International Journal of Engineering

Journal Homepage: www.ije.ir

Modification of the Fast Global *K*-means Using a Fuzzy Relation with Application in Microarray Data Analysis

Z. Shaeiri [a], R. Ghaderi* [b]

[a] *Department of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, Iran*
[b] *Department of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, Iran*

PAPER INFO

Paper history:
Received 6 June 2012
Accepted in revised form 30 August 2012

Keywords:
Gene Expression Data Clustering
Global K-means Algorithm
Fuzzy Entropy Measure

ABSTRACT

Recognizing genes with distinctive expression levels can help in prevention, diagnosis and treatment of the diseases at the genomic level. In this paper, fast Global *k*-means (fast GKM) is developed for clustering the gene expression datasets. Fast GKM is a significant improvement of the *k*-means clustering method. It is an incremental clustering method which starts with one cluster. Iteratively new clusters are added. Since in each epoch, all data points are examined for the next cluster center, it is believed that fast GKM attains a near global solution. In the gene expression clustering problem, genes with significant differential expression levels, across the output disease classes, are important for the accurate classification of samples. Thus a fuzzy entropy measure which is designated based on maximum within class and minimum between class relevance is exerted in to the search procedure of the fast GKM. As a result, the search procedure of the proposed method is conducted in such a way to provide clusters which assembles the most discriminative genes closer to their centers. Therefore, capacity of the fast GKM which is its ability to find global clusters is managed in a profitable way. To demonstrate the usefulness of the proposed method, three published microarray datasets are used: Leukemia, Prostate, and Colon. Classification results are found robust and accurate using three public classification methods: K-NN, SVM, and Naïve Bayesian.

doi: *10.5829/idosi.ije.2012.25.04c.03*

1. INTRODUCTION

To handle the vast amounts of data that are rapidly generated, exploratory data analysis methods such as clustering methods are being increasingly popular [1]. In biomedical and pharmaceutical researches, with typical examples including gene expression microarray data analysis [2-4], functional genomic data mining [5], functional MRI image analysis [6] and etc, application of clustering methods is ubiquitous. In particular, DNA microarray data allow monitoring the expression levels of thousands of genes under several experimental conditions. A gene expression data can be viewed as a data matrix, with each row corresponding to one condition/sample and each column corresponding to one gene/feature. Each entry of this data matrix is the expression level of a gene under corresponding condition. Among the overwhelming number of genes, only a fraction of them is important in biomedical viewpoint. Beside, many conventional classifiers are developed in such a way that cannot handle a huge number of features. For this, feature selection and feature clustering methods are often used to reduce the data dimensionality in gene expression datasets [7]. Clustering can be used to group genes which share similar expression patterns. Based on similarity between objects, several clustering methods have been introduced in the literatures [8]. They are often roughly divided in two categories: partitional methods such as *k*-means, Kohonen's self organizing maps (SOM) and hierarchical approaches such as decision trees (like C4.5) and Random forests [9-11]. In hierarchical clustering (HC), clusters are generated successively based on previously generated clusters in an agglomerative (bottom-up, clumping) or divisive (top-down, splitting) procedure. In HC a tree called the dendrogram is used to represent how the objects are similar. HC has been extensively applied in gene expression data clustering [12-13]. In the partitional clustering, clusters are defined all at a time based on

*Corresponding Author Email: r_ghaderi@nit.ac.ir (R. Ghaderi) some criterion without using any hierarchical structure. *k*-means is one example of the partitional clustering methods which has been successfully used in gene expression data clustering [14]. It usually starts with *K* randomly selected cluster centers. Iteratively, the cluster centers are updated and moved toward the optimal cluster centers. In the *k*-means algorithm the resulting cluster centers are not unique. They vary depending on the choice of the starting point for the cluster centers. Thus no satisfactory robustness is observed in the further classification task. Requiring a predefined number of clusters, *K*, which has to be preset, using the hidden natural structure of the dataset, is another drawback of the *k*-means clustering method. The Fuzzy counterpart of the *k*-means clustering, is the Fuzzy *c*-means (FCM) clustering, which was proposed by Bezdek [15]. FCM allows the data points to be associated with all clusters with a degree of membership. It relaxes the hard clustering of *k*-means and it makes the clusters much believable and almost more accurate especially in the boundary region of each cluster.

The global *k*-means algorithm which first was proposed in reference [16], is one of the modified versions of the *k*-means algorithm. It is also an iterative algorithm. In GKM new clusters are generated iteratively until a predefined number of clusters is achieved. Taking in to account all the data points as the cluster center for the *k-th* cluster in each epoch, a near global solution is achieved. Although it has been shown that GKM is an improved version of the *k*-means, its applicability is limited. It's very time consuming because of *N* applications of the *k*-means in each epoch (*N* is the number of data points to be clustered). Therefore, it is not applicable for middle-sized and large datasets. To reduce the complexity of the GKM two methods have been proposed. We consider one of them, the fast GKM method, which is applicable to large datasets. In the fast GKM method instead of applying the *k*-means *N* times in each epoch, an upper bound is exerted into the criterion function. As a result, the time and space requirements are decreased. Fast GKM is less sensitive to initialization in comparison with the *k*-means clustering, and is faster in comparison with the GKM clustering. In this paper a modified version of the fast GKM is introduced. A fuzzy entropy relation is used to guide the choice of the new cluster center in each epoch. To some extent, it can be comparable with the FCM as both improve the *k*-means using fuzzy principals, but there are some differences; FCM starts with *c* clusters while the proposed method starts with one cluster. In FCM all data points have a degree of membership in all clusters but in the proposed method, like the *k*-means, any data point is associated with only one cluster. The results show that the proposed method is more accurate and more robust in comparison with the *k*-means, the global *k*-means, and the fast global *k*-means methods. The rest of the paper is as follows: in the next section the details of the *k*-means, the GKM, and the fast GKM algorithms are presented. In section 3 the modified GKM is introduced. Section 4 includes the experimental results and section 5 concludes this paper.

2. *K*-MEANS AND THE GLOBAL *K*-MEANS

One of the well-known and basic partitional clustering algorithms is the *k*-means algorithm. Having influence on many other clustering methods, this algorithm is one of the staple and very popular clustering algorithms. Suppose that a finite set of data points $X=\{x_1,x_2,...,x_N\}$ in the *f* dimensional space is given, where $x_i \in R^f$. *k*-means is an iterative clustering algorithm and successively seeks for the optimal clusters by minimizing the sum-of-squared-error criterion function below:

$$J_k(W,M)=\sum_{j=1}^{K}\sum_{i=1}^{N} w_{ij}\left\|x_i - m_j\right\|^2 \tag{1}$$

In which:

$W=\{w_{ij}\}$,

$w_{ij}=\begin{cases}1 & if\ \ x_i \in cluster\ \ j\ , \\ 0 & otherwise\end{cases}$

$\sum_{j=1}^{K} w_{ij}=1\ \ \forall i$’

$M=[m_1,m_2,...,m_K]$, is the cluster center matrix, and $m_j=(1/N_j)\sum_{i=1}^{N} w_{ij}x_i$ is the center of the *j-th* cluster with N_j objects. The *k*-means algorithm proceeds as follows:

- Choose *K* arbitrary data points for the *K* cluster centers, (not necessarily belonging to *X*).
- Allocate data points to their nearest cluster center and produce *K* clusters from *X*.
- Recompute the cluster centers, using the criterion function, and go back to step 2 and continue until no more data points change their clusters.

Using a hill-climbing approach for finding the clusters, *k*-means suffers from some drawbacks such as easy getting in to the local minima, and sensitivity to the initialization. Requiring the number of clusters to be predetermined is another drawback of the *k*-means.

The global *k*-means (GKM) algorithm is an incremental algorithm which starts with one cluster. The center of this cluster is the mean of all data points in the dataset. In each epoch, one cluster center is added. The number of epochs is equal to the number of clusters, *K*, which can be determined automatically from the hidden knowledge within the dataset. For instance, a criterion such as the mean or the maximum of between data point's distances can be employed to decide about the number of the clusters. Details of the GKM are as follows:

- Set $c = 1$ and compute the mean of the set X : $m_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad x_i \in X$
- Let $c = c + 1$, if $c > K$ then stop, otherwise go to the next step.
- Consider the $c - 1$ previously made cluster centers $m_1, m_2, \ldots, m_{c-1}$, check all the data points $(x_1, x_2, \ldots, x_N)$ as a candidate for the *c-th* cluster center. Apply the *k*-means algorithm to each of them and choose the best clusters and their centers $m_{new,1}, m_{new,2}, \ldots, m_{new,(c-1)}, m_{new,c}$.
- Set $m_i = m_{i,new}$ and go back to step 2.

As in the GKM, in each epoch, every data point is considered as the *k-th* cluster center; a near global solution is expected [17]. However, because of the application of the *k*-means in a couple of times, in each iteration, this algorithm is very time consuming and is not efficient for high dimensional datasets. Two methods have been proposed to reduce the computational time. We describe here only one of them, the fast GKM method, which is applicable for high dimensional datasets.

Consider the $c - 1$ clusters obtained so far, and the corresponding value of the criterion function $J^*_{c-1} = J_c(W, [m_1, m_2, \ldots, m_{c-1}])$ in Equation (1). Instead of applying the *k*-means for each data point, the fast GKM method computes an upper bound on J^*_c:

$$J^*_c \le J^*_{c-1} - r_i \tag{2}$$

where:

$$r_i = \sum_{j=1}^{N} \max\{0, l^j_{c-1} - \|x_i - x_j\|^2\}, \qquad i = 1, 2, \ldots, N \tag{3}$$

The l^j_{c-1} is the distance between *j-th* data point and its nearest cluster center among the $c - 1$ cluster centers which are obtained so far. For each data point, r_i is computed. r_i shows the value by which the criterion function defined in Equation (1) decreases if x_i is chosen as the *c-th* cluster center. The data point with maximum r_i is chosen as the *c-th* cluster center.

3. THE MODIFIED GKM METHOD FOR GENE EXPRESSION DATA CLUSTERING

The main contribution of this paper is a modification to improve the cluster validity of the fast GKM method clustering. As it is known, proceeding of the *k*-means, the global *k*-means, and the fast global *k*-means are all based on solving the optimization problem defined in Equation 1. This optimization is based on minimizing the distance between within cluster objects. When applied to gene expression datasets for dimensionality reduction purposes, the fast GKM method easily and rapidly find a predefined number of clusters of genes. The genes in each cluster are similar to each other, but there is no guarantee that they are able to be effective and informative for further classification tasks. In Figure1, two of the ten clusters containing four genes which are found by the fast GKM method, are shown. As can be seen, in these clusters, genes are very coherent and similar in terms of their expression levels. However, they are not differentially expressed across the output classes to be efficient and effective for the classification.

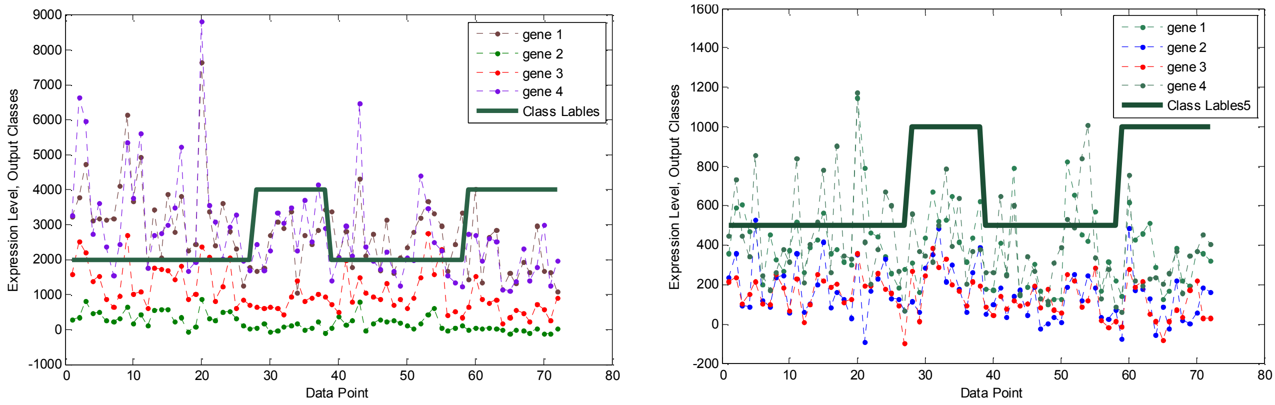


Figure 1. Two clusters found by the fast GKM method for Leukemia dataset. It may be considerable that in each cluster, there are many non-differentially expressed genes. Although, they are coherent and similar, they may not be effective in classification procedure as they don't show any interesting changes during the experiments of the time spots.

To improve the ability of the fast GKM method to be able to produce the clusters with sufficiently differentially expressed features, in this regard genes, a modification based on a fuzzy entropy relation is proposed [18]. The only parameter that has to be adjusted is a fuzzy neighborhood distance, α , which can be derived automatically from the dataset.

3. 1. Fuzzy Relation Matrix

Let X be a nonempty finite set of N real-valued data points with F features as in (4). A fuzzy entropy relation matrix M(R) on X is proposed as follows:

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^F \\ \vdots & \ddots & \vdots \\ x_N^1 & \cdots & x_N^F \end{pmatrix} \tag{4}$$

$$M(R) = \begin{pmatrix} e_1^1 & \cdots & e_1^F \\ \vdots & \ddots & \vdots \\ e_N^1 & \cdots & e_N^F \end{pmatrix} \tag{5}$$

In which:

$$e_i^f = \begin{cases} 0 & \forall x_j^f , \left| x_i^f - x_j^f \right| \geq \alpha \\ \sum_j \left(1 - \dfrac{\left| x_i^f - x_j^f \right|}{\alpha} \right) & \forall x_j^f , \left| x_i^f - x_j^f \right| < \alpha \end{cases} \tag{6}$$
$$f = 1,2,...,F \qquad i,j = 1,2,...,N$$

where α is a predefined fuzzy neighborhood parameter. Feature *f* is sufficiently differentially expressed feature, if for each *i* and all *j* of between class data points, e_i^f is close to zero and for each *i* and all *j* of the within class data points it is close to one. For every feature, the sums of e_i^f of between class, E_{fb} , and within class data points, E_{fw} are computed as follows:

$$E_{fw} = \sum_{i=1}^{N_w} e_i^f \qquad f = 1,2,...,F \tag{7}$$

$$E_{fb} = \sum_{i=1}^{N_b} e_i^f \qquad f = 1,2,...,F \tag{8}$$

The number of within class data points is N_w , and E_{fw} is the sums of e_i^f over all of them. The number of beetween class data points is N_b , and E_{fb} is the sums of e_i^f over all of them. It is clear that $N = N_b + N_w$. Features with higher value of E_{fw} and smaller value of E_{fb} are significant choices for the classification.

3. 2. Modified Fast GKM

Given a gene expression data matrix containing *N* data points and *F* genes as in Equation (4), the proposed method proceeds as follows:

Step 1: Set $c = 1$ and compute the mean of all genes in the set *X* as the first cluster center: $m_1 = \frac{1}{F}\sum_{i=1}^{F} x_i \qquad x_i \in X$

Step 2: Let $c = c + 1$, if $c > K$ then stop, otherwise go to the next step. (*K* is a predefined number of clusters)

Step 3: Consider the $c - 1$ cluster centers obtained so far, $m_1, m_2, ..., m_{c-1}$. For each gene, calculate the r_f defined in Equation 3. Then for each gene, compute the Er_f as follows:

$$Er_f = \left(\frac{E_{fw}}{E_{fw} + E_{fb}} \right) \times r_f \tag{9}$$

Compute $Er_{\max} = \max(Er_f), f = 1,2,...,F$ and choose the gene with the largest value of Er_f for the next cluster center.

Step 4: Refine the clusters by applying the *k*-means algorithm to solve the *k*-partition problem. Take the refined cluster centers:

$$m_{new,1}, m_{new,2}, \ldots, m_{new,(c-1)}, m_{new,c} \ .$$

Step 5: Set $m_i = m_{i,new}$ and go back to step 2.

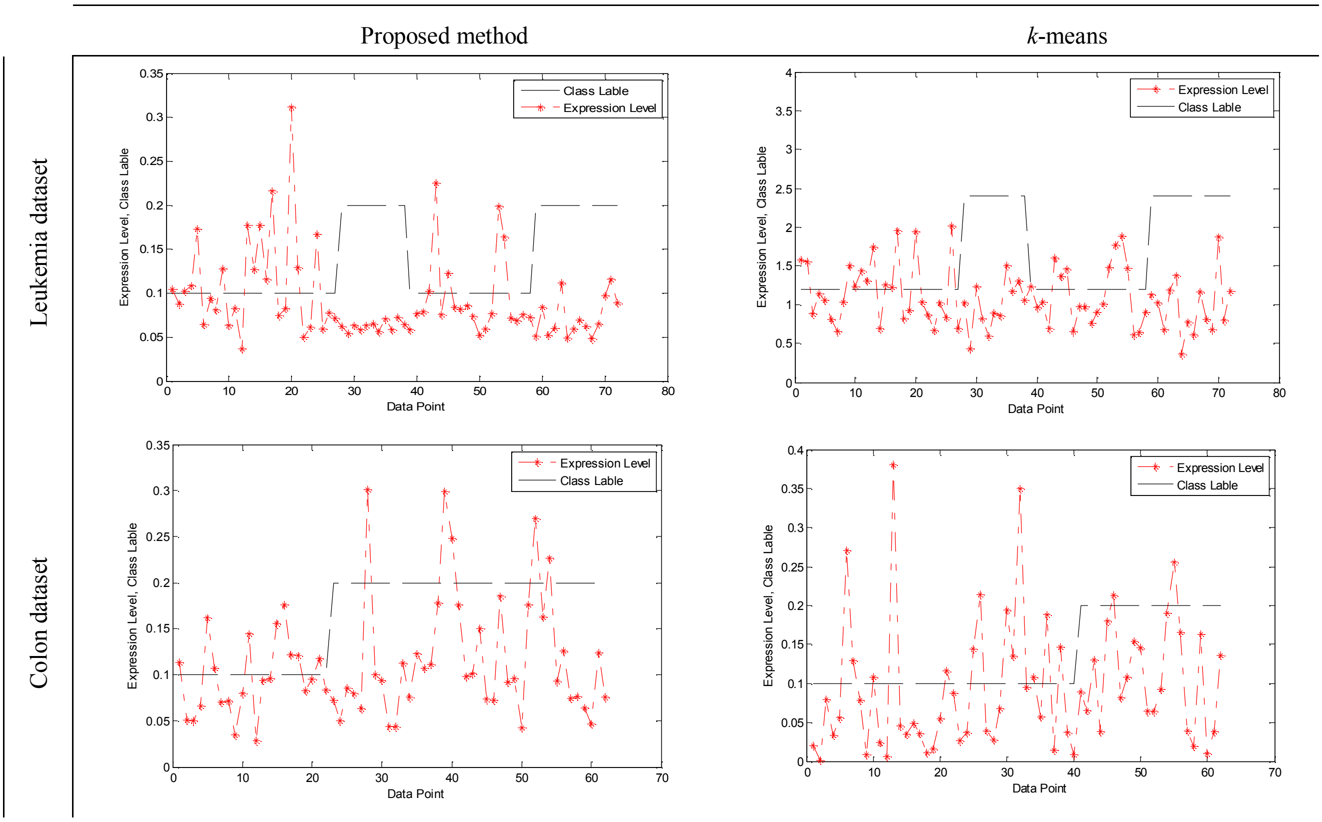
4. EXPERIMENTS ON GENE EXPRESSION DATASETS

The performance evaluation of the proposition is done, using three published micro array datasets: Leukemia [19], Colon [20], and Prostate [21]. Details of the datasets are presented in Table 1.

After running the proposed clustering method on the datasets, *K* (a predefined parameter) clusters are obtained. From each cluster, five top genes are selected (five genes with minimum distance to their cluster centers). As a result, the dimensionality of the datasets was reduced significantly. In Figure 2 and 3, one of the top genes selected with the proposed clustering method is compared with that of the *k*-means and the fast GKM methods. As can be observed, in the proposed method, there is a meaningful correlation between the top gene and the output class labels. It is expected because of imposing the term $E_{fw} / (E_{fw} + E_{fb})$ in to the searching procedure in the proposed method. For evaluating the classification performance the reduced datasets are fed to three well-known classifiers: KNN, SVM, and Naive Bayesian. 10-fold cross validation is used to compute the classification performances. The results are presented in Figure 4 and Table 2. In Figure 4 the classification accuracies of KNN, SVM, and Naïve Bayesian classifiers are depicted against the fuzzy neighborhood parameter, α. From the Figure 4 it can be implied that, changing α has little influence on the classification performance of the Leukemia dataset. For the prostate and colon datasets the classification performance of the Naïve Bayesian and SVM classifiers are somehow sensitive to α. Naïve Bayesian classifier discriminates the samples based on maximum a posteriori probability, while considering the features independent. Thus, dependency among features can increase the number of misclassified samples. The proposed clustering method doesn't absolutely and directly seek the independent features. Its search procedure to find a cluster center is based on two different and maybe independent criteria: r_f and $E_{fw}/(E_{fw}+E_{fb})$. Among them r_f doesn't take in to consideration the independency of features. It shows the amount one can reduce the criterion function defined in Equation 1, if f is chosen as the *k-th* cluster center. Therefore, using the Naïve Bayesian classifier, some misclassified data points are expected. There is no difference taking a large or small value for α. SVM classification is based on searching for the optimal separating hyper plane which maximizes the margin between two classes. A small value of α means that the between class separation and within class correlation is stronger. Taking a proper value for α, the reduced dataset would be a linearly separable set. SVM classifier can find a unique optimal hyper plane for which the margin between the projections of the training points of the two different classes is maximized. KNN classifier has a similar reaction to changing α. KNN classifies the data points using a majority voting procedure of the K nearest neighbors data points. Thus, a proper value for α can produce features considerably compatible to the ability of KNN to find the output classes. Table 2 contains the simulation results of the comparison between the proposed method, *k*-means, and the fast GKM methods. It can be seen that the proposed clustering method can outperform the previous similar methods.

TABLE 1. Summary of the gene expression datasets used for classification

Gene expression dataset	# Samples	# Features	# Classes
Leukemia	72	7129	2
Prostate	102	12600	2
Colon	62	2000	2



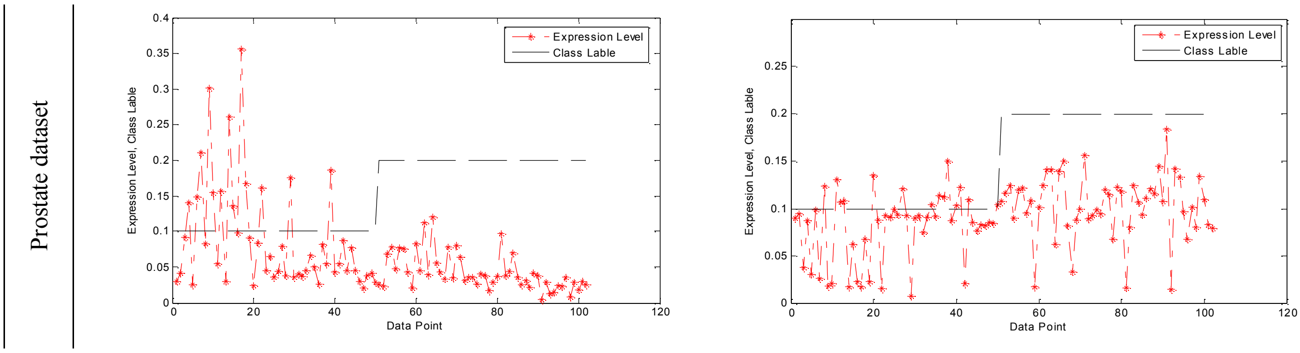


Figure 2. Nearest gene to the first cluster center, found by the *k*-means, right, and found by the proposed method, left. The gene which is selected using the *k*-means doesn't change significantly in the time spots, while the selected gene of the proposed method has a meaningful correlation with the output classes and therefore can be effective for classification procedure. (The expression levels are being normalized.)

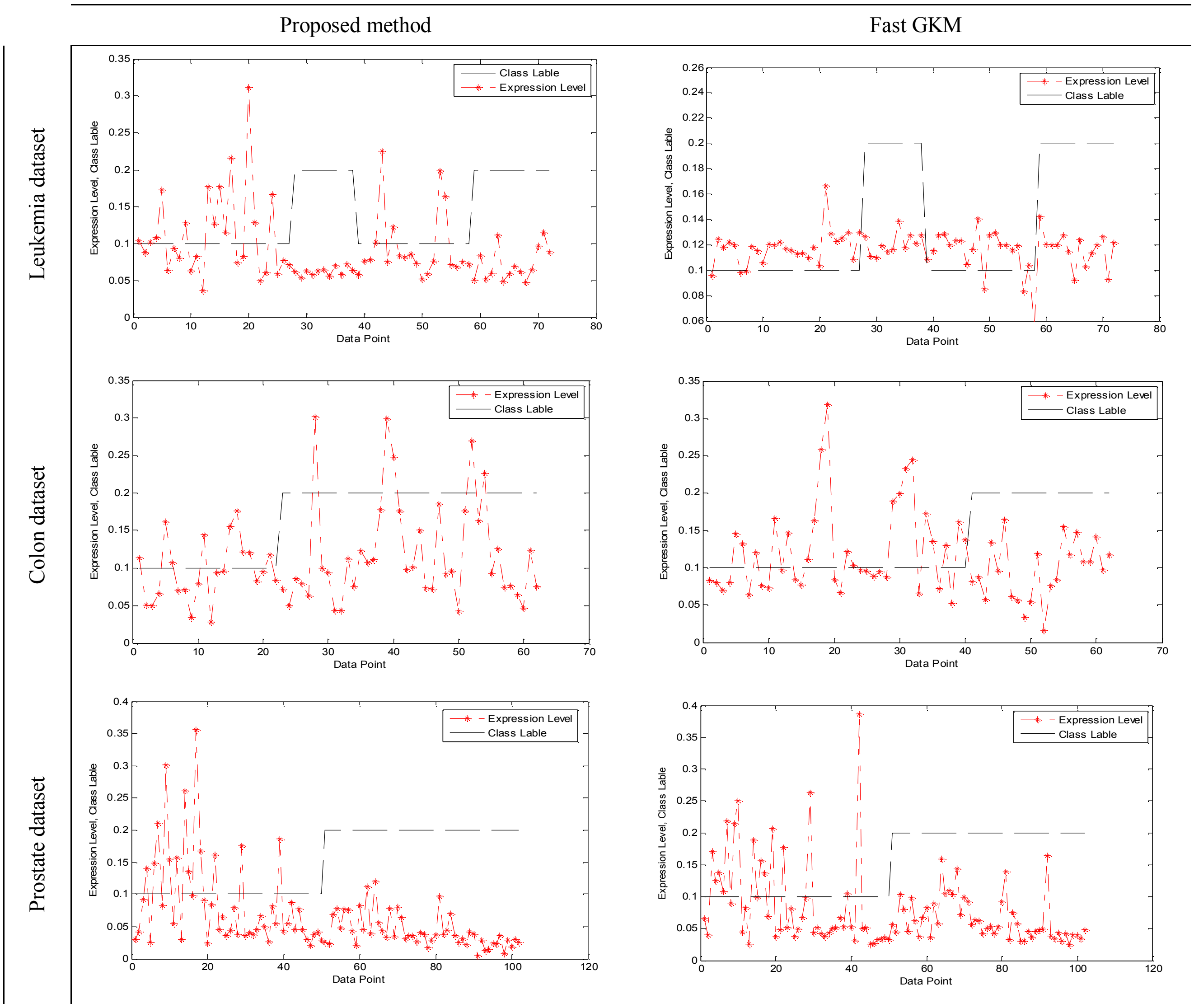
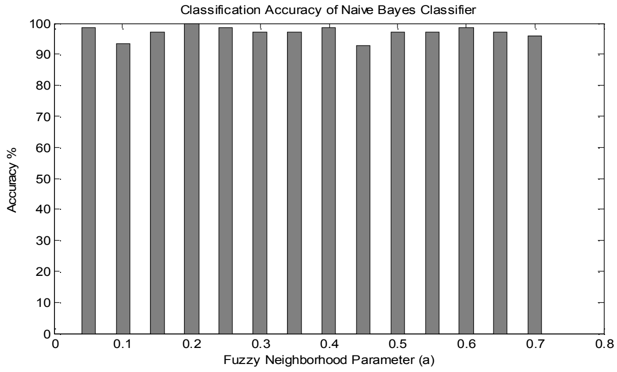
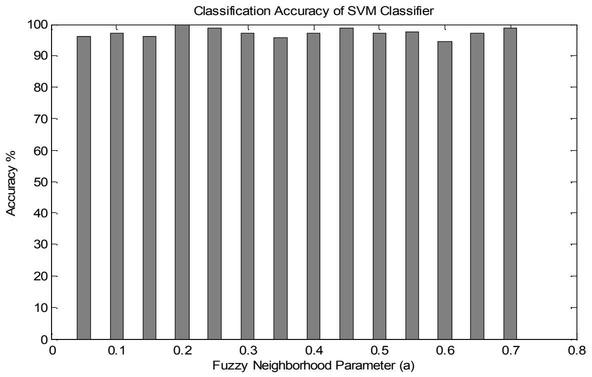
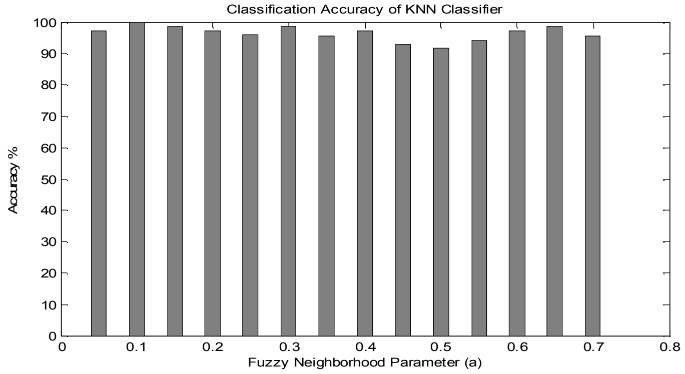


Figure 3. Nearest gene to the first cluster center, found by the fast GKM, right, and the proposed method, left. (The expression levels are being normalized.)

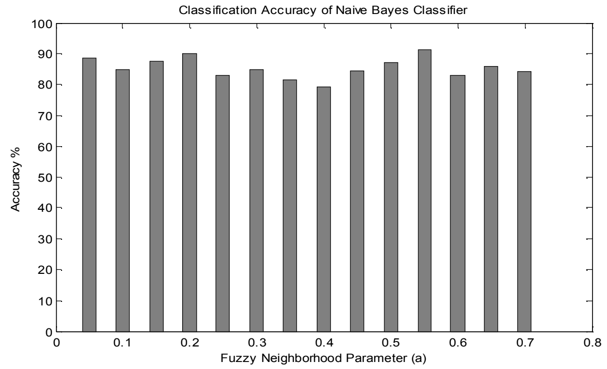
Leukemia

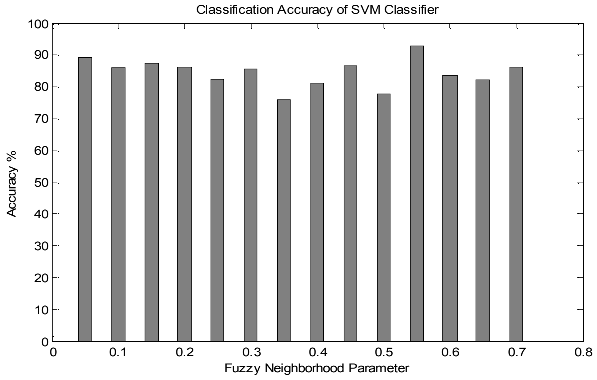


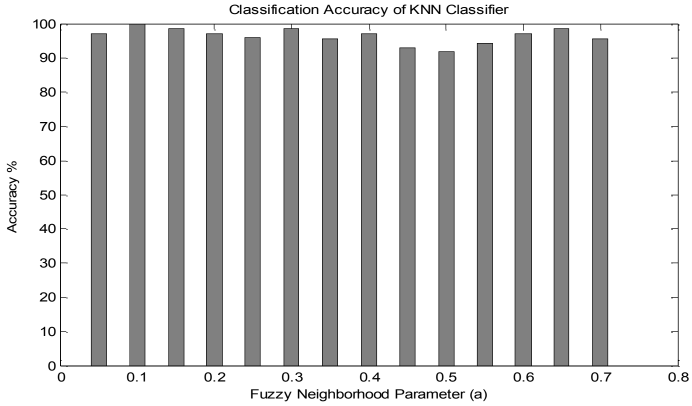




Prostate







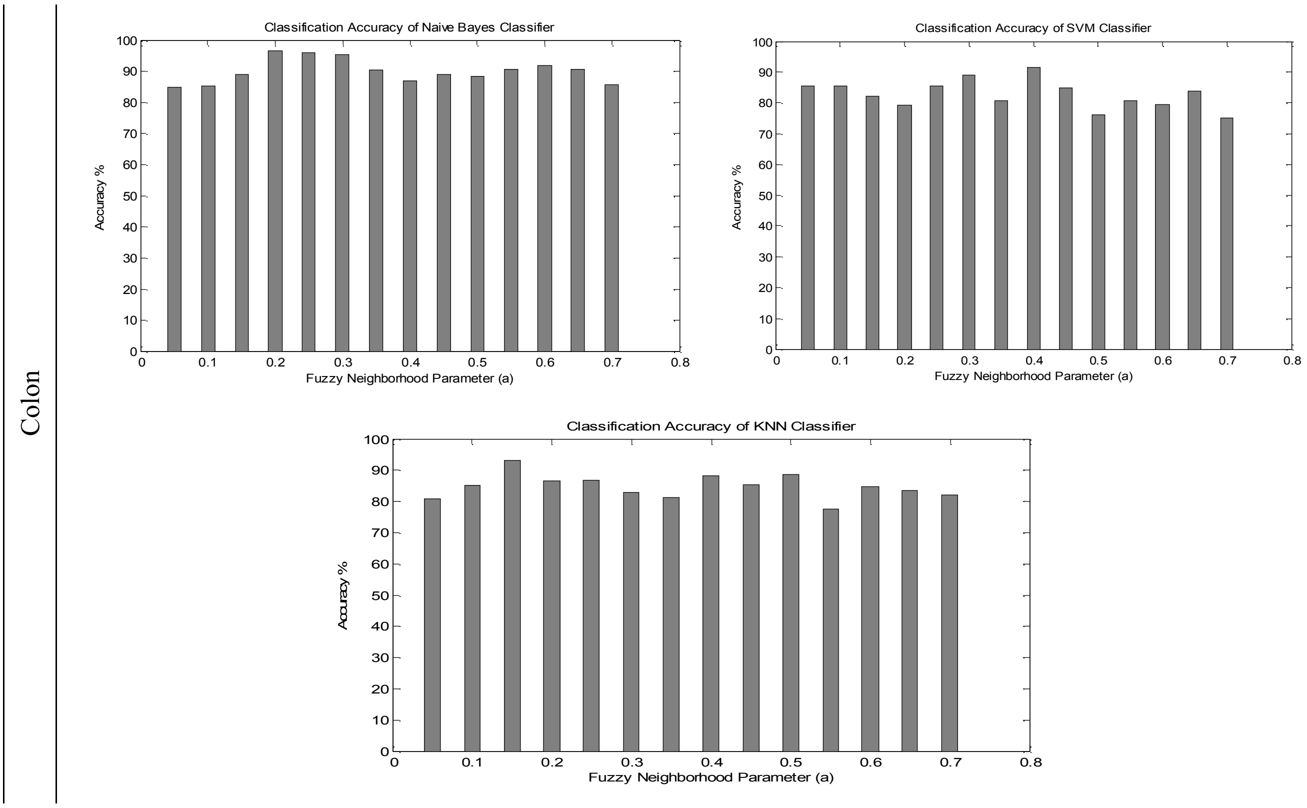


Figure 4. Classification performance of the proposed method, using Naïve Bayes, SVM, and KNN classifiers.

TABLE 2. Classification accuracy of SVM, Naïve Bayes, and KNN classifiers.

Method / Dataset		**Leukemia**	**Prostate**	**Colon**
k-means	KNN	80.75	70	75
	SVM	85.53	76	85.23
	NB	84.28	73.27	66
Fast GKM	KNN	90.35	64.5	77.3
	SVM	90	78.43	82.68
	NB	87.32	73.32	83.64
Proposed method	KNN	100 (α=0.1)	100 (α=0.1)	93 (α=0.15)
	SVM	100 (α=0.2)	92 (α=0.55)	95.23 (α=0.4)
	NB	100 (α=0.2)	91.5 (α=0.55)	97.12 (α=0.2)

5. CONCLUSION

A modified version of the fast global *k*-means (fast GKM) clustering method for clustering the gene expression datasets is proposed. The fast GKM algorithm is a modified version of the *k*-means algorithm. It is an iterative algorithm which starts with one cluster containing the whole data points. Its proceeding is based on minimizing the criterion function of the Equation 1 which is based on minimizing the distance between any data point and its cluster center. When applying on gene expression datasets, clusters of sufficiently coherent and similar genes are produced; Although similarity of genes are important, it's not enough to conclude about the classification accuracy of the classifiers on the reduced datasets. Genes with no considerable changes during the experiments are often redundant because of having no interesting information. In this paper a modified version of the GKM is proposed. A fuzzy entropy relation is used to control the selection of new cluster centers in each epoch. The experimental results are found accurate and robust in comparison with some similar clustering methods.

6. REFERENCES

1. Xu, R. and Wunsch, D.C., "Clustering algorithms in biomedical research: A review", ***IEEE Reviews in Biomedical Engineering***, Vol. 3, (2010), 120-154.
2. Newman, A.M. and Cooper, J.B., "AutoSOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number", ***BMC Bioinformatics***, Vol.11(117), (2010), 1-48.
3. Au, W.H., Chan, K.C.C., Wong, A.K.C. and Wang, Y., "Attribute clustering for grouping, selection, and classification of gene expression data", ***IEEE Transactions on Computational Biology and Bioinformatics***, Vol.2, (2004), 83-101.
4. Baya, A.E. and Granitto, P.M., "Clustering gene expression data with a penalized graph-based metric", ***BMC Bioinformatics***, Vol.12(2), (2011), 1-18.
5. Kohane, I.S., Kho, A.T. and Butte, A.J., "Microarrays for an integrative genomics", Londo, The MIT Press, (2003).
6. Cheng, Q., Zhou, H. and Cheng, J., "The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with application to high-dimmensional data", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol.33(6), (2011), 1217-1233.
7. Montazer, G.A., "A new approach for knowledge-based systems reduction using rough sets theory", ***International Journal of Engineering (IJE Transaction A)***, Vol.16(2), (2003), 157-162.
8. Duda, R.O., Hart, P.E. and Stork, D.G., "Pattern classification", New York, Wiley, (2001).
9. Hartigan, J., "Clustering algorithms", New York, Wiley, (1975).
10. Wang, H., Zheng, H. and Azuaje, F., "Poisson-based self-organizing feature maps and hierarchical clustering for serial analysis of gene expression data", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, Vol.4(2), (2007), 163-175.
11. Zhang, B. and Pham, T.D., "Phenotype recognition with combined features and random subspace classifier ensemble", *BMC Bioinformatics*, Vol.12(128), (2011), 1-13.
12. Handl, J. and Kell, D.B., "Computational cluster validation in post genomic data analysis", ***BMC Bioinformatics***, Vol.21, (2005), 3201-3212.
13. Tan, T.Z. and Quek, Ch., "A novel biologically and psychologically inspired fuzzy decision support system: hierarchical complementary learning", ***IEEE Transactions on Computational Biology and Bioinformatics***, Vol.5(1), (2008), 67-79.
14. Du, Z., Wang, Y. and Ji, Z., "PK-means: A new algorithm for gene clustering", ***Computational Biology***, Vol.32, (2008), 243-247.
15. Bezdek, J., "Pattern recognition with fuzzy objective function algorithms", New York, Plenum, (1981).
16. Likas, A., Vlassis, M. and Verbeek, J., "The global k-means clustering algorithm", ***Pattern Recognition***, Vol.36, (2003), 451-461.
17. Bagirov, A.M., "Modified global *k*-means algorithm for minimum sum-of-squares clustering problems", ***Pattern Recognition***, Vol.41, (2008), 3192-3199.
18. Hu, Q., Yu, D. and Xie, Z., "Information-preserving hybrid data reduction based on fuzzy-rough techniques", ***Pattern Recognition letters***, Vol.27, (2005), 414-423.
19. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. and Bloomfield, C.D., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", ***Science***, Vol.286, (1999), 531-537.
20. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", ***National Academy of Sciences of the United States of America***, Vol.96, (1999), 6745-6750.
21. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Colub, T.R. and Sellers, W.R., "Gene expression correlates of cilinical prostate cancer behavior", ***Cancer Cell***, Vol.1(2), (2002), 203-209.

Modification of the Fast Global *K*-means Using a Fuzzy Relation with Application in Microarray Data Analysis

Z. Shaeiri [a], R. Ghaderi [b]

[a] *Department of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, Iran*
[b] *Department of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, Iran*

PAPER INFO

Paper history:
Received 6 June 2012
Accepted in revised form 30 August 2012

Keywords:
Gene Expression Data Clustering
Global K-means Algorithm
Fuzzy Entropy Measure

چکیده

تعیین ژن‌هایی که در نمونه‌های مختلف دارای بیان افتراقی هستند، می‌تواند در جلوگیری، تشخیص و درمان بیماری‌ها در سطح ژنی مفید واقع شود. در این مقاله، الگوریتم Global *k*-means (GKM) سریع، برای خوشه‌بندی داده‌های بیان ژنی بهبود داده می‌شود. خوشه بندی GKM سریع یکی از نسخه‌های بهبودیافته‌ی خوشه‌بندی *k*-means است. در این الگوریتم تکرار شونده در گام اول یک خوشه تعریف می‌شود. در هر گام بعدی یک خوشه اضافه می‌گردد. در هر گام برای انتخاب مرکز خوشه‌ی بعدی تمام نمونه‌های فضای داده‌ها بررسی می‌شوند. بنابراین این الگوریتم دارای پاسخی نزدیک به پاسخ بهینه‌ی کلی است. در بحث خوشه‌بندی ژن‌ها، ژن‌هایی که در میان نمونه‌های بیمار و سالم و یا نمونه‌های با انواع مختلف یک بیماری، دارای سطوح بیان افتراقی قابل توجهی هستند، دارای اهمیت می‌باشند. با توجه به این موضوع، در این نوشته جهت بهبود خوشه‌بندی GKM سریع، یک رابطه‌ی آنتروپی فازی استفاده می‌شود. با این روش قدرت خوشه‌بندی GKM سریع در جهت کاربرد مزبور، برنامه‌ریزی و هدایت می‌شود. جهت نشان دادن نتایج، سه مجموعه داده‌ی بیان ژن لوکمی، پروستات و کولن مورد استفاده قرار می‌گیرند. بهبود دقت و پایداری روش پیشنهادی، براساس نتایج طبقه‌بندی با سه روش متداول SVM، Naïve Bayesian و KNN نسبت به روش‌های مشابه پیشین، قابل توجه می‌باشد.

doi: *10.5829/idosi.ije.2012.25.04c.03*