A Novel Approach to Gene Selection of Leukemia Dataset Using Different Clustering Methods

P. Prasath, K. Perumal, K. Thangavel and R. Manavalan

Abstract Gene datasets from microarray comprise large number of genes. Clustering is a widely used approach for grouping similar kind of genes. The main objective of this paper is to identify the optimal subset of genes from the leukemia dataset in order to classify the leukemia cancer. Different clustering approaches such as *K*-means (KM) clustering, fuzzy *C*-means (FCM) clustering, and modified *K*-means (MKM) clustering have been adopted in this research. The clusters obtained from these methods are further clustered using *K*-means sample-wise (by omitting class values), and the results are compared with ground truth value to evaluate the performance of the different clustering methods. The highly correlated genes are selected from the cluster that produces more accurate classification results. It is observed that the FCM (gene-wise clustering) with *K*-means (sample-wise clustering) produces better accuracy, and the resultant genes have been identified.

Keywords FCM · MKM · *K*-means · Leukemia · Microarray

P. Prasath (✉) · K. Perumal
Department of Biotechnology, Periyar University, Salem 636 011, India
e-mail: prasathbiotech@rediffmail.com

K. Perumal
e-mail: perumaldr@gmail.com

K. Thangavel
Department of Computer Science, Periyar University, Salem 636 011, India
e-mail: drktvelu@yahoo.co.in

R. Manavalan
Department of Computer Science, K. S. Rangasamy College of Arts and Science, Thiruchengode, India
e-mail: manavalan_r@rediffmail.com

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_7, © Springer India 2014

1 Introduction

Gene expression data can be obtained by high-throughput technologies such as microarray and oligonucleotide chips under various experimental conditions, at different developmental stages. This technique promises to allow for the detection of networks of correlated genes, which are characteristic of phenomena such as diseases. However, classification of samples according to phenotypes or other criteria is not necessarily precise; therefore, it is desirable to use unsupervised methods to classify samples according to gene expression similarity and to detect networks of correlated genes that discriminate those sample classes [1].

The gene expression data are usually organized in a matrix of *n* rows and *m* columns, which is known as a gene expression profile. Due to the large amount of gene expression data available on various cancerous samples, it is important to construct classifiers that have high predictive accuracy in classifying cancerous samples based on their gene expression profiles [2]. Microarrays contain precisely positioned DNA probes that are designed to specifically monitor the expression of genes in parallel. Data mining often utilizes mathematic techniques that are traditionally used to identify patterns in complex data. Here, the unsupervised benchmark *K*-means clustering method is adopted from [7] for comparative analysis.

The rest of the paper is organized as follows: Sect. 2 describes the FCM clustering method, Sect. 3 deals with MKM clustering, Sect. 4 proposes novel proposed approach to gene selection, Sect. 5 provides experimental environment, Sect. 6 includes experimental results, and Sect. 7 concludes this paper with direction for further research.

2 Fuzzy *C*-Means Clustering

In FCM clustering method, an object can simultaneously be a member of multiple clusters. The objective function, which is minimized iteratively, is a weighted within-group sum of distances. The weight is computed by multiplying the squared distances with membership values. After computing the membership values for all calibration objects, the cluster centers are described by prototypes, which are fuzzy weighted means. This fuzzy clustering method allows intermediate logical assignments whereby genes are placed into multiple groups by assigning a membership value for each group that is compared between 0 (not in group) and 1 (completely in group). The use of membership values has the advantage of allowing a gene or sample to belong to multiple clusters, which may better reflect the underlying biology [4].

3 Modified *K*-Means Clustering Algorithm

This algorithm calculates the cluster centers that are quite close to the desired cluster centers. It first divides the dataset into K subsets according to some rule associated with data space patterns and then chooses cluster centers for each subset [5].

4 Proposed Approach

In this paper, the gene datasets have been clustered using K-means, modified K-means, and fuzzy C-means by setting $K = 5$, $K = 10$, and $K = 15$. 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.

5 Experimental Environment

The description of leukemia cancer dataset is as follows [6]: This has 7,129 genes with 34 samples and consists of 2 classes: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). They are various types of cancer, and each of them has different characteristics. Each patient is represented as one row. First column is the patient number in the dataset, columns 2 to 34 denote the gene expression values corresponding to each patient, and 7,130th column 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 two-dimensional matrix S of size 7,129 × 34; the entry s_{ij} 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 [3].

6 Computational Results

The K value is arbitrarily fixed as 5, 10, and 15, FCM clustering is performed, and the results are provided in Tables 1, 2, and 3, respectively. The best results are indicated in bold letters. Similarly, the K value is arbitrarily fixed as 5, 10, and 15, MKM clustering is performed, and the results of the clusters are provided in Tables 4, 5, and 6, respectively. The best results are indicated in bold letters.

Table 1 Experimental results for $K = 5$

Run(s)	K	FCM clustering		
		Sensitivity	Specificity	Accuracy
1	5	0.94	0.76	0.85
2	5	0.76	0.69	0.74
3	5	0.76	0.69	0.74
4	5	0.90	0.92	0.91
5	5	0.94	0.76	0.85
6	5	0.59	0.42	0.53
7	5	0.90	0.92	0.91
8	5	0.90	0.92	0.91
9	5	0.76	0.69	0.74
10	5	0.90	0.92	0.91

Table 2 Experimental results for $K = 10$

Run(s)	K	FCM clustering		
		Sensitivity	Specificity	Accuracy
1	10	0.84	0.73	0.79
2	**10**	**0.95**	**0.87**	**0.91**
3	10	0.62	0.46	0.56
4	10	0.86	0.85	0.85
5	10	0.95	0.93	0.94
6	10	0.94	0.72	0.82
7	10	0.86	0.85	0.85
8	10	0.94	0.72	0.82
9	10	0.86	0.85	0.85
10	10	0.86	0.85	0.85

Table 3 Experimental results for $K = 15$

Run(s)	K	FCM clustering		
		Sensitivity	Specificity	Accuracy
1	15	0.90	0.92	0.91
2	15	0.93	0.65	0.76
3	15	0.75	0.56	0.65
4	**15**	**0.81**	**0.77**	**0.79**
5	15	0.95	0.87	0.91
6	15	0.95	0.93	0.94
7	**15**	**0.95**	**0.93**	**0.94**
8	**15**	**0.83**	**1.00**	**0.88**
9	15	0.73	0.53	0.62
10	**15**	**0.86**	**0.85**	**0.85**

Table 4 Experimental results for $K = 5$

Cluster(s)	K	MKM clustering		
		Sensitivity	Specificity	Accuracy
1	5	0.36	0.30	0.32
2		*0.87*	*1.00*	*0.91*
3		0.62	0.60	0.62
4		0.66	0.80	0.68
5		0.61	0.45	0.56

Table 5 Experimental results for $K = 10$

Cluster(s)	K	MKM clustering		
		Sensitivity	Specificity	Accuracy
1	10	0.36	0.30	0.32
2		0.79	0.67	0.74
3		*0.91*	*1.00*	*0.94*
4		0.55	0.20	0.50
5		0.56	0.33	0.50
6		0.58	0.38	0.53
7		0.65	0.63	0.65
8		0.57	0.33	0.53
9		0.50	0.33	0.41
10		0.67	0.54	0.62

Table 6 Experimental results for $K = 15$

Cluster(s)	K	MKM clustering		
		Sensitivity	Specificity	Accuracy
1	15	0.36	0.30	0.32
2		0.79	0.67	0.74
3		**0.83**	**1.00**	**0.88**
4		0.77	0.52	0.62
5		0.55	0.20	0.50
6		0.67	0.71	0.68
7		0.52	0.27	0.44
8		0.60	0.44	0.56
9		0.72	0.78	0.74
10		0.68	0.83	0.71
11		0.62	0.60	0.62
12		0.59	0.40	0.56
13		0.67	0.54	0.62
14		0.57	0.38	0.50
15		0.50	0.33	0.41

Table 7 Relative performance measure of experimental analysis

Run	*K* value	Sensitivity	Specificity	Accuracy	Number of gene selected
1	5	0.94	0.76	0.85	75
4	5	0.90	0.92	0.91	203
5	5	0.94	0.76	0.85	42
7	5	0.90	0.92	0.91	75
8	5	0.90	0.92	0.91	75
10	5	0.90	0.92	0.91	42
2	10	0.94	0.72	0.82	23
5	10	0.95	0.93	0.94	219
6	10	0.94	0.72	0.82	37
8	10	0.94	0.72	0.82	37
10	10	0.86	0.85	0.85	34
1	15	0.90	0.92	0.91	104
5	15	0.95	0.87	0.91	20
6	15	0.95	0.93	0.94	19
7	15	0.95	0.93	0.94	189
8	15	0.83	1.00	0.88	29

Table 8 Significant genes selected

FCM algorithm	42, 43, 291, 1032, 1370, 1930, 2159, 2290, 2727, 2797, 2801, 3314, 4624, 5105, 5308, 5716, 6168, 6184, 6209

The accuracy of 91 % is achieved when ***K* = 5**; 203 genes are selected in Run 4, and the same is achieved for selecting 75, 75, and 42 genes in Runs 7, 8, and 10, respectively, in FCM clustering. The accuracy of 94 % is achieved when ***K* = 10**; 219 genes are selected in Run 5, and the accuracy of 85 % is achieved for 34 genes in Run 10. Also, an accuracy of 82 % is achieved for 23, 37, and 37 genes in Runs 2, 6, and 8, respectively, in FCM clustering. The accuracy of 91 % is achieved when ***K* = 15**; 104 genes are selected in Run 1, and 20 genes are selected in Run 5 in FCM clustering. The accuracy of 94 % is achieved when $K = 15$; 19 genes are selected in Run 6, and 189 genes are selected in Run 7. The accuracy of 88 % is achieved for the same value of *K*; 29 genes are selected in Run 10. The results with best accuracy obtained for $K = 5$, $K = 10$, and $K = 15$ using FCM clusters are given in Table 7.

The results of the significant genes selected by FCM algorithm are given in Table 8.

7 Conclusion

In this paper, clustering-based gene selection methods have been proposed and analyzed. It is a novel approach, since genes were selected through sequence processing of clustering approaches. The FCM and MKM clustering algorithms have been applied for different values of *K*. Again KM clustering algorithm has been performed for all the clusters produced by FCM and MKM methods. The highly correlated genes were selected from the clusters of the high accurate classification results. Out of 7,129 genes, 19 genes were selected by the proposed novel gene selection method, and it is enough to consider only such 19 genes to predict the leukemia cancer. It was observed that FCM clustering method outperformed.

The further research direction is to identify a single gene for diagnosing the leukemia cancer.

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

References

1. Stanislav Busygin, Gerrit Jacobsen, and Ewald Kramer. Double conjugated clustering applied to leukemia microarray data. In Proceedings of the 2nd SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data, 2002.
2. Aik Choon Tan and David Gilbert, Ensemble machine learning on gene expression data for cancer classification: Applied Bioinformatics 2003:2 (3 Suppl) S75–S83.
3. Cherie H. Dunphy (2006) Gene Expression Profiling Data in Lymphoma and Leukemia: Review of the Literature and Extrapolation of Pertinent Clinical Applications. Archives of Pathology & Laboratory Medicine: April 2006, Vol. 130, No. 4, pp. 483–520.
4. Yoo CK, Vanrolleghem PA. Interpreting patterns and analysis of acute leukemia gene expression data by multivariate statistical analysis. In: Barbosa Povoa A, Matos H, editors. Computer-Aided Chemical Engineering. Elsevier Science; 2004. pp. 1165–70.
5. Wei Li, Modified K-means clustering algorithm, Congress on Image & Signal Processing, IEEE, 2008, pp. 618–621.
6. T.R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 1999, Vol. 286, pp. 531–537.
7. Palanisamy, P.; Perumal; Thangavel, K.; Manavalan, R., “A novel approach to select significant genes of leukemia cancer data using K-Means clustering,” Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference on, pp. 104, 108, 21–22 Feb. 2013.