Comparison Of Different Filtering Approaches On Gene Expression Data For Clustering

Nisha Singh
Dept. of Information Technology
GGSIP University
Delhi, India

Khushboo Guliani
Dept. of Information Technology
GGSIP University
Delhi, India

Prashant Prabhat
Dept. of Information Technology
GGSIP University
Delhi, India

ABSTRACT - **Clustering** is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other groups. A **microarray** is a collection of microscopic DNA spots attached to a solid surface. Microarrays are used to measure the expression levels of large numbers of genes simultaneously. A **gene cluster** is a set of two or more genes that serve to encode for the same or similar products. Because common ancestors tend to possess the same varieties of gene clusters, they help to trace back recent evolutionary history. In this work, comparative analysis of the proposed work is made with different filtering approaches like **Entropy filtering, Genevarfilter** and **Genelowvalfilter**.

Keywords: Gene expression data, K-means, Entropy filtering, genevarfilter, **genelowvalfilter**

1. INTRODUCTION

1.1 Clustering

Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support [8]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [7]. Several data mining techniques are pattern recognition, clustering, association, classification and clustering [9]. There might be a thought in our mind as to "Why clustering?" A few good reasons are simplifications, pattern detection, useful in data concept construction and unsupervised learning process. We have used ***k*-means clustering which** is a method of cluster analysis aiming to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. It tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

1.2 Gene Expression Data

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's biochemical or physiological properties.

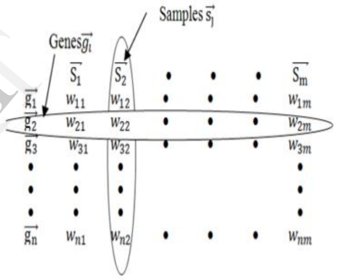


Figure 1: Gene Expression Matrix

Gene expression represents the activation level of each gene within an organism at a particular point of time. The expression value provides activity of a gene under certain biochemical conditions (also represents Samples).

2. Problem Statement

The problem in particular is a comparative study of filtering techniques (entropy filter, genevarfilter and genelowvalfilter) with an integration of Kmeans, on gene expression dataset Leukemia containing 22 attributes and 2337 instances.

3. Proposed Work

Clustering is the process of organizing objects into groups whose members are similar in some way. A *cluster* is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. In this paper we have applied K-means on a data set obtained by applying three different filtering techniques. Fig 2 represents the process flow of the gene clustering approach.

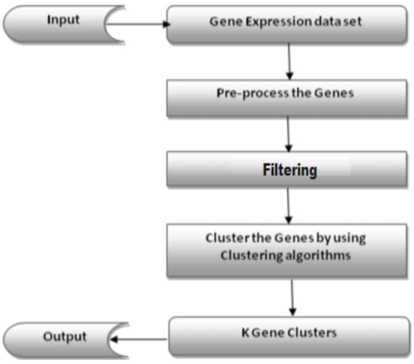


Figure 2. Methodology

4. Literature Survey

The pre processing of gene expression data obtained from several platforms routinely includes the aggregation of multiple raw signal intensities to one expression value. Cluster analysis technique is used to identify sets of genes that are co-ordinately regulated. Many clustering techniques have been applied for gene expression data [1].

Data clustering analysis has been extensively applied to extract information from gene expression profiles obtained with DNA microarrays [2].

Parvesh kumar, Siri Krishan Wasan applied both K-means and Rough K-means algorithms for Cancer data sets [3].

Preliminary Gene Selection

One of the interesting features of microarray experiments is the fact that they group information on a large number of genes. These issues will affect biologist in many ways and we face lot of problems while go for convergence. So, the dimensionality reduction of gene expression data sets should be considered. One of the characteristics of gene expression data is that it is meaningful to reduce dimension in both genes and samples, but in our thesis work we perform only the gene based clustering. The gene selection has three different approaches such as Filtering Approach, Wrapper Approach, and Embedded Approach. We have followed here, the Filtering Approach.

5. Filtering

5.1 geneentropyfilter

The effectiveness of the genes is calculated by using entropy filter method. Entropy measures the uncertainty of a random variable. For the measurement of interdependency of two random genes X and Y we used a direct function:
[Mask, FData, FNames] = geneentropyfilter(Data,Names)

5.2 genevarfilter

Gene profiling experiments have genes that exhibit little variation in the profile and are generally not of interest in the experiment. These genes are commonly removed from the data. For the measurement of interdependency of two random genes X and Y we used a direct function:
[Mask, FData, FNames] = genevarfilter(Data,Names)

5.3 genelowvalfilter

Gene expression profile experiments have data where the absolute values are very low. The quality of this type of data is often bad due to large quantization errors or simply poor spot hybridization. For the measurement of interdependency of two random genes X and Y we used a direct function:
[Mask, FData, FNames] = genelowvalfilter(Data, Names)

6. Clustering Algorithm

6.1 K Means Clustering

K-means clustering [4] is a partitioning method. The function kmeans partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. It operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters.

K-means treats each observation in your data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. You can choose from different distance measures, depending on the kind of data you are clustering.

Algorithm

Input: Set of sample patterns of genes{X1,X2,…,Xm}, Xi Rn

Step 1: Choose K initial cluster centers z1,z2,…,zk, randomly from the m patterns
{X1,X2,…Xm} where K<m.

Step 2: Assign pattern Xi to cluster center Zj, where I = 1, 2,…, m and j {1,2,…,K},
If and only if , p=1,2,…K and j p. These are resolved arbitrarily, and compute cluster center for each point xi as follows, Zi= (1/ni) , i = 1, 2, …, K. xj Zi.. where ni is the number of elements belonging to cluster Zi.

Step 3: Repeat this step 2 until there are no changes in centroid values.

7. EXPERIMENTAL RESULTS OF GENE FILTERING

To evaluate the performance of the Clustering algorithm with different filtering techniques described in this paper such as K-means algorithm with Entropy filtering, genevar filtering and genelowval filtering. These are implemented for cluster genes from various gene expression data sets and the experiment is performed using MATLAB.

Filtering Table is a table of data that has been filtered i.e. the number of total genes and samples in the original set are compared to the number of genes and samples left after the dataset has been filtered.

Error Chart is another output of our experiment which shows that among the following techniques applied, which technique has removed the maximum errors from the dataset and has given the best filtered data.

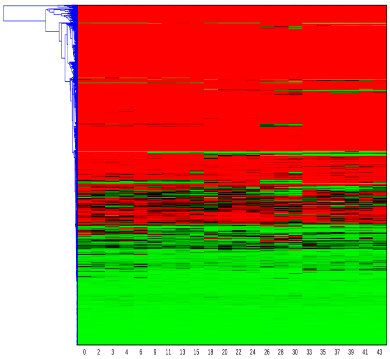
Clustergram is a type of genes expression patterns plotting method that automatically clusters genes and samples based on gene expression patterns.

* DATA SET

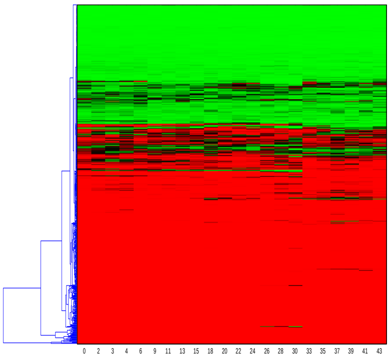
The data set is collected from the broad institute database [5]

DATASETS	GENES & SAMPLES
Leukaemia Cancer	**(2337,22)**

* CLUSTERGRAM



Clustergram of Original Data



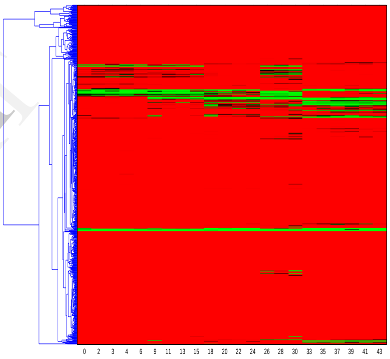
Clustergram after Geneentropyfilter

* FILTER TABLE

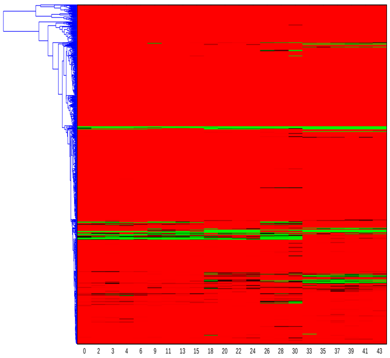
Filtering Technique	Original Datasets	Filtered Datasets
Gene Entropy Filtering	**(2337,22)**	**(1504,22)**
Gene Variance Filtering	**(2337,22)**	**(1519,22)**
Gene Low Value Filtering	**(2337,22)**	**(1805,22)**

* ERROR CHART

ORIGINAL DATASET	GENE ENTROPY FILTER	GENEVAL FILTER	GENLOWVAL FILTER
6.0500	1.7947	4.1350	5.5925
2.2919	0.6804	1.3939	1.5914
1.0603	0.3079	0.7795	0.8341
1.7501	0.2034	0.5249	0.5607
0.5293	0.1411	0.4241	0.4520



Clustergram after Genevarfilter



Clustergram after Genelowvalfilter

8. CONCLUSION

By observing the Filter Table and the Error Chart, we conclude that minimum error is in **Gene Entropy Filtering** which proves that it is the best among the three chosen techniques. After applying the different filtering techniques, the best among them is Gene Entropy Filtering since it removed the maximum waste and noisy data.

9. REFERENCES

[1] Daxin Jiang Chun Tang Aidong Zhang, “Cluster Analysis for Gene Expression Data: A Survey”, 2009.

[2] K. Y. Yeung and W.L. Ruzzo, “Principal Component Analysis for clustering gene expression data”, Oxford University press, Vol. 17 , pp. 763-774, 2001.

[3] Parvesh kumar, Siri Krishan Wasan, “Comparative Study of Kmeans, Pam and Rough K-means Algorithms using Cancer Datasets”, ISCCC-2009, vol.1, 2011.

[4] Abdul Nazeer. K. A, Sebastian.M.P, “Improving the accuracy and efficiency of the k-means clustering algorithm”, Proceedings of the World Conference on engineering 2009, Vol.1, pp.1-3, 2009.

[5]http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi/

[6] Dhanalakshmi.K, Hannah Inbarani, “FUZZY SOFT ROUGH K-MEANS CLUSTERING APPROACH FOR GENE EXPRESSION DATA”, 2012.

[7]J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.

[8] Desouza, K.C. (2001) ,Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands Institute for Healthcare Technology Management.

[9]Ritu Chauhan, Harleen Kaur, M.Afshar Alam, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.