Fuzzy Clustering Algorithm of Kernel for Gene Expression Data Analysis

Wenyuan Liu, and Bin Zhang

Abstract—Many clustering techniques have been proposed for the analysis of gene expression data obtained from microarray experiments. However, choice of suitable method(s) for a given experimental dataset is not straightforward. KFCM algorithm has been widely applied in gene expression data analysis, but it is sensitive to the class center migration. Therefore, the fuzzy kernel clustering algorithm based on the kernel function and a dynamic weighted is proposed. The algorithm for the kernel function marking the introduction of the effectiveness function, missing data be filtered from the data sets and the original data is mapped to the high-dimensional feature space in each category of each gene and the right to assign a dynamic value, so that in the clustering algorithm iterative process of adaptive adjustment of weights, after analysis of the value of the size of the right to identify outliers, finally, getting the more ideal clustering result. Experiments showed that the proposed methods achieve better clustering effect than the fuzzy kernel clustering algorithm and possibility clustering algorithm, and is fast in convergence

Index Terms—microarray, clustering, effective function, outliers, kernel function

I. INTRODUCTION

Microarray technology is the field of life sciences in recent years developed a high-tech, for molecular biology, bio-medicine research provides strong support. Gene chip technology application process produce large amounts of gene expression data, how to effectively process and analyze these data and extract valuable biological information has become the biology and information science a hot issue [1].

Currently, many clustering algorithms are widely used in gene expression data [2] analysis, for example: K-means algorithm, self-organizing map (SOM), fuzzy C-means clustering (FCM) [3] algorithms, Kernel fuzzy C-means clustering (KFCM) [4] algorithm. Which KFCM cluster analysis algorithm is known for its deal with large volumes of data, the characteristics of high accuracy are widely used in gene expression data analysis, KFCM algorithm can only use cluster center to represent each class, if the data set there is missing data and outliers, then the center of each class will occur during the cluster analysis of large offset, the impact cluster analysis results. In order to resolve this problem, a new kernel function [5] and dynamic weighted [6] Kernel Fuzzy C Means clustering method (KW-KFCM) is proposed in this paper. In KW-KFCM, each gene is assigned effective function.

KW-KFCM algorithm is extensions of kernel fuzzy C-means algorithm with the data in the original space are mapped to a high-dimensional feature space, and an additional effective function is assigned to each vector in the feature space. The noise and outliers could find through analysis effective function of a data, then, clusters can be identified.

II. GENE EXPRESSION DATA

The microarray technology enables to analyze the expression of thousands of genes in a single experiment and provides quantitative measurements of the differential expression of these genes. The raw microarray data images are transformed into gene expression matrix $A = (a_{ij})_{m 	imes n}$ [8].

The rows $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ in the matrix correspond to genes, and the columns $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})^T$ represent samples or experimental conditions, each element $a_{ij}$ expresses the gene i in the experimental condition (sample) the j expression value. By the microarray chip's experimental principle, $a_{ij}$ takes for relative fluorescence intensity ratio as follows:

$$a_{ij} = \log_2 \frac{I_j}{I_i}$$

(1)

From the formula (1), $I_j$ is on the microarray chip the sample group gene (red fluorescent agent) intensity, $I_i$ is on the microarray chip the control group gene (green fluorescent agent) intensity.

III. KERNEL FUZZY C-MEANS CLUSTERING ALGORITHM

A. Mercer kernel function

Mercer kernel function [7] The principle is to sample data space for non-linear mapping to a new high-dimensional feature space $H : \phi : X^N \rightarrow H, x \rightarrow \phi(x)$, availability of feature space $H$ in a group of vector set : $\phi(x_1),\ldots,\phi(x_n)$. According to the nature of the mapping $\phi$, the input space dot product in feature space $H$ can be expressed with the Mercer kernel functions:

$$k(x_i,x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

(2)

where $k(x_i,x_j)$ is the kernel function, from formula (2) Analysis of input samples on the formation of a nuclear function matrix $K_{ij} = k(x_i,x_j)$.
thus, \( K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \), and four widely used basic kernel functions are as follows:

1. Linear: \( K(x_i, x_j) = x_i^T x_j \)
2. Polynomial: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \)
3. Sigmoid: \( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \)
4. Radial basis: \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \)

where \( r, \gamma \) and \( d \) are kernel parameters. Since, 

\[
d_m(x, y) = \sqrt{\phi(x) - \phi(y)} = \sqrt{[\phi(x) - \phi(y)]^2} = \sqrt{[\phi(x) - 2\phi(x) \phi(y) + \phi(y)] \phi(y)} \tag{3}\]

when the kernel function is chosen as radial basis function, \( K(x_i, x_j) = 1 \), \( K(v_i, v_j) = 1 \), then 

\[
\|\phi(x_i) - \phi(v_j)\|^2 = 2(1 - K(x_i, v_j)) \tag{4}\]

B. KFCM algorithm

The KFCM algorithm modifies the objective function of FCM to 

\[
J_m(U, V) = \sum_{k=1}^{c} \sum_{n=1}^{N} u_{kn}^{m} \left[\phi(x_n) - \phi(v_k)\right]^2 \tag{5}\]

The parameters of KFCM are estimated according to the following formulas:

\[
u_n^{m} = \frac{(1 - K(x_n, v_k))^{2m-1}}{\sum_{k=1}^{c} (1 - K(x_n, v_k))^{2m-1}} \tag{6}\]

\[
v_k = \frac{\sum_{n=1}^{N} u_{kn}^{m} K(x_n, v_k) x_n}{\sum_{n=1}^{N} u_{kn}^{m} K(x_n, v_k)} \tag{7}\]

when \( J_m(U, V) - J_{m+1}(U, V) \leq \varepsilon \), the iteration stopped and the final matrix \( U \) and \( V \) are obtained.

C. KFCM Algorithm steps

The algorithm steps of modified KFCM algorithm are as follows:

1) Initialize the cluster number \( c \), the exponential weight \( m \), the terminate number \( c \) and the max iteration times \( t_{\text{max}} \), where \( 2 \leq c \leq n \). Initialize \( U^{(0)} \).

2) Compute the kernel function of \( k(x, y) \) with formula (10).

3) At step \( i = 1 \), calculate the cluster center \( v^{(i)} \) and \( U^{(i)} \) with formula (6) and (7).

4) For step \( i = i + 1 \), update \( v^{(i+1)} \) \( U^{(i+1)} \).

5) If \( J_m(U, V) - J_{m+1}(U, V) \leq \varepsilon \) or \( i > t_{\text{max}} \), stop; Else, go to step 4.

IV. PROPOSED ALGORITHM FOR CLUSTERING

As a result of KFCM algorithm is sensitive to noise and outliers. Therefore, the paper presents a new approach to a kernel fuzzy clustering method of gene expression data.

A. Filter the flaw genes

At present a large-scale gene expression data are derived from the hybrid experiment, a variety of analysis techniques can be a one-time expression of hundreds of thousands of genes for testing. From the data pre-processing to the final process of information collection in this series of experiments the inevitable existence of various internal and external factors, resulting in missing data (expression values do not conform to quality standards). The analyses of these data follow-up methods have a great impact, even fatal. In order to solve such problems, proposed a new Gaussian kernel function, for the validity of each gene is assigned an identification function:

\[
H_{A_i} = \begin{cases} 1 & \text{the ith sample meets the requirement} \\ 0 & \text{the ith sample flaw or not meet standards} \end{cases} \tag{8}\]

then, the radial basis function formula change is:

\[
k'(x, y) = \exp \left( \frac{-1}{2\beta^2} \left[ \sum_{i=1}^{M} Z(x, y)(x_i - y_i) \right]^2 \right) \tag{9}\]

\[
Z(x, y) = \sum_{i=1}^{M} H_{x_i} H_{y_i} \tag{10}\]

Therefore, from the formula (3), high dimensionality space \( H \) distance function can by change is:

\[
d_m(x, y) = \sqrt{k(x, x) - 2k(x, y) + k(y, y)} \tag{11}\]

Improved radial basis function through analyzes each gene the validity, removes the flaw data from the analysis, therefore, raised the kernel function operation efficiency.

B. Identify the outlier genes

In general, the outlier genes are concentrated sample of some genetic similarities with other genes from very different data, outlier genes are likely to contain very important information, for such outlier genes should focus on analysis and research. However, outlier genes likely to be in the process of data collection errors in operations caused by the genes to get rid of this kind of outlier. The traditional approach often ignored all of the outlier genes, resulting in the loss of important information. Therefore, profits from reference [8] each gene to assign a dynamic weight to the feature space, through to weight analysis discovery stray gene method. However, the method did not overcome the traditional fuzzy kernel clustering is not suitable for complex data structure defects, that is very different in a variety of data classes of cases, often leads to a small data class be mistaken for large data sub-category errors or the consequences of the annexation of . In view of the above question, proposed in the feature space again
for each kind of dynamic assignment weight, the improvement difference very big data class cluster effect.

Improved kernel fuzzy C-means objective function as follows:

\[
J_{im}(U,V) = \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{qij}^{m} a_{i}^{p} \frac{1}{w_{j}^{q}} d_{ij}^{2} (x_{j} - v_{i})
\]  

(12)

where \( d_{ij}^{2} (x_{j} - v_{i}) \) is the use of formula (11) calculate the feature space the first \( j \) genes in the sample to the first \( i \) class center distance. The formula (12) is the KW-KFCM algorithm optimization objective function.

Outlier information is often the most important characteristics of data sets, therefore, \( w_{j} \) is the first \( j \) genes in samples of the dynamic distribution of weights. When assigns small weight \( w_{j} \) to the ordinary gene sample, then \( 1/w_{j} \) is big; when assigns big weight \( w_{j} \) to the stray gene sample, then \( 1/w_{j} \) is small. Parameter \( q \in (0, \infty) \), when \( q \to 0 \), the weight value will influence achieve in a big way. \( w_{j} \) iterative formula through under formula computation:

\[
w_{j} \leftarrow w_{j} \frac{\left( \sum_{i=1}^{C} \mu_{qij}^{m} a_{i}^{p} d_{ij}^{2} (x_{j} - v_{i}) \right)^{1/q}}{\sum_{i=1}^{C} \left( \sum_{j=1}^{N} \mu_{qij}^{m} a_{i}^{p} d_{ij}^{2} (x_{j} - v_{i}) \right)^{1/q}}
\]  

(13)

where \( \sum_{j=1}^{N} w_{j} = w \), \( w \) is the user definition constant.

\( a_{i} \) is the \( i \) th kind of dynamic allocation weight, indicates this kind in the parsing process importance. When the iterative process samples that contain more genes or gene samples of dense cluster, its importance degree greater distribution of a class should have a bigger; when those composed of elements in the class or less than the sparse distribution of elements in the class has a smaller Importance. Parameter \( q \in (0, \infty) \), when \( q \to 0 \), the weight value will influence achieve in a big way. \( a_{i} \) iterative formula through under formula computation:

\[
a_{i} \leftarrow \frac{\left( \sum_{j=1}^{N} \mu_{qij}^{m} a_{i}^{p} d_{ij}^{2} (x_{j} - v_{i}) \right)^{1/p}}{\sum_{i=1}^{C} \left( \sum_{j=1}^{N} \mu_{qij}^{m} a_{i}^{p} d_{ij}^{2} (x_{j} - v_{i}) \right)^{1/p}}
\]  

(14)

where \( \sum_{i=1}^{C} a_{i} = 1 \).

C. Algorithm steps

The algorithm steps of modified KW-KFCM algorithm are as follows:

1) Initialize the cluster number \( c \), the exponential weight \( m \), data weighted index of \( q \), class weight index \( p \), the terminate number \( c \) and the max iteration times \( t_{max} \), where \( 2 \leq c \leq n \). Initialize \( U^{(0)} \).

2) Compute the kernel function of \( k(x,y) \) with formula (10).

3) At step \( i=1 \), calculate the cluster center \( v^{(i)} \) and \( U^{(i)} \) with formula (6) and (7).

4) For step \( i=i+1 \), update \( v^{(i+1)}, U^{(i+1)} \).

5) Compute each sample weighting factor \( w_{j}^{(i)} \) with formula (13).

6) Compute each kind of weighting factor \( a_{j}^{(i)} \) with formula (14).

7) If \( \| U^{(i)} - U^{(i+1)} \| \leq \epsilon \) or \( i > t_{max} \), stop; Else, go to step 4.

V. DATABASES AND EXPERIMENTAL RESULTS

In this article, in order to confirm the KW-KFCM algorithm the validity and the feasibility, uses the yeast gene expression data to carry on the cluster analysis, and carries on the comparison with the commonly used cluster algorithm. The experiments are performed on MatLab7.5 platform.

A. Databases

We used one database in our experiments. The data set is standard Yeast data, which can be downloaded from the Standford University biology department website (http://smd.stanford.edu). As the relatively large experimental data, from which 428 are known to select the function of genes and cells, the strongest cyclical clustering, each gene 27-dimensional, representing the different experimental conditions and time point combination. There are 428 data samples with five phases (G1 (171), S (69), S/G2 (17), G2/M (95), M/G1 (76)) in the cell cycle.

B. Experimental Results

Compare the recognition performance of FCM, KFCM, and KW-KFCM on Yeast data. The cluster number \( c=5 \). The exponential weight \( m=2 \). The terminate number \( \epsilon =0.001 \). The max iteration times \( t_{max} =100 \). The experimental results are listed from table 1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>G1</th>
<th>S</th>
<th>S/G2</th>
<th>G2/M</th>
<th>M/G1</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW-KFCM</td>
<td>93.22</td>
<td>77.24</td>
<td>68.47</td>
<td>86.07</td>
<td>74.11</td>
<td>84.68</td>
</tr>
<tr>
<td>KFCM</td>
<td>90.56</td>
<td>66.91</td>
<td>20.38</td>
<td>79.15</td>
<td>62.51</td>
<td>76.45</td>
</tr>
<tr>
<td>FCM</td>
<td>80.89</td>
<td>65.83</td>
<td>15.49</td>
<td>63.34</td>
<td>60.22</td>
<td>68.30</td>
</tr>
</tbody>
</table>

From Table 1, we can conclude that KW-KFCM has better performance than FCM or KFCM, and the performances of FCM and KFCM are very low accuracy for the S/G2. The reason is because some classes have more genes swallow up the 17 genes belonging to the S/G2, lead to the S/G2 is not gathered to one cluster well. But KW-KFCM is assigns for each gene the effectiveness of function, avoided the small data cluster by the large data cluster swallow, therefore, the S/G2 cluster accuracy
had the remarkable enhancement. KW-KFCM algorithm total average accuracy achieves 84.68%. This accuracy was considered very high in the biology research.

In addition, compare the recognition performance of KW-KFCM and commonly used cluster algorithm. The cluster number \( c=8 \) so we used K-means (\( k=8 \)) and SOM (\( 4 \times 2 \) neurons) cluster the above yeast expression data. The experimental results are shown from Fig. 2.

![Graph](image)

**Figure 1.** Comparison of the results of KW-KFCM, K-means and SOM.

From the Fig. 2, we can conclude that KW-KFCM has better performance than FCM or KFCM, and the performances of K-means and SOM are smaller Radius and lower inner cluster deviation. Therefore, the experiment results show that KW-KFCM algorithm is more effective.

**VI. CONCLUSION**

In this paper, Through carries on the analysis to the KFCM algorithm's nuclear function and the iterative process, has improved the algorithm operation process from computational method's angle, and introduces the valid marking function and the dynamic weighting method, reduced the flaw data and the stray gene to the cluster effect influence. KW-KFCM algorithms does not limit applies in the gene expression data analysis, it may also expand applies in the biological information sciences other research direction. Regarding the different biological information sciences question, may solve the gene expression data cluster analysis model based on this article to carry on the corresponding change, in which dynamic weighting thought may use for reference or the quotation. In the later research, how according to a data feature selection more appropriate nuclear function as well as the determination initial cluster integer will be the key point which further studies.

**ACKNOWLEDGMENT**

The authors wish to thank the National Natural Science Foundation of China (Grant Nos.60773100) and the National Information Generic Technology Service Platform of Torch Plan based on SAAS (Grant Nos.2008 GH540088) under which the present work was possible.

**REFERENCES**


