Proceedings of the 2013 International Conference on
Pattern Recognition, Informatics and Mobile Engineering, February 21-22

A Novel Approach to Select Significant Genes of Leukemia Cancer Data Using K-Means Clustering

Prasath Palanisamy
Department of Biotechnology
Periyar University
Salem 636011, India
prasathbiotech@rediffmail.com

Perumal
Department of Biotechnology
Periyar University
Salem 636011, India
perumaldr@gmail.com

K.Thangavel
Department of Computer Science
Periyar University
Salem 636011, India
drktvelu@yahoo.co.in

R.Manavalan
Department of Computer Science
KSR Institute of Arts & Science
Thiruchengodu, India
manavalan_r@rediffmail.com

***Abstract*-DNA microarray technologies are leading to an explosion in available gene expression data which simultaneously monitor the expression pattern of thousands of genes. All the genes may not be biologically significant in diagnosing the disease. In this paper, a novel approach has been proposed to select significant genes of leukemia cancer using K-Means clustering algorithm. It is an unsupervised machine learning approach, which is being used to identify the unknown patterns from the huge amount of data. The proposed K-Means algorithm has been experimented to cluster the genes for K=5,10 and 15. The significant genes have been identified through the best accuracy obtained from the clusters generated. The accuracy of the clusters are determined again by using K-Means algorithm compared with ground truth values.**

Keywords: K-Means; Leukemia; Microarray, Clustering, Specificity, Sensitivity, Accuracy

I. INTRODUCTION

Microarray technology can either be used to investigate the functions of genes, or be used in the diagnosis of diseases. In the past years, many existing data analysis methods from other fields have been applied to gene expression data; also many novel methods are developed or under developing particularly for gene expression data analysis.

DNA micro arrays are also commonly known, as gene chips, DNA chip, or biochip. In which it is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA micro arrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Important knowledge can be extracted from these data by the use of data analysis techniques. It focuses on various techniques and methods proposed for biological data analysis. Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously.[1]

Clustering involves dividing a set of data points into non-overlapping groups, or clusters, of points, where points in a cluster are "more similar" to one another than to points in other clusters. The term "more similar," when applied to clustered points, usually means closer by some measure of proximity. When a dataset is clustered, every point is assigned to some cluster, and every cluster can be characterized by a single reference point, usually an average of the points in the cluster. Any particular division of all points in a dataset into clusters is called a partitioning.[2]

"Good," or representative, clustering: Consider a single cluster of points along with its centroid or mean. If the data points are tightly clustered around the centroid, the centroid will be representative of all the points in that cluster. The standard measure of the spread of a group of points about its mean is the variance, or the sum of the squares of the distance between each point and the mean. If the data points are close to the mean, the variance will be small. A generalization of the variance, in which the centroid is replaced by a reference point that may or may not be a centroid, is used in cluster analysis to indicate the overall quality of a partitioning; specifically, the error measure E is the sum of all the variances:

$$E = \sum_{i=1}^{K} \sum_{j=1}^{n_j} ||x_{ij} - z_i||^2$$

where xij is the jth point in the ith cluster, zi is the reference point of the ith cluster, and ni is the number of points in that cluster. The notation ||xij - zi|| stands for the distance between xij and zi. Hence, the error measure E indicates the overall spread of data points about their reference points. To achieve a representative clustering, E should be as small as possible.[2]

Among these, clustering methods are a large family of commonly used data analysis methods. Clustering methods are particularly useful in the analysis of gene expression data and organization of gene expression data and the results from clustering can be used as a starting point for further analysis.

978-1-4673-5845-3/13/$31.00©2013 IEEE

A natural basis for organizing gene expression data is to group together genes with similar patterns of expression i.e. co-expressed genes. Co-expressed genes may reveal much about co regulatory mechanisms.

An increasingly common objective in the analysis of genetic microarray data is to investigate the association between genomic profiles and disease class or outcome (for example, tumor or tissue type). A clinical goal of such efforts would be the ability to predict disease class based solely upon a sample's gene expressions. To accomplish this, we must first select a subset of genes from among all those considered, with the optimal subset being that which best predicts disease class using as few genes as possible.

Gene expression data are characterized by a very high dimensionality (genes),a relatively small number of samples (observations), irrelevant features, and it leads to a co-linearity and multivariate problem. In this paper, we propose a systematic approach to gene selection based on K-means clustering method. The proposed method was applied to microarray data from Leukemia patients; specifically, it was used to interpret the gene expression pattern.

The rest of the paper is organized as follows: Section 2 discusses the K-Means clustering algorithm, Section 3 the proposed approach, Section 4 provides the necessary data for the experimental analysis, Section 5 presents the computational results and discussion. Section 6 concludes this paper.

II. K-MEANS CLUSTERING

K-Means is a prototype-based, simple partition clustering technique which attempts to find a user specified K number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. The algorithm consist of two phases: the first phase is to define K centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to nearest centroid. The K-Means algorithm works as follows:

a) Select initial centroid of the K clusters. Repeat steps b through c until the cluster membership stabilized.

b) Generate a new partition by assigning each data to its closest cluster centroid.

c) Compute new cluster centroid for each cluster

The most widely used convergence criteria (1) for the K-Means algorithm is minimizing the SSE (Sum Squared Error).

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in c_j} ||x_i - \mu_j||^2$$

Where

$$\mu_j = \frac{1}{n_j} \sum_{x_i \in c_j} x_i$$

denotes the mean of cluster cj and nj denotes the no. of instances in cj. The K-Means algorithm always converges to a local minimum. The particular local minimum found depends on the starting cluster centroids. The K-Means algorithm updates cluster centroids till local minimum is found.[3]

Before the K-Means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of kmeans iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational complexity of the algorithm is O(nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations. The time complexity for the high dimensional data set is O(nmkl) where m is the number of dimensions.[3]

K-means algorithm is iteratively convergent, and, if the initial “center points” are selected well, that is to say, they are close to the true center points,then K-means will converge more rapidly, and the clustering result will be more accurate [4].

III. PROPOSED APPROACH

The K-Means clustering is applied to cluster the genes of the Leukemia dataset for different values of K, since there is no class label for genes. In this paper, the novelty is achieved by applying the K-Means algorithm to cluster the samples by omitting the class labels in each gene cluster to determine the accuracy by comparing with the ground truth values.

IV. EXPERIMENTAL ENVIRONMENT

The description of leukemia expression datasets are as follows: This has 7129 samples with 34 genes and consists of 2 classes, Acute lymphoblastic leukemia & Acute myeloid leukemia. They are kinds of cancer, and each of them has different characteristics.

Each patient is represented as one row. Column 1 is the patient number in the dataset, columns 2 to 34 denote the gene expression values corresponding to each patient, column 7130 indicates the type of cancer (ALL,AML) that each patient is classified. A simple measure of the genomic difference between two patients can be obtained by resorting to the Euclidean distance of two points.

In order to ease the algebraic manipulations of data, the dataset can also be represented as a real 2-D matrix S of size 7129×34; the entry sij of S measures the expression of the jth gene of the ith patient. Each patient is determined by a sequence of 34 real numbers, each measuring the relative expression of the corresponding gene.

V. COMPUTATIONAL RESULTS

The K value is arbitrarily fixed as 5, 10 and 15 and the clustering is performed and the results are provided in table 1,

table 2 and table 3 respectively. The best results are indicated in Bold letters.

TABLE I. EXPERIMENTAL RESULTS FOR K=5

Run(s)	K-Means Clustering		
	Sensitivity	**Specificity**	**Accuracy**
1	0.71	0.50	0.59
2	0.61	0.45	0.56
3	0.64	0.56	0.62
4	0.61	0.45	0.56
5	0.74	1.00	0.79
6	0.78	0.82	0.79
7	0.71	0.50	0.59
8	0.71	0.70	0.71
9	0.59	0.43	0.56
10	0.74	1.00	0.79

The graphical representation of the results shown in table 1 is provided in fig. 1. The X axis represents the number of runs and the Y axis represents the accuracy.

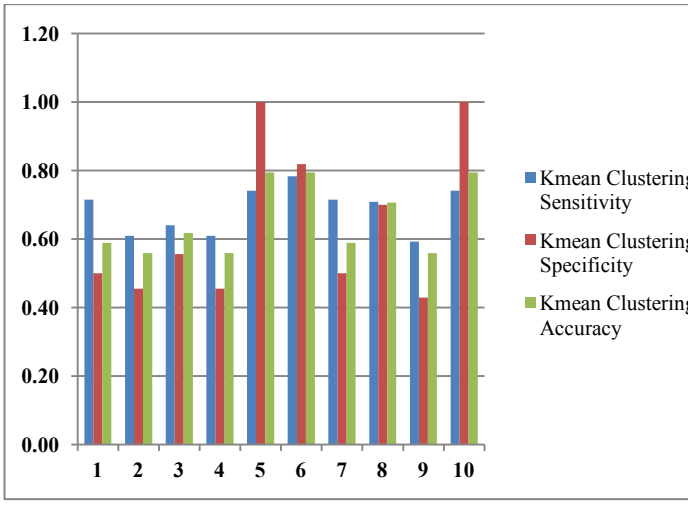

Figure 1. Accuracy Results for K=5

TABLE II. EXPERIMENTAL RESULTS FOR K=10

Run(s)	K-Means Clustering		
	Sensitivity	**Specificity**	**Accuracy**
1	1.0	1.0	1.0
2	0.8	0.8	0.8
3	0.8	0.9	0.8
4	0.8	0.9	0.8
5	0.8	0.9	0.8
6	0.8	0.8	0.8
7	0.6	0.7	0.6
8	1.0	0.8	0.9
9	0.7	1.0	0.8
10	0.8	0.5	0.5

The graphical representation of the results shown in table 2 is provided in fig. 2. The X axis represents the number of runs and the Y axis represents the accuracy.

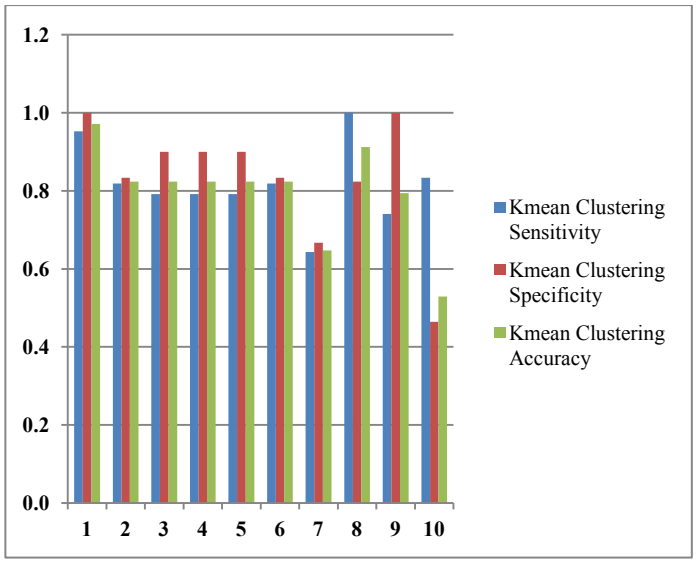

Figure 2. Accuracy Results for K=10

TABLE III. EXPERIMENTAL RESULTS FOR K=15

Run(s)	K-Means Clustering		
	Sensitivity	Specificity	Accuracy
1	0.79	0.90	0.82
2	0.77	0.75	0.76
3	0.73	0.88	0.76
4	0.71	1.00	0.76
5	0.70	0.86	0.74
6	0.74	1.00	0.79
7	0.95	0.93	0.94
8	0.69	1.00	0.74
9	0.79	0.90	0.82
10	1.00	0.82	0.91

The graphical representation of the results shown in table 3 is provided in fig. 3. The X axis represents the number of runs and the Y axis represents the accuracy.

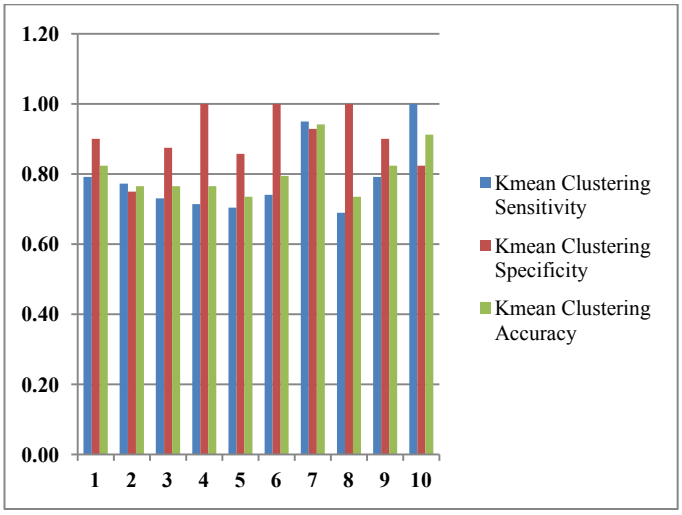

Figure 3. Accuracy Results for K=15

The accuracy of 79% is achieved for 873 genes and the same is achieved for 6633 genes when K=5.(Run 5 and Run 6).

The accuracy of 97% is achieved for 434 genes in Run 1 and the accuracy of 82% is achieved for 114 genes (Run2) and 1491 genes when K=10 in Run 3, Run 4 and Run 5.

The accuracy of 82% is achieved for 1642 genes in Run 1. The accuracy of 94% is achieved for 668 genes in Run 7. The accuracy of 82% is achieved for 1552 genes in Run 9. The accuracy of 91% is achieved for 256 genes in Run 10.

The best accuracy results obtained for K=10 and K=15 are tabulated in table 4.

TABLE IV. BEST RESULTS FOR K=5, K=10 AND K=15

Run	K Value	Sensitivity	Specificity	Accuracy	Number of Genes Selected
5	5	0.74	1.00	0.79	873
6	5	0.78	0.82	0.79	6633
1	10	0.95	1.00	0.97	434
2	10	0.82	0.83	0.82	114
3	10	0.79	0.90	0.82	1491
4	10	0.79	0.90	0.82	1491
5	10	0.79	0.90	0.82	1491
6	10	0.82	0.83	0.82	4524
8	10	1.00	0.82	0.91	114
1	15	0.79	0.90	0.82	1642
7	15	0.95	0.93	0.94	668
9	15	0.79	0.90	0.82	1552
10	15	1.00	0.82	0.91	256

The graphical representation of the best accuracy results are illustrated in the fig. 4.

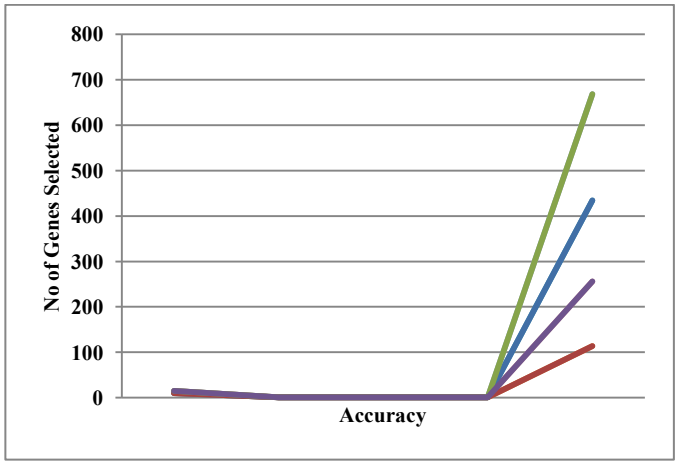

Figure 4. Best Accuracy Results

VI. CONCLUSION

In this paper, K-Means algorithm has been studied for leukemia gene expression dataset.

The K-Means algorithm has been implemented to find the clusters. The gene clusters obtained by using K-means algorithm are used for classification. The performance of the clusters was evaluated using accuracy by comparing against the ground truth values. Out of 7000 genes in the dataset it is enough to consider only 114 genes to predict the disease.

REFERENCES

[1] P.Rajeswari et al, A Survey of Human Cancer Classification using Micro Array Data, Int. J. Comp. Tech. Appl., Vol 2 (5), 1523-1533 ISSN:2229-6093

[2] Vance Faber, Clustering and the Continuous k- Means Algorithm, Number 22 1994 Los Alamos Science.

[3] An Improved Method Of Unsupervised Sample Clustering Based On Informative Genes For Microarray Cancer Data Sets. TAJUNISHA N .1*, SARAVANAN V.2 International Journal of Computational Biology ISSN: 2229–6700, E-ISSN: 2229–6719, Vol. 2, Issue 1, 2011, pp-24-31

[4] Clustering of Leukemia Patients via Gene Expression Data Analysis. 12-15-2006, Zhiyu Zhao University of New Orleans.