerr, H.J. Ruskin, M. Crane, and P. Doolan, "Techniques for Clustering Gene Expression Data," Computers Biology Medicine, vol. 38, no. 3, pp. 283-293, 2008.

[7] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[8] A.A. Alizadeh et al., "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling." Nature, vol. 403, no. 6769, pp. 503-511, 2000.

[9] S. Ramaswamy, K.N. Ross, E.S. Lander, and T.R. Golub, "A Molecular Signature of Metastasis in Primary Solid Tumors," Nature Genetics, vol. 33, no. 1, pp. 49-54, Jan. 2003.

[10] J. Lapointe et al., "Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer," Proc Nat'l Academy Sciences USA, vol. 101, no. 3, pp. 811-816, 2004.

[11] I.G. Costa, A. Scho¨nhuth, and A. Schliep, "The Graphical Query Language: A Tool for Analysis of Gene Expression Time- Courses," Bioinformatics, vol. 21, no. 10, pp. 2544-2545, 2005.

[12] J. Ernst, G.J. Nau, and Z. Bar-Joseph, "Clustering Short Time Series Gene Expression Data," Bioinformatics, vol. 21, pp. i159-i168, 2005.

[13] T.J. Hestilow and Y. Huang, "Clustering of Gene Expression Data Based on Shape Similarity," EURASIP J. Bioinformatics Systems Biology, article 12, 2009.

[14] A. Ben-Dor and Z. Yakhini, "Clustering Gene Expression Patterns," Proc. Third Ann. Int'l Conf. Computational Molecular Biology, pp. 33-42, 1999.

[15] E.P. Xing and R.M. Karp, "Cliff: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts," Bioinformatics, vol. 17, no. suppl 1, pp. S306-S315, 2001.

[16] R. Sharan, A. Maron-Katz, and R. Shamir, "Click and Expander: A System for Clustering and Visualizing Gene Expression Data," Bioinformatics, vol. 19, no. 14, pp. 1787-1799, 2003.

[17] X. Wu et al., "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

[18] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[19] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice- Hall, 1988.

[20] S. Datta and S. Datta, "Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data," Bioinformatics, vol. 19, no. 4, pp. 459-466, 2003.

[21] I.G. Costa, F.A.T.d. Carvalho, and M.C.P. de Souto, "Comparative Analysis of Clustering Methods for Gene Expression Time Course Data," Genetics Molecular Biology, vol. 27, pp. 623-631, 2004.

[22] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G.C. Tseng, "Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis," Bioinformatics, vol. 22, pp. 2405-2412, 2006.

[23] M. Pirooznia, J. Yang, M.Q. Yang, and Y. Deng, "A Comparative Study of Different Machine Learning Methods on Microarray Gene Expression Data," BMC Genomics, vol. 9, no. Suppl 1, article S13, 2008.

[24] M.C.P. de Souto, I.G. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering Cancer Gene Expression Data: A Comparative Study," BMC Bioinformatics, vol. 9, no. 1, article 497, 2008.

[25] E. Freyhult, M. Landfors, J. Onskog, T. Hvidsten, and P. Ryden, "Challenges in Microarray Class Discovery: A Comprehensive Examination of Normalization, Gene Selection and Clustering," BMC Bioinformatics, vol. 11, no. 1, article 503, 2010.