Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.212 – 215

RESEARCH ARTICLE

Unsupervised Gene Data Using Clustering Method

Mr. S.B.Lanjewar, Mr. Pratik R.Tadas, Mr. Arvind B.Kadam, Mr. Gaurao B.Mankar, Mr. Gajendra Pise

Lec. of CSE dept.,	CSE,	CSE,	CSE,	CSE,
DBACER,.	DBACER.	DBACER.	DBACER.	DBACER.

ABSTRACT:

Microarrays are made it possible to simultaneously monitor the profiles of thousands of genes under various experimental conditions. Identification of co-expressed genes and coherent patterns is the central goal in microarray or gene expression data analysis and is an important task of research in bioinformatics. Feature selection is a procedure to select features which gives more information. The problem is that all features are not important. In this some of the features may be repeated or redundant, and others may be unrelated and noisy. In this features selection work the unsupervised Gene selection method and Center Initialization (Enhanced) Algorithm (ECIA) with K-Means algorithms have been applied for clustering of Gene Data. The proposed clustering algorithm reduced the drawbacks in terms of specify the optimal number of clusters and initialization of excellent cluster centroids. Data of gene expression show that could recognize dense or compact clusters with performs well in terms of the Silhouette Coefficients cluster measure.

INTRODUCTION:

Data of gene expression can be recognized in two way:

Unsupervised analysis: In this supervised analysis, information about the structure or arrangement of the object is suppose known or at least known.

Supervised analysis: In this type of analysis, preceding knowledge is unknown.

© 2015, IJCSMC All Rights Reserved

Clustering of gene data expression can be split into two types:

1. Gene-based clustering : In gene based clustering, genes are treat as objects and samples are treated as attributes for clustering. The goal of gene-based clustering is to recognize differentially expressed genes and sets of genes with similar expression pattern or profiles, and to generate a list of expression measurements.

2. Sample-based clustering. In Sample based clustering, samples are treated as objects and genes are features for clustering.

2. Sample based clustering: This clustering can be used to reveal the phenotype arrangement or substructure of samples.

Applying the usual clustering methods to cluster samples using all the genes as features may degrade the quality and reliability of clustering results. Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes. DNA microarrays are emerged as the leading technology to measuring gene expression levels first and foremost, because of their high throughput results.

The most popular clustering algorithms in microarray gene expression analysis are Hierarchical clustering, K-Means clustering [3]. Of these K-Means clustering is very simple and fast efficient. Numerous methods have been proposed to solve clustering problem. The most popular clustering methods are K-Means clustering algorithm which is developed by Mac Queen [5]. The K-Means algorithm is very effective in producing clusters for many applications. But the time and computational complexity of the original K-Means algorithm is effective and very high, specially for large data sets. The K-Means clustering algorithm is a partitioning clustering method.

LITERATURE REVIEW:

Discriminant analysis is now widely used in bioinformatics, such as differentiating the cancer tissues from normal tissues or among one cancer subtype vs. another one [4]. A critical issue in discriminant analysis is feature selection: instead of using all available variables (features or samples/attributes) in this data, one by one selectively choose a subset of features to be used in the discriminant system. There are number of advantages of feature selection: (1) dimension reduction to reduce the computational cost; (2) reduction of noise to improve the classification accuracy; (3) more interpretable features or characteristics that can help identify and monitor the target diseases and function types. These advantages are typify in DNA microarray gene expression profiles. Of the tens thousands of genes in research , only a smaller number of them show strong bond or correlation with the targeted phenotypes [4, 5].

The most popular clustering algorithms in microarray gene expression analysis are Hierarchical clustering, K-Means clustering [3], and SOM [6]. Of these K-Means clustering is simple and fast efficient. Numerous methods have been proposed to solve clustering problem. The most popular clustering methods are K-Means clustering algorithm which is developed by Mac Queen [8]. The K-Means algorithm is effective in producing clusters for many practical applications. But the computational complexity of the original K-Means algorithm is very high, especially for large data sets.

© 2015, IJCSMC All Rights Reserved

ARCHITECTURE:

In this work of unsupervised gene expression of data we have download the dataset. The dataset is in the matrix form. In matrix the row represent the genes and the column represents the condition or attributes.

The purpose of clustering gene expression data is to reveal the natural structure inherent data and extracting useful information from noisy data. The two class cancer subtype classification problem, 50 informative genes are usually sufficient. There are studies suggesting that only a few genes are sufficient. Thus, computation is reduced while prediction accuracy is increased via effective feature selection.

When a small number of genes are selected, their biological relationship with the target diseases is more easily identified.

In unsupervised gene based clustering we have done the three modules. These are as follows:

1) Preprocessing Module.
2) Unsupervised Quick Reduct Algorithm.
3) Clustering Module.

1) Preprocessing Module:

The preprocessing module consists of three sub-modules . They are as:

1. **Missing Values:**

   Sometimes the downloaded dataset contains the missing values. This missing values affecting the clustering. So we have filled this missing values by row averaging.

2. **Min –Max Normalization:**

   In min-max normalization we scaling the values between 0.0 and 1.0(0.0~1.0). The advantage of min-max normalization is we do not compute large values of gene in fraction.

3. **Discritazation:**

   In discritization part we converted the values of min-max normalized into the 1,0,-1 means(1,-1,0).

2) Unsupervised Quick Reduct Algorithm :

In this project after preprocessing part we have applied the USQR algorithm to quickly reduct the genes in dataset by using the it measures the values of one to another gene by Euclidian-distance algorithm.

3) Clustering :

Clustering is the process of grouping the similar objects into one cluster and dissimilar objects into another cluster.

© 2015, IJCSMC All Rights Reserved

The clustering module consists two algorithms. They are as:

1. **K-Means clustering algorithm:**

The k-means The main objective in cluster analysis is to group objects that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters. One of the most popular clustering methods is K-Means clustering algorithm [3, 6, 7]. The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids [8]. The Euclidean distance between two multi-dimensional data points .

2. **Euclidian-Distance algorithm.**

WORK FLOW:

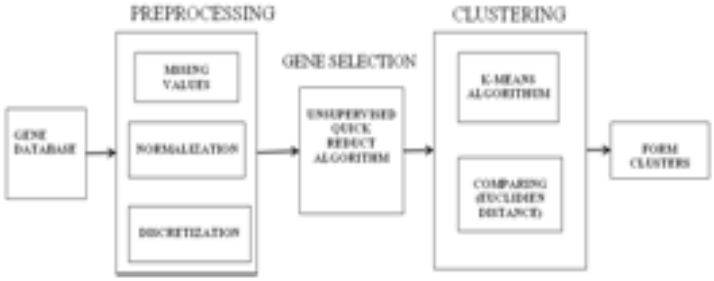

Fig: Working Diagram of Unsupervised Gene Expression

References:

[1] M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced K-Means clustering algorithm", journal of Zhejiang University, 10 (7): 1626 - 1633, 2006.

[2] Bashar Al-Shboul and Sung-Hyon Myaeng, "Initializing K-Means using Genetic Algorithms", World Academy of Science, Engineering and Technology 54, 2009.

[3] Chen Zhang and Shixiong Xia, " K-Means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[4] Chris Ding and Hanchuna Peng, "Mininmum Redundancy Feature Selection from Microarray Gene Expression Data", proceedings of the International Bioinformatic Conference, Date on 11-14, August - 2003.

[5] Doulaye Dembele and Philippe Kastner, "Fuzzy C means method for clustering microarray data", Bioinformatics, vol.19, no.8, pp.973- 980, 2003.

[6] Dongxiao Zhu, Alfred O Hero, Hong Cheng, Ritu Khanna and Anand Swaroop, "Network constrained clustering for gene microarray Data", doi:10.1093 /bioinformatics / bti 655, Vol. 21 no. 21, pp. 4014 – 4020, 2005.

[7] Kohei Arai and Ali Ridho Barakbah, " Hierarchical K-Means: an algorithm for centroids initialization for K-Means", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 25-31, 2007.

[8] K.R De and A. Bhattacharya , "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles," bioinformatics, Vol. 24, pp. 1359-1366, 2008.

© 2015, IJCSMC All Rights Reserved