IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 3, Issue 3, June-July, 2015
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

Gene Data Analysis and Techniques: A Survey

R. Rajeswari[1], K. Upendra Babu[2], Dr. G. GunaSekaran[3]

[1]Research scholar, Department of Computer Applications,
St. Peter's University, Chennai.

[2]Research Scholar, Department of Computer science and Engineering
Manonmanim Sundaranar University

[3]Principal, Meenakshi College of Engineering, Chennai

Abstract—

Analysing gene data to uncover the hidden features thereby, predicting the functioning and behaviour of the gene. In this paper we have presented a new ant-based clustering algorithm for knowledge discovery. Without any pre assumptions regarding the number of clusters or the features of the clusters, thus a reliable clustering algorithm not relying on any pre assumed constants. This paper also provides a survey on the major approaches used for clustering gene data, while summarizing their features and drawbacks.

Keywords –Data mining, clustering, ant colony algorithm, swarm inteligence, bio infomatics, Gene expression data,

I. INTRODUCTION

Gene is a combination of DNA and RNA chains which controls the levels of proteins that form the structure of any living cell or tissue. This Gene protein controls the functionality, behavior and traits of the life cell while also storing the heredity information. A normal human gene sequence might consist of anywhere around 20,000 to 25,000 strands of gene protein. For the purpose of studying the gene, this sequence of strands is expressed in the form of Gene expression. This Gene expression in other words can also be described as sequence of data that is encoded by the different levels of different protein types that constitute the Gene. This encoded sequence control the functionality of the Gene that includes control functions, behavior, traits, and heredity information from the parent Gene. The role of each protein in this control mechanism has always been a mystery which we have been trying to understand for a long time. The earlier but fail safe method was to switch off one protein at a time and study the resultant behavior of the test sample to identify the role of the protein that had been switched off. This process could not be pursued on human genes due to the destructive nature of the testing and also that it was time consuming yielding very little or sometimes confusing results. Hence a new domain in bioinformatics for generating predictions based on data-intensive computational procedures. Clustering techniques has been proving itself as a reliable and useful technique over the years to solve this problem. Genes with similar expression patterns can be clustered together to reveal hidden information to help us understand the gene functions, gene regulation and cellular processes as whole.

II. LITERATURE SURVEY

Xiao Zhang et al 2012 propose several validation techniques for gene expression data analysis. Normalization and validity strategies are proposed to improve the prediction about the number of relevant clusters [1]. Guiquan Liu et al 2010 suggests a novel genetic programming clustering system for gene data based on hierarchical statistical model and an appropriate fitness function that can largely eliminate the

infection of data scale and dimension [2]. Wai-Ho Au et al 2005 proposes a method that groups interdependent attributes into clusters by optimizing a criterion function derived from an information measure. And also suggests that by applying this algorithm to gene expression data, meaningful clusters of genes can be discovered. [3]. Marco A. Mendez et al 2002 describes an innovative procedure that provides statistical support for gene clusters by a hierarchical clustering method while allowing the correct classification of genes within the groups. While using principal component analysis (PCA) and linear discriminate analysis [4]. Zhe Liu et al 2014 suggests a new scheme for clustering gene expression data based on the multivariate elliptical contoured mixture models with an improved expectation maximization (EM) algorithm that can adding or deleting initial value of the classical EM algorithm and the number of clusters as a known parameter. To overcome the problem of over-reliance on the initialization [5]. Harun Pirim et al 2012 suggest that the essential cellular molecules for a biological system to function and interact with its surroundings include DNA, RNA, proteins and metabolites, all of which are under physiological and environmental control. Many different interaction layers exist among these molecules such as PPI networks, I.E, Intractomes, Gene Regulatory Networks, bio-chemical networks and gene co-expression networks. Thus a powerful clustering approach as well as a predictive model is required to detect patterns or relationships in the expression data. However a predictive model should be guided by biological facts, meaning that results of predictive models should be validated by biological knowledge [6]. G. Kerr et al 2008 stresses on the need of a disciplined information-driven clustering algorithm that can integrate cluster and meta-information, thus providing a basis for validation independent of the current problem and also simplify interpretation of clustering results. [7]. Joon Jin Song et al 2007 proposes a innovative method based on functional data analysis to cluster time-dependent gene expression profiles using Hidden Markov Models that can take time dependent data into account [8]. Miin-Shen Yang n et al 2012 proposes a robust EM clustering algorithm for Gaussian mixture models, which construct a schema to automatically obtain an optimal number of clusters. Therefore, creating a new way to solve the initialization problems [9].Urszula Boryczka proposes an ant-based clustering algorithm, a particular kind of a swarm intelligent system, and on the effects on the final clustering by using during the classification different metrics of dissimilarity: Euclidean, Cosine, and Gower measures. Clustering with swarm-based algorithms is emerging as an alternative to more conventional clustering methods, such as e.g. k-means, etc. Among the many bio-inspired techniques, ant clustering algorithms have received special attention, especially because they still require much investigation to improve performance, stability and other key features that would make such algorithms mature tools for data mining [10]

III. EXISTING MODEL

Data mining and Clustering has evolved as a collection of different methodologies and techniques over the period of time, with a common goal of extracting meaning full predictions from a extensive data sets of gene microarrays. A microarray typically consists of a large number of DNA sequences from a collection of similar or different tissue samples. In this paper, we will focus on the cluster analysis methods used for analyzing gene expression data of DNA sequences to identify the role of a particular gene in the gene expression data. The different clustering techniques applicable to gene data can be categorized into few broad categories.

K-Means Algorithms

The K-means algorithm is a typical partition-based clustering method. The algorithm partitions the data set into K pre-determined disjoint subsets. Which uses the optimizing the following objective function:

$$E = \sum_{i}^{k} \sum_{O \in c_i} \lfloor O - \mu_i \rfloor^2$$

Where,

$O \Rightarrow$ is a data object
$C_i \Rightarrow$ is a cluster
$\mu_i \Rightarrow$ is the centroid of C_i

Thus, the objective function E tries to minimize the sum of the squared distances of objects from their cluster centers.The K-means algorithm is simple and fast. The time complexity of K-means is

$$O(l * k * n)$$

Where,

l ⇒ is the number of iterations

k ⇒ is the number of clusters.

It has been observed that the K-means algorithm typically converges after a fixed number of iterations.[11]

Self-Organizing Map

The Self-Organizing Map (SOM) is based on a single layered neural network. The data objects are represented as neurons organized within a simple neighborhood structure such as a two-dimensional m * n grid. The data objects are presented at both input and output. Each neuron of the neural network is associated with a reference vector, and each data point is mapped to the neuron with the closest reference vector. Each data object acts as a training sample for rest of the data in the algorithm, which directs the movement of the reference vectors towards the denser areas of the input vector space, so that reference vectors are trained to fit the distributions of the input data set. When the training is complete, clusters are identified by mapping all data points to the output neurons. It has been observed that SOM is relatively Robust while dealing with noisy data, and is capable of generating an appealing map of data sets in both 2D and 3D space.[12],[13]

Hierarchical Clustering

Hierarchical clustering generates a series of nested clusters that can represent in a tree format known as dendrogram. Wherein the root contains the entire data and leafs consist of individual data. Cutting the tree at some level provides a specified number of clusters, and by reordering the branches similar clusters can be placed together. Hierarchical clustering algorithms can be further divided into bottom-up approach (agglomerative) and top-down approach (divisive) based on how the hierarchical Tree is formed. The Bottom-up approach treats each data object as an individual cluster, and at each step, closest pair of clusters are merged together until entire data stet is merged into one single cluster. The Top-down approach treats entire data set as one single cluster, and at each step splits the cluster with similar data's until each cluster contains a single datum.[14]

Graph-Theoretical Approaches

A graph theoretical approach treats the entire data set as a graph. It considers each data as node (vertex) and the relation with other nodes (data) is represented as a weighted edge between the nodes. The weight of each edge is determined by the level of proximity between the nodes. Clustering methods group node based on the weight of the edges connecting them. Interesting patterns can be derived by mapping the edges as 0 or 1 based on a pre defined threshold levels for the weighs.[15],[16]

Model-Based Clustering

Model-based clustering approach views the data set as a finite mixture of probability distributions, with each datum corresponding to a different cluster. Data elements are mapped into clusters based on probabilistic function determining the probability that the specific data element belongs to a cluster.

$\Theta = \{\theta_i | 1 \leq i \leq k\}$

And

$$\Gamma = \{\gamma_r^i | 1 \leq i \leq k, 1 \leq r \leq n\}$$

And

$$L_{(mix)}(\Theta, \Gamma) = \sum_{i=1}^{k} \Upsilon_r^i f_i(x_r|\theta_i)$$

Where
n is the number of data objects
k is the number of components
x_r is a data object (i.e., a gene expression pattern),
$f_i(x_r|\theta_i)$ is the density function of x_r of component C_i
θ_i is the unknown set of parameters or model parameters
and γ_r^i Represents the probability that x_r belongs to C_i.

An important advantage of model-based approaches is that they provide an estimated probability γ_r^i that data object i will belong to cluster k, thus providing a highly adaptable method while mining highly connected data.[17],[18],[19]

A Density-Based Hierarchical Approach

Density-Based Hierarchical clustering DHC is developed based on the density and attraction of data objects. The entire data set is considered as a high-dimensional dense cluster, where data objects are attracted with each other. Once the density and attraction of data objects are defined, DHC maps the data set into a two-level hierarchical cluster structure. At the first level is an attraction tree that represents the relationship between the data objects. Each data objects become the node on the attraction tree and the attracting node becomes the patent node of the attracted node. Thus the node with highest density becomes the root and data objects with least attraction become the leaf node of this attraction tree. Initially, the entire data set is considered as a single cluster and is represented by the root node of the density tree. This single tree is then split into several sub-dense sub-trees based on predefined criteria where each sub-tree area is represented by a child node of the root node. These sub-trees areas are further split, until each sub-tree represents a single cluster. It has been observed that the density tree method is capable of detecting co-expressed genes even in noisy environment.[20],[21]

IV. DRAWBACKS

K-Means Algorithms

K-Means has several drawbacks as a gene-based clustering algorithm. First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of k and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical. Second, gene expression data typically contain a huge amount of noise; however, the K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise.[1a]

Self-Organizing Map

Self-Organizing Map requires users to input the number of clusters and the grid structure of the neuron map beforehand, and improperly specified parameters can prove disastrous in extracting the clusters. Furthermore, if the data set has lots of irrelevant data's this will result in lots of irrelevant clusters masking any interesting relevant clusters [2a]

Hierarchical Clustering

Hierarchical Clustering is very taxing in the computation complexity; construction of a tree requires at least $\frac{n^2-n}{2}$ splits or merge's resulting in time complexity of $(n^2 \log n)$.]. Furthermore, the greedy nature of hierarchical clustering prevents the refinement of the previous clustering i.e.. If a bad decision is made in

the initial steps, it can never be corrected in the following steps, and also it lacks robustness i.e. any small change in data will change the entire tree.[4a]

Graph-Theoretical Approaches

Graph-theoretical approaches can sometimes generate highly unbalanced partitions, furthermore it might fail to split highly co-expressed clusters and report them as one highly connected cluster.[5a]

Model-Based Clustering

Model based clustering relies on the assumption that the data set fits a specific distribution but this may not be true in many cases.[9a]

A Density-Based Hierarchical Approach.

The density tree structure would become too hard to interpret when the data set becomes large and the data structure becomes complicated. further while calculating the distance between each object pair to determine the density of data objects, this algorithm requires a computation complexity of $O(n^2)$ which makes this not so efficient algorithm, and furthermore the algorithm requires two global parameters to be set at the beginning and cannot be altered at a later stage, thus the reliability of the result is dependent on the pre defined parameters.[10a]

V. PROPOSED WORK

We have observed that most of the algorithms though are fast and robust but leaves a doubt regarding the reliability or correctness of the outcome, since they require us to determine certain initial parameters that determine the reliability of the entire algorithms. Inspired by the natural behavior of ants in finding the shortest path to the best food source even in the absence of any previous knowledge about the sources, we proposes a new clustering algorithm in this paper based on the ant colony to cluster the genes on the basis of the enzymes that can be used to detect a group of genes whose expression level changes in the same pattern. This is a population based Meta heuristic that can be used to find approximate solutions to difficult optimization problems. We expect that the algorithm can effectively uncover the hidden patterns for accurate identification of gene function and predicting its role in gene behavior.

The basic environment of the algorithm consists of randomly placed high-dimensional data objects, having several attributes in a bi-dimensional grid. In this method objects are placed as pick up drop method.

Ant Colony Algorithm:

- Ants (blind) navigate from nest to food source
- Shortest path is discovered via pheromone trails

$$\frac{n - ants * \frac{pheromone}{x-distance}}{\frac{distance_{(longer\ path)}}{time}} < \frac{n - ants * \frac{pheromone}{x-distance}}{\frac{distance_{(shoter\ path)}}{time}}$$

- Each ant moves at random
- Pheromone is deposited on path
- Ants detect lead ant's path, inclined to follow
- More pheromone on path increases probability of path being followed

Algorithm in Pseudo code:

Create construction graph
Initialize pheromone values
while not stop-condition do
    Create all ants solutions
    Perform local search
    Update pheromone values
end while End Do

For the purpose of gene mining we propose to implement this process in three steps

Step:1 Feature Extraction

In this stage the main features of objects are extracted and the method of comparison.

Step:2 Similarity Computation

The similarity between the objects taken into consideration in term of these chosen features attributes.

Step:3 Grouping

The result of similarity or dissimilarity computation is presented in the next step grouping, the form of partitioning these objects into groups.

VI. CHALLENGES

The performance of ant colony optimization algorithms largely depends on the local search subroutines and population topologies have influences on the performance of the ant swarm optimization algorithms. The parameters of ant's behavior needed to fine adapt for the performance of clustering. This is because of the lack of understanding of the global behavior of a colony of simulated insect like agents. And also since these algorithms are designed problem specific it is difficult to tell if these methods would be suitable for certain problems or not. Furthermore it is very hard to analyze the computational complexity of these algorithms due to the nature of these algorithms Moreover we expect the overall performance of the ant colony algorithm to provide high reliability which will speak for itself over its drawbacks.

VII. CONCLUSION

In this paper we have presented a new ant-based clustering algorithm for knowledge discovery. The ACO introduces new ideas and modifications to improve the convergence providing a reliable method for optimization. The main features of this algorithm are that, it does not require pre-establishing the number of clusters or any other information about the feature of the clusters. Though we are not certain on the time complexity of the algorithms it certainly provides an angle for improvising in the phoneme updating algorithms.

VIII. REFERENCES

[1] Xiao Zhang, Aichen Li,You Zhang and Yongpeng Xiao, "Validity of Cluster Technique for Genome Expression Data",in proceedings of 24th Chinese Control and Decision Conference, pp 3737 – 3741,2012.

[2] Guiquan Liu, Xiufang Jiang and Lingyun Wen," A Clustering System for Gene Expression Data Based upon Genetic Programming and the HS-Model", in the proceedings of Third International Joint Conference on Computational Science and Optimization, IEEE computer society, pp 238-241, 2010.

[3] Wai-Ho Au, Keith C.C. Chan, Andrew K.C. Wong, and Yang Wang," Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", Transactions On Computational Biology And Bioinformatics, Vol. 2, No. 2,pp 83 – 101, 2005.

[4] Marco A. Me¤ndez, Christian Hodar, Chris Vulpe, Mauricio Gonzalez and Veronica Cambiazo, "Discriminant analysis to evaluate clustering of gene expression data", FEBS Letters, European Biochemical Societies, Elsevier Science, Vol 522, pp 24 – 28, 2002.

[5] Zhe Liu, Yu-qing Song, Cong-hua Xie, Feng Zhua and Xiang Bao, "Clustering gene expression data analysis using an improved EMalgorithm based on multivariate elliptical contoured mixture models", Optik, Elsevier,Vol 125, pp 6388-6394, 2014.

[6] Harun Pirim, BurakEks-ioglu, AndyD.Perkins and Cetin Yuceer, "Clustering of high throughput gene expression data", Computers & Operations Research, Elsevier,Vol 39, pp 3046-3061,2012.

[7] G. Kerr, H.J. Ruskin, M. Crane and P. Doolan," Techniques for clustering gene expression data", Computers in Biology and Medicine, Elsevier, Vol 38, pp 283 – 293, 2008.

[8] Joon Jin Song, Ho-Jin Lee, Jeffrey S. Morris and Sanghoon Kang," Clustering of time-course gene expression data usingfunctional data analysis", Computational Biology and Chemistry, Elsevier, Vol 31, pp 265–274, 2007.

[9] Miin-Shen Yang n, Chien-YoLai and Chih-YingLin, "A robust EM clustering algorithm for Gaussian mixture models", Pattern Recognition, Elsevier, Vol 45, pp 3950–3961, 2012.
[10] Urszula Boryczka,"Finding groups in data: Cluster analysis with ants", Applied Soft Computing, Elsevier, Vol 9, pp 61–70,2009.
[11] Kaiyang Liao, Guizhong Liu, Li Xiao, and Chaoteng Liu "A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval", Knowledge-Based Systems, Elsevier,Vol 49, pp 123-133,2013.
[12] Petri Toronen, Mikko Kolehmainen, Garry Wong, and Eero Castren," Analysis of gene expression data using self-organizing maps", FEBS Letters, Federation of European Biochemical Societies,Vol 451,pp 142-146,1999.
[13] Janne Nikkila, Petri Toronen, Samuel Kaski, Jarkko Venna, Eero Castren, and Garry Wong," Analysis and visualization of gene expression data using Self-Organizing Maps", Neural Networks Elsevier,Vol 15, pp 953-966,2002.
[14] Md. Bahadur Badsha, Md. Nurul Haque Mollah, Nusrat Jahan, and Hiroyuki Kurata, "Robust complementary hierarchical clustering for gene expression data analysis by β-divergence", Journal of Bioscience and Bioengineering, Elsevier,Vol 116,No3, pp 397-407,2013.
[15] Zhan C.T.," Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters ", Computers, IEEE Transactions,Vol C-20, issue 1, pp 68-86, Jan 1971.
[16] Wu, Z. and Leahy, R., "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation ", Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol 15 , Issue 11 ,pp 1101-1113,Aug 2002.
[17] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo," Model-based clustering and data transformations for gene expression data", *Bioinformatics,* Oxford Journals, Vol 17, issue 10,pp 977-987, 2001.
[18] Adrian E Raftery and Nema Dean, "Variable Selection for Model-Based Clustering", Journal of the American Statistical Association, Vol 101, Issue 473, pp 168-178,2006
[19] Shi Zhong and Joydeep Ghosh," A unified framework for model-based clustering", The Journal of Machine Learning Research, ACM, Vol 4, pp 1001-1037,2003.
[20] Daxin Jiang, Jian Pei, and Aidong Zhang," DHC: a density-based hierarchical clustering method for time series gene expression data", In Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering, pp 393–400, March 2003.
[21] Kriegel, H.-P. and Pfeifle, M.," Hierarchical density-based clustering of uncertain data", In Proceedings of fifth IEEE conference on Data Mining, Nov 2005.