IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 4 Ver. V (Jul -Aug. 2014), PP 12-16
www.iosrjournals.org

Informative Gene Selection for Leukemia Cancer Using Weighted K-Means Clustering

Prasath Palanisamy[1] P. Perumal[2] , K. Thangavel[3] , R. Manavalan[4]
[1]*Quality Control Biologics Syngene International Ltd Biocon ParkBangalore -560 099 India*
[2]*Department of Biotechnology Periyar University Salem-636 011 India.*
[3]*Department of Computer Science Periyar University Salem- 636 011 India.*
[4]*Department of Computer Science KSR College of Arts & Science Thiruchengodu-637 215 Namakkal. India.*

Abstract: *Gene expression data analysis is playing a vital role in diagnosing the diseases and drug designing. Many researchers realized that most of the cancers could be diagnosed based on the gene expression data. This paper focusses on identifying the prominent genes, which are mainly causing the Leukemia cancer using computational methods. Clustering methods are used to identify the components of a data set without the prior knowledge of the data set. The Weighted K-Means clustering method is proposed in a novel manner to analyze the Leukemia Cancer data set. However, the resultant clusters are again clustered sample wise using the K-Means clustering approach to understand more meaningful biological inferences. The proposed method selects the most significant genes and produces high accuracy in cancer classification.*

Keywords: *Accuracy, Clustering, FC-Means, K-Means, Leukemia, MK-Means, Specificity, Sensitivity, Weighted K-Means.*

I. Introduction

Microarray technology is used to monitor the expressions of thousands of genes simultaneously. In microarray data analysis, there is a big challenging problem, the dimension of gene expressions is much larger than the sample size, which makes it be a hot and hard research topic. In the field of Biotechnology and Bioinformatics, a large amount of efforts have also been made to identify relevant or important genes that have influential effects on diseases including variety of cancers, which is a class of diseases for which a group of cells undergoes uncontrolled growth. It causes destruction of adjacent tissues and sometimes spreads to other locations in the body via lymph or blood. American Cancer Society stated that about 7.6 million people died from cancer in the world during 2007, and nearly all cancers are caused by abnormalities in the genetic material of the transformed cells. With the current deluge of data, computational methods have become indispensable to biological investigations.[1]

Clustering is a process of partitioning a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some define criteria. It means grouping the data into clusters according to their internal character, the element in each cluster should have as similar character as possible, the difference between clusters should be as big as possible. [2].Clustering methods can be divided into two basic types: hierarchical and partitional clustering. Within each type there exists a wealth of subtypes and different algorithms. Hierarchical clustering proceeds successively either by merging smaller clusters into larger ones (bottom-up), or by splitting larger clusters into smaller clusters (top-down). The hierarchical clustering methods differ in the rules used to decide which two small clusters are merged or which large cluster is split. The final result of the algorithm is a binary tree of clusters called a dendrogram, which shows how the clusters are related to each other. By cutting the dendrogram at a desired level, a clustering of objects in a dataset into disjoint groups is obtained.

On the other hand, partitional clustering – K-Means, for example – attempts to directly divide a dataset into a number of disjoint groups. All partitional clustering algorithms need as input the number of clusters and a cost (criterion) function to define the quality of a partition. The partitional clustering method aims at optimizing the cost function to minimize the dissimilarity of the objects within each cluster, while maximizing the dissimilarity of different clusters. In general, the partitional clustering algorithms are iterative and hill-climbing, and thus they are sensitive to the choice of the initial partition. Furthermore, since the associated cost functions are nonlinear and multimodal, usually these algorithms converge to a local minimum [3].

Most of the researchers have applied K-Means clustering to analyze the gene expression data.[4] In this method, all the samples will be treated equally at the time of clustering. But some genes of the samples might have been over expressed and more weightage will be given for such samples. Hence, in this paper the Weighted K-Means (WKM) algorithm is proposed and a novel procedure is adopted to analyze the gene expression profile of Leukemia dataset.

The rest of the paper is organized as follows: Section II discusses the Weighted K-Means clustering algorithm. Section III the proposed novel approach to gene selection. Section IV provides the necessary data for the experimental analysis. Section V presents the computational results and discussion. Section VI concludes this paper. Section VII presents the acknowledgement.

II. Weighted K-Means Clustering

In K-Means algorithm, every data point has equal importance in locating the centroid of the cluster as well as every members of the data points carry unit weight. This property does no longer hold in the case of density-based sample clustering, for which each data point represents varied density in the original data. Therefore, the clustering algorithm has to consider a weight associated with each data point in the computation of grouping similar data points. The weight function, w(xi) was introduced which represents density of the original data points.[5] The WKM algorithm helps to find out those genes, which are most important by calculating the weight of genes during each iteration. It attempts to decompose a set of genes into a set of disjoint clusters and clustering is performed on those genes to obtain significant clusters.

Weighted K-Means Clustering Algorithm
Input:
D = { d_1, d_2, d_3, . . ., d_n } // Set of n data points.
W = { w_1, w_2, w_3, . . ., w_n } // Set of weights associated with each data point.
K // Number of desired clusters.
Output:
K distinct clusters.
Steps:
1. Using the standard K-means algorithm, assign points to clusters and centers to appropriate positions.
Repeat
2. Compute the distance between each weighted data point d_i to all the centroids c_j.
3. Update cluster center mean i.e., calculate the mean value of the weighted objects for each cluster.
4. Set point energies
(1) f (i) = $\sum^{n}_{i=1}$(di-cj). ^2
5. Set cluster energies
(2) ce (i) = ce (i) + w (i) * f (i)
6. Adjust point energies by weight factor
If (w (i) < ce (i)) then
(3) f (i) = f (i) * ce (i) / (ce (j) – w(i))
End
7. Until no change

III. Proposed Approach

In this paper, the gene datasets have been clustered using WK-Means by setting *K*=10, *K*=15, and *K*=20. The clusters which are obtained using the above methods, are further clustered using K-means clustering by taking *K*=2. Hence, it is a novel approach to gene selection using clustering methods.

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 (ALL) and Acute Myeloid Leukemia (AML). The classes represent cancer, and each of them has different characteristics. Each patient is represented as one column. 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.

In order to ease the algebraic manipulations of data, the dataset can also be represented as a real 2-D matrix d of size 7129×34; the entry d_{ij} of d measures the expression of the j^{th} gene of the i^{th} 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 10, 15 and 20 and the Weighted K-Means clustering is performed and the results are provided in, the below TABLE 1, TABLE 2 and TABLE 3 respectively. The best results are indicated in Bold letters.

Table I: Experimental Results For K=10

Run	K	Weighted K-Mean		
		Sensitivity	Specificity	Accuracy
1	10	0.95	1	0.97
		1	0.78	0.88
		0.81	0.61	0.71
		1	0.64	0.76
		1	0.5	0.59
		0.25	0.23	0.24
		0.59	NaN	0.59
		1	0.47	0.53
		0.9	0.86	0.88
		0.83	1	0.88

The graphical representation of the results shown in "TABLE 1" is provided in "Fig. 1". The X axis represents the number of clusters and the Y axis represents the measures.

Figure 1 Results for K=10

Table Ii: Experimental Results For K=15

Run	K	Weighted K-Mean		
		Sensitivity	Specificity	Accuracy
1	15	0	0	0
		0.83	1	0.88
		0.79	0.55	0.65
		1	0.56	0.68
		0.59	NaN	0.59
		0	0	0
		1	0.93	0.97
		0.93	0.68	0.79
		1	0.64	0.76
		0.95	1	0.97
		0.93	0.68	0.79
		0.59	NaN	0.59
		1	0.54	0.65
		0.86	0.48	0.56
		0.42	0.32	0.35

The graphical representation of the results shown in "TABLE II" is provided in "Fig. 2". The X axis represents the number of clusters and the Y axis represents the measures.

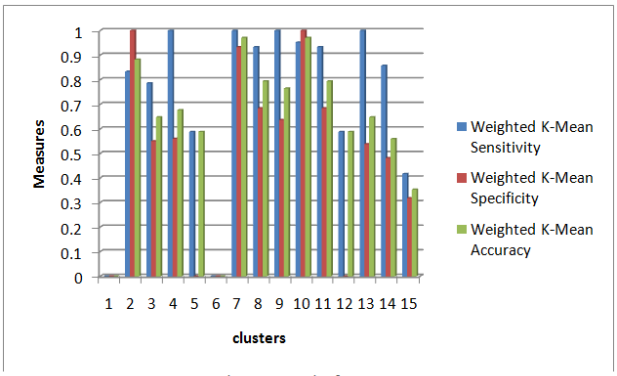

Figure 2 Results for K=15

Table Iii: Experimental Results For K=20

Run	K	Weighted K-Mean		
		Sensitivity	Specificity	Accuracy
1	20	0.95	1	0.97
		0.83	1	0.88
		1	0.52	0.62
		1	0.58	0.71
		0	0	0
		0.93	0.68	0.79
		0.23	0.19	0.21
		1	0.47	0.53
		0.52	0.22	0.44
		0	0	0
		1	0.93	0.97
		0.59	NaN	0.59
		0	0	0
		0.63	0.57	0.62
		1	0.64	0.76
		0.8	1	0.85
		0	0	0
		0.69	1	0.74
		0.61	0.67	0.62
		0.71	1	0.76

The graphical representation of the results shown in "TABLE III" is provided in "Fig" 3. The X axis represents the number of clusters and the Y axis represents the measures.

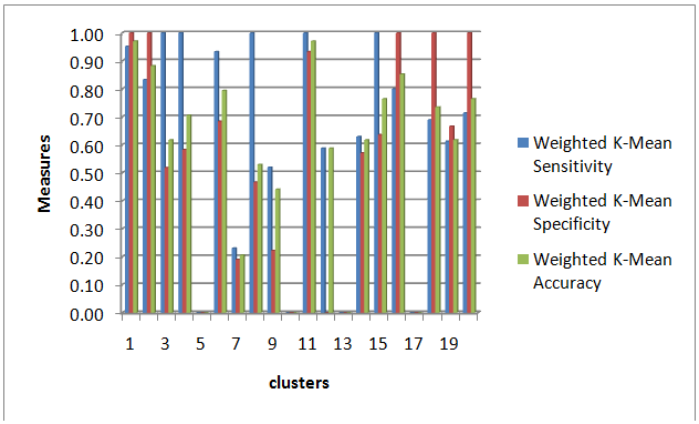

Figure 3 Results for K=20

K=10:

The accuracy of 97% is achieved for 2 genes which was obtained in the cluster 1 and the genes selected were 19 & 1222 respectively in Weighted K-Means Clustering.

K=15:

The accuracy of 97% is achieved for 6 genes which was obtained in the cluster 7 & 10 and the genes selected were 5710, 5711, 19, 1222, 5058 & 5997 respectively in Weighted K-Means Clustering.

K=20:

The accuracy of 97% is achieved for 4 genes which was obtained in the cluster 1 & 11 and the genes selected were 5710,5711,19 & 1222 respectively in Weighted K-Means Clustering.

The best accuracy results obtained for K=10, K=15 and K=20 Weighted K-Means Clusters along with the genes selected are tabulated in "TABLE IV".

TABLE IV- BEST ACCURACY RESULTS

Run(s)	K	Weighted K-Mean Clustering				
		Sensitivity	Specificity	Accuracy	Number of Genes Selected	Selected Gene(s)
1	**10**	0.95	1	0.97	2	19, 1222
	15	1	0.93	0.97	6	5710,5711,19,1222,5058,5997
	15	0.95	1	0.97	6	5710,5711,19,1222,5058,5997
	20	0.95	1	0.97	4	19, 1222, 5710, 5711
	20	1	0.93	0.97	4	19, 1222, 5710, 5711

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

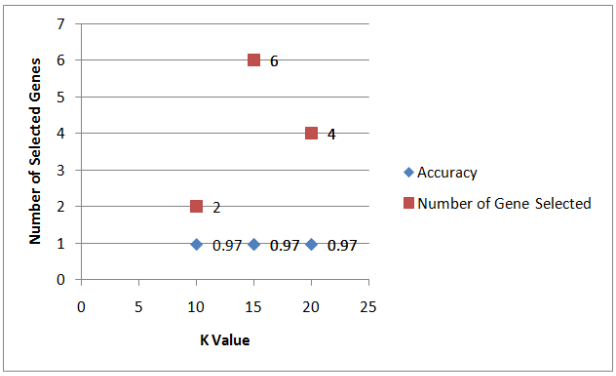

Figure 4 best accuracy results

VI. Conclusion

In this paper, Weighted K-Means algorithm has been studied and implemented for Leukemia gene expression dataset. The gene clusters obtained by using WKM algorithm are further classified using K-Means algorithm and accuracy was evaluated by comparing against ground truth values. The identified significant genes 19 and 1222 were presented for all the different values of K. Hence the Leukemia cancer could be diagnosed with the identified genes 19 and 1222 by the proposed WKM method instead of analysing the 7129 genes.

Acknowledgement

The third Author immensely acknowledges the UGC, New Delhi for partial financial assistance under UGC-SAP (DRS) Grant No: F3-50/2011.

References

[1] Liping Jing, Michael K. Ng, and Tieyong Zeng, Novel Hybrid Method for Gene Selection and Cancer Prediction World Academy of Science, Engineering and Technology 62 2010

[2] Anand M. Baswade, Kalpana D. Joshi, Prakash S. Nalwade, A Comparative Study Of K-Means And Weighted K-Means For Clustering, International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 10, December- 2012

[3] Fang-Xiang Wu, Genetic weighted k-means algorithm for clustering large-scale gene expression data, BMC Bioinformatics 2008, 9(Suppl 6):S12 doi:10.1186/1471-2105-9-S6-S12

[4] Prasath Palanisamy[#1], Dr. Perumal[#2], Dr.K.Thangavel[#3], R.Manavalan[#4], A novel approach to select significant genes of leukemia cancer data using K-Means clustering, International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 Page(s): 104 – 108,

[5] M Chitralegha[#1], Dr K Thangavel[*2], Protein Sequence Motif Patterns using Adaptive Fuzzy C-Means Granular Computing Model IEEE-2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), PRIME 2013