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. 3, Issue. 5, May 2014, pg.359 – 364

RESEARCH ARTICLE

Clustering Techniques Analysis for Microarray Data

Shweta Srivastava[1], Nikita Joshi[2]

[1]CSE Department, ABES Engineering College, Ghaziabad, India

[2]CSE Department, ABES Engineering College, Ghaziabad, India

[1] shweta.shrivastava@abes.ac.in; [2] nikita.joshi@abes.ac.in

Abstract: Microarray data is gene expression data which consists of the protein level of various genes for some samples. It is a high dimensional data. High dimensionality is a curse for the analysis of gene expression data. Thus gene selection process is used in which most informative genes are selected from the pool of gene expression data set. All the genes are not relevant in each case. First we need to select those genes which are relevant as well as there should be least redundancy among them. For this purpose various approaches can be used such as: Filter methods, wrapper methods, embedded approach and clustering. In this paper embedded approach for gene selection and clustering method will be used for performing the sample clustering to refine the classification and will be compared with each other on the basis of various parameters.

Keywords- Clustering; Microarray Data; Gene Selection; Data Mining; Statistical Analysis

I. INTRODUCTION

Microarray technology and statistical analysis techniques have made it possible to analyse thousands of genes at one go. Clustering analysis is one of the statistical techniques that play an important role for elucidating the hidden patterns in gene expression data. How clustering can be useful for the gene expression data analysis? Clustering technique can be applied on microarray data for sample clustering, gene clustering or subspace clustering. Clustering is an unsupervised learning technique but in case of microarray data it can be useful for supervised data also. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms [1]. A detailed comparison of feature selection method is given in the previous work [9]. According to [9], the embedded approach is least prone to over fitting, thus we will be using this method for gene selection. This paper is divided in various sections as: Section 2 is about the background study done in which the study about gene expression data, clustering techniques for gene selection and how the clusters will be evaluated is discussed. Section 3 illustrates the proposed model for the given problem. Section 4 will discuss about the experimental results and Section 5 gives conclusion of the work and possible future work.

II. BACKGROUND STUDY

2.1 Gene Expression data:

There are two major types of microarray experiments: cDNA microarray and oligonucleotide arrays [3]. Both the experiments consist of basic three steps: first is chip manufacturing, second are target preparation, labelling and hybridization and third is the scanning process. Gene expression data is expressed in form of expression matrix having real values showing the protein level of a particular gene. Gene expression data contains thousands of genes but less number of samples. There are

various problems with microarray data such as: (a) Microarray data is high dimensional data characterized by thousands of genes for small sample size, which grounds significant problems such as irrelevant and noise genes, complexity in constructing classifiers, and multiple gene-expression values are missing due to inappropriate scanning. (b) Another drawback is mislabeled sample data or doubtful sample results by experts. (c) Biological relevancy result is another integral criterion that should be taken into account in analyzing microarray data rather than only focusing on accuracy of cancer classification. [6]

2.2 Clustering techniques:

In gene expression data, it is worth to cluster both genes and samples. There are three types of clustering that can be applied on microarray data: gene based clustering, sample based clustering and subspace clustering where genes and samples are treated in same manner. In case of gene clustering, the clustering is used to reduce the search dimension of the dataset. In case of sample based clustering, the clustering is used to group the samples of same kind whereas in subspace based clustering both the tasks are performed. Gene based clustering can be applied on the supervised dataset where the samples are already classified. The distinctive characteristic of gene expression data allows clustering both gene and samples. The clustering analysis of sampled data is to find new biological classes or to refine the existing ones [2].
There are different types of clustering: [7]
1. Hierarchical Clustering: (a) Agglomerative hierarchical clustering –In this each object initially represents a cluster of its own. Then clusters are recursively merged until the desired cluster formation is obtained.
(b) Divisive hierarchical clustering - All objects initially belong to one cluster. Then the cluster is divided into sub-clusters which are successively divided into sub clusters. This process continues until the desired cluster structure is obtained.
Some commonly used metrics for hierarchical clustering are: Euclidean distance, Squared Euclidean distance, Manhattan distance, Maximum distance, Mahalanobis distance and cosine similarity.
2. Partitioning Algorithms: They are iterative relocation algorithm. They are non hierarchical or flat methods. This method divides the data objects into non overlapping clusters such that each data object is in exactly one subset. There are several methods which are used to implement partitioning clustering such as: (a) K-medoids, (b) K-means, (c) Probabilistic.
3. Density based clustering: The clusters in this are dense regions of objects in space that are separated by low density regions where cluster density is defined as each point must have a minimum number of points in its neighborhood.
(i) Based on density based connectivity e.g. DBSCAN
(ii) Based on density distribution functions e.g. DENCLUE
4. Constraint based clustering: Constraints are strong background information that should be satisfied. Constraints also reduce the search space and all the data in dataset has common property. e.g. in gene expression data set we have a constraint of low and high expressed genes.
5. Evolutionary Clustering: It is used to process time stamped data to produce a series of clustering. The similarity among existing data points varies along with time. Present clusters mainly depend on the current data features. Data is likely to change not too rapidly. Evolutionary clustering is useful for the following reasons: (i) consistency, (ii) noise removal (iii) smoothing (iv) cluster correspondence.[8] Mostly used for online document clustering.
6. Graph Partitioning based Algorithms: It depends on finding the minimum cut or minimum cliques in the proximity graph G. [3] Many other graph partitioning algorithms depends on eigen vectors and eigen values also. It consists of three steps: (i) preprocessing i.e. to covert data into graph and finding similarity between the nodes. (ii) partitioning of the graph. (iii) performing clustering until required number of clusters are not obtained.
Each clustering algorithms belongs to one of the clustering types listed above. So that, Partitioning method is *exclusive clustering*, Fuzzy C-means is an *overlapping clustering* algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a *probabilistic clustering* algorithm.

2.3 Clustering Techniques for Sample clustering:

Co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicating co-regulation [3]. Unsupervised sample-based clustering is not good as supervised clustering since there no training set of samples that can be used as a reference to decide whether clustering is done correctly or not. Thus the class label is also taken into consideration in this paper.

2.4 Evaluation of clusters:

Evaluation of cluster will be done on the basis of predefined classes for samples and class obtained for the corresponding sample after performing the clustering.

III. PROPOSED MODEL

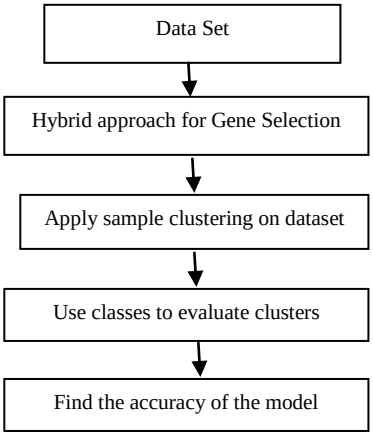

Fig.1. Proposed Approach

Step 1: Data Collection: Dataset is referred from reference 4. It consists of values of 2880 gene attributes along with a class attribute. There are total 39 samples in this dataset.
Step 2: The hybrid approach for gene selection will be same as in [4].
Step 3: Apply various clustering techniques on data set and find the clustered model.
Step 4: Use already allocated class labels for evaluating the clusters.
Step 5: Find the accuracy of the model on the basis of predefined classes.

IV. EXPERIMENTAL RESULTS

(1) Expectation Maximization (EM):

Clustered Instances:
0 14 (36%)
1 4 (10%)
2 21 (54%)
Class attribute: Class
Classes to Clusters:
0 1 2 <-- assigned to cluster
5 2 17 | Relapse
9 2 4 | Non-relapse
Cluster 0 <-- Non-relapse
Cluster 1 <-- No class
Cluster 2 <-- Relapse
Incorrectly clustered instances: 13.0 33.3333 %

(2) Cobweb:

Clustered Instances
0 2 (5%)
1 36 (92%)
2 1 (3%)
Class attribute: Class
Classes to Clusters:
0 1 2 <-- assigned to cluster
1 22 1 | Relapse
1 14 0 | Non-relapse
Cluster 0 <-- Non-relapse
Cluster 1 <-- Relapse
Cluster 2 <-- No class
Incorrectly clustered instances: 16.0 41.0256 %

(3) DBSCAN:

Clustered Instances
0 37 (100%)
Unclustered instances : 2
Class attribute: Class
Classes to Clusters:
0 <-- assigned to cluster
23 | Relapse
14 | Non-relapse
Cluster 0 <-- Relapse
Incorrectly clustered instances: 14.0 35.8974 %

(4) Hierarchical Clustering:

Clustered Instances
0 35 (90%)
1 1 (3%)
2 3 (8%)
Class attribute: Class
Classes to Clusters:
0 1 2 <-- assigned to cluster
22 1 1 | Relapse
13 0 2 | Non-relapse
Cluster 0 <-- Relapse
Cluster 1 <-- No class
Cluster 2 <-- Non-relapse
Incorrectly clustered instances: 15.0 38.4615 %

(5) Density Based clustering using Farthest First method:

Clustered Instances
0 38 (97%)
1 1 (3%)
Class attribute: Class
Classes to Clusters:
0 1 <-- assigned to cluster
23 1 | Relapse
15 0 | Non-relapse
Cluster 0 <-- Relapse
Cluster 1 <-- No class
Incorrectly clustered instances: 16.0 41.0256 %

(6) K-means:

Clustered Instances
0 14 (36%)
1 21 (54%)
2 4 (10%)
Class attribute: Class
Classes to Clusters:
0 1 2 <-- assigned to cluster
4 18 2 | Relapse
10 3 2 | Non-relapse
Cluster 0 <-- Non-relapse
Cluster 1 <-- Relapse
Cluster 2 <-- No class
Incorrectly clustered instances: 11.0 28.2051 %

Result Analysis:

Number of instances in cluster Relapse in actual dataset: 24
Number of instances in cluster Non relapse in actual dataset: 15
The clusters are evaluated using actual classes to assigned cluster comparisons.

TABLE 1
PERFORMANCE PARAMETERS OF CLUSTERING ALGORITHMS

Clustering Technique	Cluster Non Relapse found	Cluster Relapse found	Any new cluster found	Number of instances in Non relapse cluster	Number of instances in Relapse cluster	Number of instances in new cluster	Incorrectly classified instances (%)
Expectation Maximization (EM)	Yes	Yes	Yes	14	21	4	13(33.33%)
Cobweb	Yes	Yes	Yes	2	36	1	16(41.03%)
DBSCAN	No	Yes	No	NA	37	NA	14(35.9%)
Hierarchical Clustering	Yes	Yes	Yes	3	35	1	15(38.46%)
Density Based clustering using Farthest First method	No	Yes	Yes	NA	38	1	16(41.03%)
K-Means	Yes	Yes	Yes	14	21	4	11(28.21%)

The above parameters show that the K-means performs the best clustering in terms of correct clustering of instances. DBSCAN and Density Based clustering using Farthest First method doesn't create the cluster for non relapsed instances which exhibits not to be perfectly fitted models for this dataset. They create only one cluster for the whole dataset. Though they are working well for relapsed instances. The number of clustered instances in each cluster of EM and K-means are same but there are some instances which are relapse in actual but are put in non relapse cluster by clustering and vice versa. K-means performs best for non relapse instances. Hierarchical clustering and Cobweb are not performing well for non relapsed instances. They are more sensitive to relapsed instances.

V. CONCLUSION AND FUTURE WORK

In this paper, various clustering techniques are performed on a microarray dataset. The clustering techniques are compared on the basis of their actual classes in given dataset and allocated cluster after clustering. It is seen that density based clustering techniques are well suited for relapsed class instances but not for non relapsed class instances. K-means performs best for the non relapsed class instances and the overall performance of k-means is also best in comparison to other methods used. Some other technique for feature selection can improve the result of sample clustering.

REFERENCES

[1]. Qinbao Song, Jingjie Ni, Guangtao Wang; "A Fast Clustering Based Feature subset Selection Algorithm for High Dimensional data "; IEEE transactions on Knowledge and Data Engineering; Volume 25, Issue 1; 2013.

[2]. Wai-Ho Au, Keith C. C. Chan, Andrew K.C. Wong, Yang Wong; "Attribute Clustering for Grouping, Selection and Classification of Gene Expression Data "; IEEE transactions on Computational Biology and Bioinformatics; Volume 2, Issue 2; 2005.

[3]. Daxin Jiang, Chun Tang, Aidong Zhang; "Cluster analysis for gene expression data: a survey"; IEEE transactions on Knowledge and Data Engineering; Volume 16, Issue 11; 2004.

[4]. Shweta Srivastava, Manisha Rathi, Prof. J. P. Gupta: "Predictive Analysis of Lung Cancer Recurrence", First International Conference on Advances in Computing and Communications; 2011

[5]. E.N. Sathishkumar, K. Thangavel, T. Chandrasekhar; "A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction"; International Journal of Scientific & Engineering Research, Volume 4, Issue 5, page no 1540-1545, 2013.

[6]. Farzana Kabir Ahmad, Safaai Deris & Nor Hayati Othman; "Toward Integrated Clinical and Gene- Expression Profiles For Breast Cancer Prognosis: A Review Paper"; International Journal of Biometrics and Bioinformatics , (IJBB), Volume (3) : Issue (4).
[7]. Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443- 491.
[8]. Deepayan Chakrabarti, Ravi Kumar, Andrew Tomkins; "Evolutionary Clustering"; KDD, 2006.
[9]. Shweta Srivastava, Nikita Joshi, Madhvi Gaur, "A Review Paper on Feature Selection Methodologies and Their Applications", IJERD, Vol. 7, Issue 6, 2013.

