2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

# A Novel Feature Measure for Fuzzy Clustering Algorithm on Microarray Data

Tian Yu and JinMao Wei

***Abstract*— Fuzzy clustering algorithm is employed in gene microarray analysis to discover the strength of the association between genes and different clusters. Gene-based fuzzy clustering algorithm just employs all instances' values of a certain gene as this gene's features. In some sense, the original feature vector can hardly provide comprehensive discriminative information of the gene. In this paper, a novel feature vector by the proposed measure for each gene is employed in fuzzy clustering algorithm. The proposed feature vector can provide information about the influence of a given gene for the overall shape of clusters. By analysis and experiment upon microarray data sets, the performance of the fuzzy clustering algorithm based on proposed feature vector is compared with that of some classical clustering algorithms. The results demonstrate that the fuzzy clustering algorithm based on proposed feature vector is capable of obtaining better clusters than other contrast algorithms. The results by classifiers based on different clustering algorithms demonstrate that the proposed feature vector can get the same or better accuracy than the original feature vector.**

## I. Introduction

WITH the rapid progress of microarray technology, gene expression data have been becoming available at clinic and patient-tailored therapy is becoming possible. To reveal natural structures and identify interesting patterns in the underlying data, clustering analysis is employed. It seeks to partition a gene data set into clusters so that genes within a cluster are more similar to each other than the genes in different clusters[1]. Coexpressed genes can be grouped in clusters based on their expression patterns[2], [3]. In such gene-based clustering, the genes are treated as the objects, while the instances are the features.

Gene-based clustering has proven to be helpful to understand gene function, gene regulation, cellular processes and subtypes of cells. Genes with similar expression patterns can be clustered together with similar cellular functions[1]. Based on this common view, many crisp clustering algorithms[4], [5], [6], [7] are employed for analysing interdependences of genes on microarray data. By these clustering algorithms, genes are divided into distinct clusters, where every gene belongs to exactly one cluster.

Jiang[9], [8] demonstrates that gene expression data are often highly connected[8], and clusters may be highly intersected with each other or even embedded one in another[9]. Fuzzy clustering algorithm[10] is an alternative method employed on microarray gene data[11], [12], [13], [14]. The fuzzy clustering analysis is an analytical method, which based on fuzziness of the things index, realized through the application of the fuzzy mathematics method. Fuzzy clustering is a process of assigning membership levels, and then using them to assign genes to one or more clusters. This indicates the strength of the association between that gene and different clusters. Dembele[11] proposed a method to choose the fuzziness parameter of FCM[11] in microarray data. Tari[13] proposed a semi-supervised clustering method called GO-FCM based on the FCM and the Gene Ontology annotations as prior knowledge to guide the process of clustering. Gasch[14] focused that the corresponding genes are often coexpressed with different groups of genes under different situations. Most gene-based fuzzy clustering algorithms just employ all instances' values of a certain gene as this gene's features. Pearson Distance[14], Euclidean distance[11], [13], Minkowshi distance[15] are widely used. The information about the instance categories or response variables in train data was ignored. These unsupervised measures are directly computed from the gene expressions without using any information about the instance categories or response variables. However, these information is significant for incorporating to find groups of coregulated genes with strong association to the instances' categories[16].

In this paper, by introducing a novel feature vector by the proposed measure for each gene, the instance category information in train data is employed in fuzzy clustering algorithm. The proposed feature vector can provide information about the influence of a given gene for the overall shape of clusters. By analysis and experiment upon microarray data sets, the performance of the fuzzy clustering algorithm based on proposed feature vector is compared with that of some classical clustering algorithms. The results demonstrate that the fuzzy clustering algorithm based on proposed feature vector is capable of obtaining better clusters than other contrast algorithms.

## II. Method

Gene expression data are often highly connected[8], and clusters may be highly intersected with each other or even embedded one in another[9]. Most algorithms for gene-based clustering cannot effectively handle this situation[1]. The clusters and the relationship between the clusters are both interested by experts. A clustering algorithm, which can provide these information, would be more favored by the

Tian Yu and JinMao Wei are with the College of Computer and Control Engineering, NanKai University, Tian Jin (email: yutian198411@163.com, weijm@nankai.edu.cn).

This work was supported by the National Natural Science Foundation of China under grant 61070089, the Science Foundation of Tianjin under grant 10JCZDJC15800 and 13ZCZDGX02200, the National Natural Science Foundation of China under Grant No.61201424, the Natural Science Foundation of Tianjin, China under Grant No.12JCYBJC10100

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

biologists. Fuzzy clustering algorithm[10] is an alternative method.

### *A. New Feature Vector for Gene*

A gene expression data set can be represented by a real-valued expression matrix $M = \{x_{ij} | 1 \leq i \leq n, 1 \leq j \leq m\}$ in Fig.1, where the rows $X$=$\{\overrightarrow{X_1}, \overrightarrow{X_2}, \ldots, \overrightarrow{X_n}\}$ form the expression patterns of genes, the columns $Ins$=$\{\overrightarrow{Ins_1}, \overrightarrow{Ins_2}, \ldots, \overrightarrow{Ins_m}\}$ represent the expression profiles of instances, and each cell $x_{ij}$ is the measured expression level of gene $\overrightarrow{X_i}$ in instance $\overrightarrow{Ins_j}$. $\overrightarrow{CA}$ is the vector of instances' categories. In this paper, the "category" specifically means the index for instances.

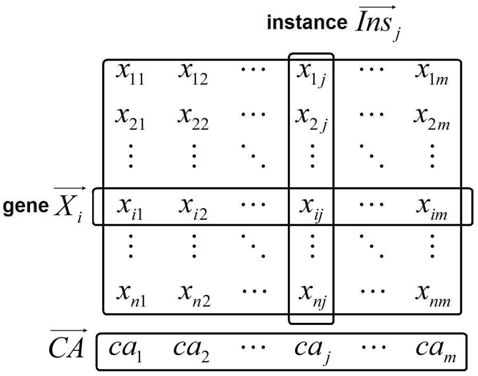
instance $\overrightarrow{Ins_j}$
gene $\overrightarrow{X_i}$
$\overrightarrow{CA}$
$x_{11}$ $x_{12}$ $\cdots$ $x_{1j}$ $\cdots$ $x_{1m}$
$x_{21}$ $x_{22}$ $\cdots$ $x_{2j}$ $\cdots$ $x_{2m}$
$x_{i1}$ $x_{i2}$ $\cdots$ $x_{ij}$ $\cdots$ $x_{im}$
$x_{n1}$ $x_{n2}$ $\cdots$ $x_{nj}$ $\cdots$ $x_{nm}$
$ca_1$ $ca_2$ $\cdots$ $ca_j$ $\cdots$ $ca_m$

Fig. 1. Gene Expression Matrix

The gene-based fuzzy clustering algorithm is a process which finds $c$ disjoint clusters, $CL_1$, $CL_2$, ..., $CL_c$, of correlated genes by linking each gene in $\{\overrightarrow{X_1}, \overrightarrow{X_2}, \ldots, \overrightarrow{X_n}\}$ to all clusters via a real-valued vector of indexes, $u_{ij}$, which lie between 0 and 1. Indexes close to 1 indicate a strong association to the cluster. Inversely, indexes close to 0 indicate the absence of a strong association to the corresponding cluster. The vector of indexes represents the relationships between a gene and different clusters. Formally, we define fuzzy gene clustering as a process that $\forall$ $\overrightarrow{X_i}$, $i$ = 1, 2, ..., $n$, $\overrightarrow{X_i}$ is linked to $\forall$ $CL_r$, $r$ = 1, 2, ..., $c$ via $u_{ij}$ where $0 \leq u_{ij} \leq 1$ and $CL_r \cap CL_s = \emptyset$ for all $r \neq s$, $r$, $s$ = 1, 2, ..., $c$.

Most gene-based fuzzy clustering algorithms just employ all instances' values in gene $i$, $\overrightarrow{X_i}$=$\{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}\}$, as this gene's features. Although, the information about the instance categories or response variables are available in a gene expression data set, they are ignored by these clustering algorithms. A novel feature vector, which can provide the instance category information, is proposed for gene-based fuzzy clustering.

For gene $i$, $\overrightarrow{R_i}$ replaces the traditional feature vector which simply uses all instances' values of gene $i$. $\overrightarrow{R_i}$ reflects the relationship between the centroid of all instances and each of all the centroids of the instances from different categories. It can be calculated as Eq.1:

$$\overrightarrow{R_i} = \{|\frac{\overline{x_{i1}} - \overline{x_i}}{m_1 S_i}|, |\frac{\overline{x_{i2}} - \overline{x_i}}{m_2 S_i}|, \cdots, |\frac{\overline{x_{iK}} - \overline{x_i}}{m_K S_i}|\}. \quad (1)$$

$$S_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \overline{x_{ik}})^2. \quad (2)$$

$\overline{x_i}$ is the mean expression value of all instances for gene $i$ and $\overline{x_{ik}}$ refers to the mean expression value of instances in category $k$ for gene $i$. $S_i^2$ represent the inter-category variance and $m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}$. There are $K$ categories of instances. $C_k$ refers to indices of the instances in category $k$ and $n_k$ is the number of instances included in category $k$. For gene $i$, $|\frac{\overline{x_{ik}} - \overline{x_i}}{m_k S_i}|$ is the $t$-statistic value between the mean of all instances and the mean of instances in category $k$. $\overrightarrow{R_i}$ records the relationships between centers of different categories and the centroid of all the instances instead of all instances' values in gene $i$.

In order to comparatively explain how $\overrightarrow{R_i}$ is used to effectively determine whether a gene is different from another through addressing discriminative capability, we use examples in Fig.2 and Fig.3. $D$ refers to the projection distance on the axis of a certain gene. We picture all instances in the two-dimensional space. Gene $w$ can be regarded as a random gene and it is used for convenient explanation. One space is formed by gene $i$ and gene $w$, the other is formed by gene $j$ and gene $w$.

To the gene-based clustering algorithms, features used by conventional measures are just instances' values of certain gene, just like what is shown in Fig.2. Gene $i$ and gene $j$ are shown respectively in the figure. There are 8 instances without information about their categories. The instances in two spaces have completely opposed distribution. For example, instance 1 has the smallest value in gene $i$ but has the biggest value in gene $j$. Similarly, instance 2 has the second smallest value in gene $i$ but has the second biggest value in gene $j$. According to conventional measures, gene $i$ and gene $j$ are compared by the original feature vector $\{D_1$, $D_2$, ..., $D_8\}$ in gene-based clustering algorithms. In this case, they are supposed to have notable differences.

In Fig.3, the information about the instance categories or response variables in train data is introduced. There are 8 instances and 3 categories. The instances of the same shape belong to the same category. For example, point 1 and point 2 belong to Category 1 in train data, so they are pictured with the same shape in Fig.3. Each of the three inverted triangles represents the mean expression value of instances in the corresponding category. The diamond is the mean expression value of all instances. For gene $i$, $C_i$ is the mean expression values of all instances. $C_{ix}$ is the centroid of the instances of category $x$ with respect to gene $i$. The feature vector about all instances for a given gene is replaced by $\overrightarrow{R_i}$ which records the relationships between the centroid of all instances and the centroids of different categories. The proposed vector provides information about the influence of a given gene for the overall shape of clusters. Gene $i$ and gene $j$ are compared

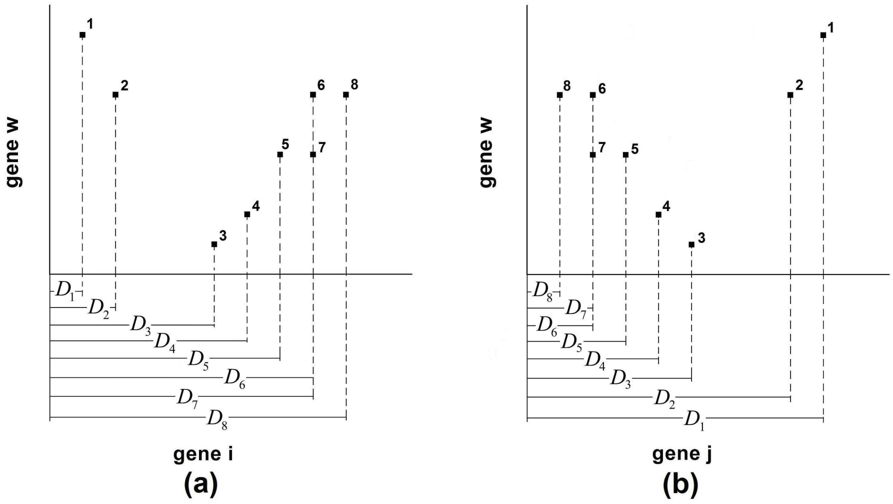

Fig. 2. Original feature vectors for gene $i$ and gene $j$

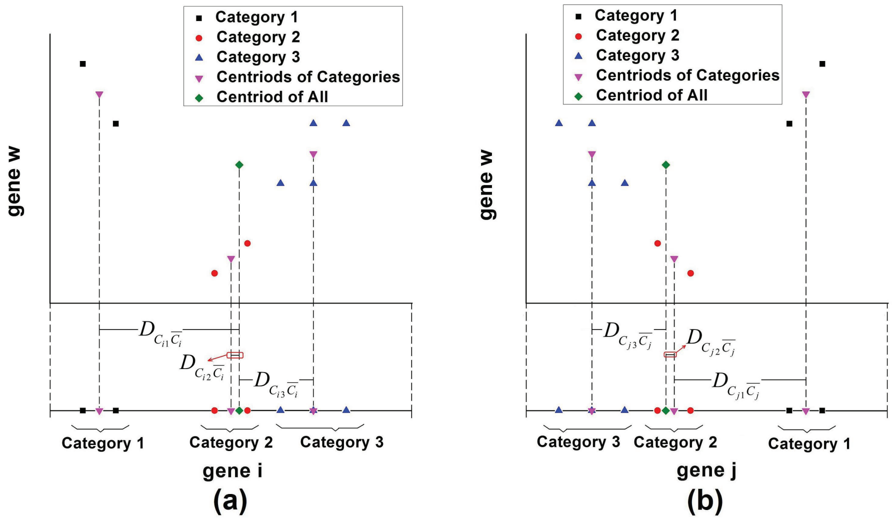

Fig. 3. Proposed feature vectors for gene $i$ and gene $j$

by the novel feature vectors, $\{|\frac{D_{C_{i1}\overline{C_i}}}{m_1 S_i}|, |\frac{D_{C_{i2}\overline{C_i}}}{m_2 S_i}|, |\frac{D_{C_{i3}\overline{C_i}}}{m_3 S_i}|\}$ and $\{|\frac{D_{C_{j1}\overline{C_j}}}{m_1 S_j}|, |\frac{D_{C_{j2}\overline{C_j}}}{m_2 S_j}|, |\frac{D_{C_{j3}\overline{C_j}}}{m_3 S_j}|\}$, instead of the instances' values in gene-based clustering algorithms. Finally, gene $i$ and gene $j$ are viewed similar according to two feature vectors.

From Fig.3, we can learn that two genes reflect the similar "discriminative capability". Gene $i$ and gene $j$ can both apparently discriminate category 1 against category 2 and category 3. However, because the information about their categories is covered up by the original feature vector in Fig.2, we cannot discover the "discriminative capability" of a certain gene. The "discriminative capability" is significant for incorporating to find groups of coregulated genes with strong association to the instances' categories.

### B. Generalized Fuzzy Algorithmic Scheme

Most of the fuzzy clustering algorithms are based on the minimization of the cost function shown below:

$$J_q(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{m} u_{ij}^q d(x_i, \theta_j) \tag{3}$$

$\theta_j$ is the parameterized representative of the $j$th cluster, $\theta \equiv [\theta_1^T, \ldots, \theta_m^T]^T$. $X_i$ is the expression pattern of the $i$th gene in the dataset. $d(x_i, \theta_j)$ is the dissimilarity between $x_i$ and $\theta_j$, $q$ is a real-valued number which controls the "fuzziness" of the resulting clusters. The fuzzy clustering algorithm links each gene to all clusters via a real-valued vector of indexes, $u_{ij}$, which lie between 0 and 1. The vector of indexes represents the relationships between a gene and different clusters.

$\theta$ and $U$ satisfies the constraint condition of Eq.4.

$$\sum_{j=1}^{m} u_{ij} = 1, i = 1, 2, \ldots, N \tag{4}$$

$$0 < \sum_{i=1}^{m} u_{ij} < N, j = 1, 2, \ldots, m \tag{5}$$

Estimates for $\theta$ and $U$ can be obtained by Alternating Optimization[17], [18]:

```
Generalized Fuzzy Algorithmic Scheme(GFAS)
• Choose $\theta_j(0)$ as initial estimates for $\theta_j$, $j$=1, $\ldots$,$m$.
• $t$=0
• Repeat
    – For $i$=1 to $N$
        ∗ For $j$=1 to $m$
          .
          $u_{ij}(t) = \frac{1}{\sum_{k=1}^{m} (\frac{d(x_i,\theta_j(t))}{d(x_i,\theta_k(t))})^{\frac{1}{q-1}}}$    (6)
        ∗ End For-j
    – End For-i
    – $t$=$t$+1
    – For $j$=1 to $m$
        ∗ Parameter updating:Solve
          $\sum_{i=1}^{N} u_{ij}^{q}(t-1) \frac{\partial d(x_i, \theta_j)}{\partial \theta_j} = 0$    (7)
          with respect to and set equal to this solution.
    – End For-j
• Until a termination criterion is met.
```

As the termination criterion we may employ $\|\theta(t) - \theta(t-1)\| < \varepsilon$, where $\| \bullet \|$ is any vector norm and $\varepsilon$ is a "small" user-defined constant.

Non-similarity measures are widely used in fuzzy clustering algorithms. When Euclidean distance is adopt, the resulting algorithm is known as Fuzzy C-Means (FCM) algorithm.

$$d_E(x_i, \theta_j) = (x_i - \theta_j)^T A (x_i - \theta_j). \tag{8}$$

where $A$ is a symmetric, positive definite matrix. We have

$$\frac{\partial d_E(x_i, \theta_j)}{\partial \theta_j} = 2A(\theta_j - x_i) \tag{9}$$

Substituting Eq.9 into Eq.7, we obtain:

$$\sum_{i=1}^{N} u_{ij}^{q}(t-1) 2A(\theta_j - x_i) = 0. \tag{10}$$

Since $A$ is positive definite, it is invertible. Premultiplying both sides of this equation with $A^{-1}$ and after some simple algebra, we obtain:

$$\theta_j(t) = \frac{\sum_{i=1}^{N} u_{ij}^{q}(t-1) x_i}{\sum_{i=1}^{N} u_{ij}^{q}(t-1)}. \tag{11}$$

TABLE I

GENE EXPRESSION DATA SETS

| Dataset | Genes | Classes | Train-samples | Test-samples |
|---|---|---|---|---|
| Leuk1 | 7129 | 3 | 38 | 34 |
| Leuk2 | 12582 | 3 | 57 | 15 |
| Leuk3 | 12558 | 7 | 215 | 112 |
| Lung1 | 7129 | 3 | 64 | 32 |
| Lung2 | 12600 | 5 | 136 | 67 |
| Breast | 9216 | 5 | 54 | 30 |
| SRBCT | 2308 | 4 | 63 | 20 |
| DLBCL | 4026 | 6 | 58 | 30 |
| Cancers | 12533 | 11 | 100 | 74 |
| GCM | 16063 | 14 | 144 | 46 |

Non-similarity measures are proper for fuzzy clustering algorithms because: 1)the result of $\frac{\partial d(x_i, \theta_j)}{\partial \theta_j}$ is easier to get the solution of $\theta$ from Eq.7 than similarity measures, 2)similarity measures and non-similarity measures have opposite optimization aims. Hence, the measures with categories information, such as mutual information, cannot directly be used in generalized fuzzy algorithmic scheme. However, by introducing the proposed feature vector for each gene, the instance category information in train data is integrated into the non-similarity measure and this vector can be directly employed in fuzzy clustering algorithm. The solution of $\theta$ from Eq.7 can be attained easily according to generalized fuzzy algorithmic scheme.

## III. EXPERIMENT

### A. Microarray Data Sets

In this paper, we present the experimental results of 10 multi-category problems, all of which have independent test sets. Three data sets about leukemia cancer(Leuk1, Leuk2, and Leuk3), two data sets about lung cancer(Lung1 and Lung2), and one data set about breast cancer (Breast) are involved. SRBCT is a data set about small, round blue-cell tumors. DLBCL is a data set about diffuse large B-cell lymphoma. Cancers and GCM comprise several kinds of human tumors. The detailed description of the data sets is shown in Table 2. The original description of the data sets can be found in the related works[19], [20]. In the experiment, we simply label the genes of each problem from 0 to the largest number as gene 0, gene 1, ..., instead of using the original biological gene labels. In Table I, "Genes" indicates the number of genes. The term "Classes" indicates the number of classes. "Train-samples" indicates how many samples the training data sets have. "Test-samples" indicates how many samples are involved in the independent test data sets.

### B. Clustering Validity Functions

FCM is employed in this paper as a representative of fuzzy clustering algorithm. Two other clustering algorithms, $K$-means and Farthest First(FF), were used to cluster genes with conventional feature vector and the proposed feature vector. Because of having no prior knowledge about the number of clusters and genes' labels, the proper clustering results were attained by heuristics experiments. And the quality of clustering results were evaluated by the clustering validity

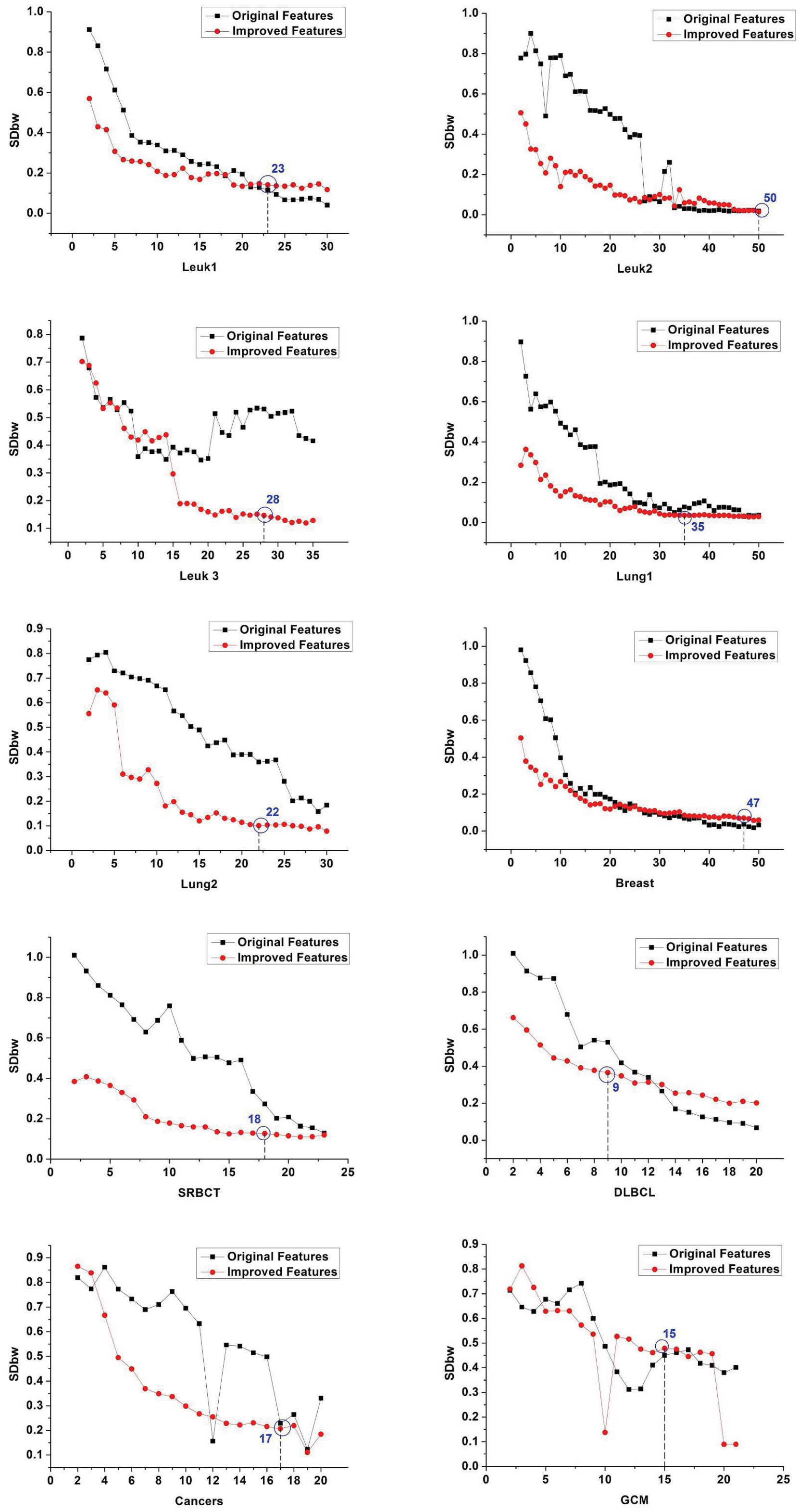

Fig. 4. $K$-means Clustering Results with Different $K$

function. SDbw[21] was employed. SDbw is proper for the crisp clustering and it is a relative criterion where the algorithm is running repetitively using different input values and the resulting clusters are compared as for their validity[21]. It has the character: the smaller value corresponds to the better clustering result. The smaller value means more compact

TABLE II

CLUSTERING RESULTS BY DIFFERENT ALGORITHMS ON MICROARRAY DATA

| | $K$-means | | FF | | FCM | |
|---|---|---|---|---|---|---|
| | Original | Proposed | Original | Proposed | Original | Proposed |
| Leukemia1(23) | **0.116** | **0.14091** | **0.07286** | **0.08716** | 0.0493 | 0.02494 |
| Leukemia2(50) | 0.01843 | 0.01438 | **0.0315** | **0.03217** | 0.02081 | 0.01112 |
| Leukemia3(28) | 0.53042 | 0.14657 | **0.07449** | **0.12254** | 0.0629 | 0.02038 |
| Lung1(35) | 0.07684 | 0.03574 | 0.04759 | 0.04596 | 0.55359 | 0.01626 |
| Lung2(22) | 0.3592 | 0.10064 | 0.0841 | 0.07427 | 0.09646 | 0.02756 |
| Breast(47) | **0.03587** | **0.07038** | 0.03528 | 0.02721 | 0.37434 | 0.00914 |
| SRBCT(18) | 0.27317 | 0.12674 | 0.11113 | 0.09547 | 0.06049 | 0.04159 |
| DLBCL(9) | 0.52929 | 0.36484 | **0.21251** | **0.21293** | 0.12472 | 0.09 |
| Cancers(17) | 0.22862 | 0.2068 | **0.09227** | **0.09271** | **0.05009** | **0.07264** |
| GCM(15) | **0.45072** | **0.47838** | **0.08467** | **0.10768** | **0.05701** | **0.0788** |

intra-class distribution and more separated inter-class distribution. All reported results based on 10-fold cross-validation for each classification task.

### C. Proper Cluster Number

In application, the cluster number can be determined experimentally. For example, to $K$-means, we assume $K' = 2$ ($K'$ is the gene cluster). All genes are clustered into $K'$ clusters and the clustering result is evaluated by SDbw. Then, $K'$ increases by 1, the clustering and validating repeats. When the validating results keep steady within the scope of $10\%$ within 5 continuous attempts of different $K'$, the last value is chosen as the proper $K'$. In Fig.4, the clustering results with original feature vector and proposed feature vector are shown. The proper $K$ are marked.

From Fig.4, we can learn that, on the proper $K$, the clustering results with proposed features are better than those with original feature vector except for Leukemia1, Breast and GCM. With the increasing of $K$, the clustering results with proposed feature vectors are much quicker and easier to attain steady trend. Overall, the procedures tend towards stability without significant fluctuations except for GCM. Specially, the proper $K$ is not the best point in all data sets from Fig.4 and the clustering results with proposed feature vectors are even worse than original ones in some data sets, such as Breast and DLBCL. But we want to propose a method which can converge stably and quickly rather than methods which get the best result totally depending on experiments.

### D. Comparative Analysis

According to the results by $K$-means, the initial cluster centers of the FCM are attained. The clustering results by $K$-means, Farthest First and FCM with original and proposed feature vector are listed in Table II.

To $K$-means, the clustering results with proposed feature vectors are better than those with original feature vector except for Leukemia1, Breast and GCM. To Farthest First, the clustering results with proposed feature vector are not good as those with original feature vector except for 4 data sets. However, the differences between results with different feature vectors are less than 2% except for Leukemia3 and GCM. To FCM, the clustering results with proposed feature vector are better than those with original feature vector only except for Cancers and GCM. Because FCM uses centers attained by $K$-means, it can be viewed as an optimization for results by $K$-means. From results with original feature vector, FCM optimizes $K$-means results except for Leukemia2, Lung1 and Breast. From results with proposed feature vector, we can learn that FCM can attain more compact intra-class distribution and more separated inter-class distribution than other two clustering algorithms.

To make further analysis of clustering validity, we introduce some classifiers. Gene subsets are selected from clusters attained by different algorithms with original feature vector and proposed feature vector. Genes in each cluster are sorted by T-test[22] and the top 5 genes in each cluster are selected and integrated into the subset. The performances of Naive Bayes(NB), J48 and SMO(a variant of SVM[23]) are compared based on these subsets in Table III. Viewed from the classifier point, SMO is better than other classifiers in most cases. View from the clustering algorithms, the best results for each data can be attained by combining some classifier with FCM. Through comparing the performances based on original feature vector and proposed feature vector, we learn that gene subsets from different clustering algorithms can attain same accuracy or better accuracy on three classifiers except for the number pairs in bold type. As shown in Table III, there are 7 cases in NB, 9 cases in J48 and 4 cases in SMO. SMO performed more stably and accurately with proposed feature vector than other two classifiers.

### E. Discussion

In application, the cluster number also can be determined by other clustering algorithms, such as subtractive clustering. However, the convergence speed of the subtractive clustering is too quick because the changes between results with close $k$ values are tiny. This phenomena is led by the characteristics of substractive clustering. Hence, $K$-means was employed in this paper. In addition, the experiment is mainly focused on microarray data sets in this paper and other data sets have not been applied. The future works will be developed by us in general data and fields.

## IV. CONCLUSION

In this paper, a novel feature vector for each gene is employed in fuzzy clustering algorithm. The proposed feature vector can provide information about the influence of a given gene for the overall shape of clusters. By analysis and experiment upon microarray data sets, the performance of

TABLE III

PERFORMANCES OF GENE SUBSET BASED ON DIFFERENT CLUSTERING ALGORITHMS

| | | Original | | | Proposed | | |
|---|---|---|---|---|---|---|---|
| | | NB | J48 | SMO | NB | J48 | SMO |
| Leukemia1 | $K$-means | 73.53% | 73.53% | 70.59% | 79.41% | 73.53% | 76.47% |
| | FF | 29.41% | 5.88% | 64.71% | 85.29% | 73.53% | 85.29% |
| | FCM | 73.53% | 73.53% | 73.53% | 73.53% | 73.53% | 76.47% |
| Leukemia2 | $K$-means | 100% | 100% | 100% | 100% | 100% | 100% |
| | FF | 93.33% | **100%** | 93.33% | 100% | **93.33%** | 100% |
| | FCM | 93.33% | 100% | 100% | 100% | 100% | 100% |
| Leukemia3 | $K$-means | 49.10% | **76.79%** | 83.04% | 56.25% | **67.86%** | 83.04% |
| | FF | 50% | 60.71% | 86.61% | 66.07% | 66.96% | 92.86% |
| | FCM | 37.5% | 66.96% | 85.71% | 60.71% | 67.86% | 86.61% |
| Lung1 | $K$-means | 43.75% | 56.25% | 84.38% | 84.38% | 81.25% | 87.5% |
| | FF | 71.88% | 28.13% | 81.25% | 78.13% | 62.5% | 81.25% |
| | FCM | 53.13% | 28.13% | 84.38% | 81.25% | 37.5% | 84.38% |
| Lung2 | $K$-means | **94.03%** | 74.63% | **97.01%** | **86.57%** | 82.09% | **95.52%** |
| | FF | **94.03%** | **85.07%** | 95.52% | **89.55%** | **80.60%** | 97.01% |
| | FCM | **97.01%** | 76.12% | 92.54% | **92.54%** | 79.10% | 97.01% |
| Breast | $K$-means | 50% | 50% | **70%** | 63.33% | 56.67% | **66.67%** |
| | FF | 56.67% | **73.33%** | 80% | 66.67% | **60%** | 80% |
| | FCM | 53.33% | 53.33% | **76.67%** | 60% | 60% | **70%** |
| SRBCT | $K$-means | **100%** | 35% | 95% | **65%** | 35% | 100% |
| | FF | 90% | **40%** | 95% | 90% | **35%** | 95% |
| | FCM | 90% | 35% | **100%** | 90% | 35% | **95%** |
| DLBCL | $K$-means | 86.67% | 76.67% | 93.33% | 86.67% | 76.67% | 93.33% |
| | FF | 83.33% | **73.33%** | 83.33% | 90% | **66.67%** | 93.33% |
| | FCM | **93.33%** | **80%** | 93.33% | **90%** | **73.33%** | 96.67% |
| Cancers | $K$-means | 43.24% | **37.84%** | 54.05% | 47.30% | **24.32%** | 56.76% |
| | FF | **54.05%** | 24.32% | 59.46% | **51.35%** | 24.32% | 75.68% |
| | FCM | 27.03% | 22.97% | 51.35% | 50% | 22.97% | 58.11% |
| GCM | $K$-means | 26.09% | 30.43% | 43.48% | 34.78% | 36.96% | 50% |
| | FF | **32.61%** | **36.96%** | 39.13% | **26.09%** | **34.78%** | 54.35% |
| | FCM | 21.74% | 34.78% | 36.96% | 32.61% | 34.78% | 54.35% |

the fuzzy clustering algorithm based on the proposed feature vector is compared with that of some classical clustering algorithms. The results demonstrate that the fuzzy clustering algorithm based on the proposed feature vector is capable of obtaining better clusters than other contrast algorithms.

## REFERENCES

[1] D. Jiang, C. Tang, A. Zhang,"Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370-1386, 2004.

[2] A. Ben-Dor, R. Shamir, Z. Yakhini,"Clustering gene expression patterns," *Journal of computational biology*, vol. 6, pp. 281-297, 1999.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, et al,"Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863-14868, 1998.

[4] J. Herrero, A. Valencia, J. Dopazo,"A hierarchical unsupervised growing neural network for clustering gene expression patterns," *Bioinformatics*, vol. 17, no. 2, pp. 126-136, 2001.

[5] H. Wang, H. Zheng, F. Azuaje,"Poisson-based self-organizing feature maps and hierarchical clustering for serial analysis of gene expression data," *Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 163-175, 2007.

[6] L. J. Heyer, S. Kruglyak, S. Yooseph,"Exploring expression data: identification and analysis of coexpressed genes," *Genome research*, vol. 9, no. 11, pp. 1106-1115, 1999.

[7] P. Maji,"Mutual information-based supervised attribute clustering for microarray sample classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 127-140, 2012.

[8] D. Jiang, J. Pei, A. Zhang,"Interactive exploration of coherent patterns in time-series gene expression data," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 565-570, 2003.

[9] D. Jiang, J. Pei, A. Zhang,"DHC: a density-based hierarchical clustering method for time series gene expression data," *Bioinformatics and Bioengineering*, pp. 393-400, 2003.

[10] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Kluwer Academic Publishers, 1981.

[11] D. Dembele, P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973-980, 2003.

[12] J. Wang, et al, "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data," *BMC bioinformatics*, vol. 4, no. 1, pp. 60, 2003.

[13] L. Tari, C. Baral, S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *Journal of biomedical informatics*, vol. 42, no. 1, pp. 74-81, 2009.

[14] A. P. Gasch, M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, no. 11, pp. 1-22, 2002.

[15] S. Theodoridis, K. Koutroumbas, *Pattern recognition*, House of Electronics Industry, 2010.

[16] M. Dettling, P. Buhlmann, "Supervised clustering of genes," *Genome biology*, vol. 3, no. 12, pp. 1-0069.15, 2002.

[17] J. C. Bezdek, R. J. Hathaway, N. R. Pal, "Norm-induced shell-prototypes (NISP) clustering,", pp. 431-449, 1995.

[18] F. Hoppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis*, John Wiley and Sons, 1999.

[19] D. G. Beer, S. L. R. Kardia, C. C. Huang, et al, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816-824, 2002.

[20] S. A. Armstrong, J. E. Staunton, L. B. Silverman, et al, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature genetics*, vol. 30, no. 1, pp. 41-47, 2001.

[21] M. Halkidi, M. Vazirgiannis, Y. Batistakis, "Quality scheme assessment in the clustering process," *Principles of Data Mining and Knowledge Discovery*, pp. 265-276, Springer Berlin Heidelberg, 2000.

[22] C. Ding, H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 2, pp. 185-205, 2005.

[23] C. W. Hsu, C. C. Chang, C. J. Lin, *A practical guide to support vector classification*, 2003.