Alternative Fuzzy C-Means Clustering for DNA Computing Readout Method Implemented on DNA Engine Opticon 2 System

Muhammad Faiz Mohamed Saaid[1], Zuwairie Ibrahim[1], Marzuki Khalid[1], and Azli Yahya[2]
[1]*Department of Mechatronic and Robotics,*
Center for Artificial Intelligence and Robotics (CAIRO),
Faculty of Electrical Engineering,
Universiti Teknologi Malaysia,
Johor, Malaysia
ninjagaban@yahoo.com, zuwairiee@fke.utm.my, marzuki@utmkl.utm.my

[2]*Microelectronic and Computer Engineering Department,*
Faculty of Electrical Engineering
Universiti Teknologi Malaysia,
Johor, Malaysia
azli@fke.utm.my

Abstract

In the previous paper, a readout approach for the Hamiltonian Path Problem (HPP) in DNA computing based on the real-time polymerase chain reaction (PCR) was proposed. Based on this approach, real-time amplification was performed with the TaqMan probes and the TaqMan detection mechanism was exploited for the design and development of the readout approach. The readout approach consists of two steps: real-time amplification in vitro using TaqMan-based real-time PCR, followed by information processing in silico to assess the results of real-time amplification, which in turn, enables extraction of the Hamiltonian path. However, the previous method used manual classification of two different output reactions of real-time PCR. In this paper, Alternative Fuzzy C-Means (AFCM) clustering algorithm is used to identify automatically two different reactions in real-time PCR. We show that AFCM clustering technique can be implemented for clustering output results of DNA computing readout method based on DNA Engine Opticon 2 System.

1. Introduction

The innovation of real-time PCR has rapidly gained popularity and plays a crucial role in molecular medicine and clinical diagnostics [1]. All real-time amplification instruments require a fluorescence reporter molecule for detection and quantitation, whose signal increase is proportional to the amount of amplified product. It has been found that the mechanism of the TaqMan hydrolysis probe shown in Figure 1 [2] is very suitable for the design and development of a readout method for DNA computing.

Previously, we have proposed a readout method tailored specifically to the HPP in DNA computing, which employs a hybrid *in vitro-in silico* approach [3]. In the *in vitro* phase, $O(|V|^2)$ TaqMan-based real-time PCR reactions are performed in parallel, to investigate the ordering of pairs of nodes in the Hamiltonian path of a $|V|$-node instance graph, in terms of relative distance from the DNA sequence encoding the known start node. The resulting relative orderings are then processed *in silico*, which efficiently returns the complete Hamiltonian path. The proposed approach is experimentally validated optical method specifically designed for the quick readout of HPP instances, in DNA computing.

As shown in Figure 2, the output of DNA computing readout method implemented on DNA Engine Opticon 2 System consist of two kinds of reactions, namely "YES" reaction and "NO" reaction. In this paper, we implement AFCM clustering algorithm for automatic classification of "YES" and "NO" reaction. We also implement the conventional FCM clustering algorithm. The results show that the conventional FCM clustering algorithm implemented on output of DNA Engine Opticon 2 system give two misclassifications of "YES" and "NO" reaction, due to the problem of outliers or noises. However, the misclassifications are overcome by implementing the AFCM clustering algorithm. We show that the "YES" and "NO" reactions can be separated for better *in silico* information processing

978-0-7695-3493-0/08 $25.00 © 2008 IEEE
DOI 10.1109/SITIS.2008.51

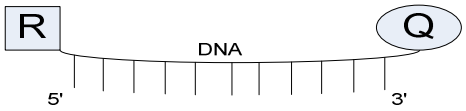

Figure 1. Illustration of the structure of a TaqMan DNA probe. Here, R and Q denote the reporter and quencher fluorophores, respectively

2. Readout approach of the DNA computing based on real-time PCR

First of all, $v_1v_2v_3v_4$ denotes a double-stranded DNA (dsDNA) which contains the base-pairs subsequences, v_1, v_2, v_3, and v_4 respectively. A reaction denoted by TaqMan(v_0,v_k,v_l) indicates that real-time PCR is performed using forward primer v_0, reverse primer v_l , and TaqMan probe v_k . Based on the proposed approach, there are two possible reaction conditions regarding the relative locations of the TaqMan probe and reverse primer. In particular, the first condition occurs when the TaqMan probe specifically hybridizes to the template, between the forward and reverse primers, while the second condition occurs when the reverse primer hybridizes between the forward primer and the TaqMan probe. Thus, we define TaqMan(v_0,v_k,v_l) = YES if the first condition occurred, while TaqMan(v_0,v_k,v_l) = NO if the second condition occurred.

Let the output of an *in vitro* computation of an HPP instance of the input graph be represented by a 140 base-pairs dsDNA(each node represents 20 base-pairs) $v_0v_1v_4v_2v_5v_3v_6$, where the Hamiltonian path V_0•V_1•V_4•V_2•V_5•V_3•V_6, begins at node V_0, ends at node V_6, and contains intermediate nodes V_1, V_4, V_2, V_5, and V_3 respectively. Note that in practice, only the identities of the starting and ending nodes, and the presence of all intermediate nodes will be known in advance to characterize a solving path. The specific order of the intermediate nodes within such a path is unknown.

The first part of the approach, which is performed *in vitro*, consists of $[(|V|-2)^2-(|V|-2)]/2$ real-time PCR reactions, each denoted by TaqMan(v_0,v_k,v_l) for all k and l, such that $0 < k < |V|-2$, $1 < l < |V|-1$, and $k < l$. For this example instance, so that the DNA template is dsDNA $v_0v_1v_4v_2v_5v_3v_6$, these 10 reactions are as follows:

(1) TaqMan(v_0,v_1,v_2) = YES
(2) TaqMan(v_0,v_1,v_3) = YES
(3) TaqMan(v_0,v_1,v_4) = YES
(4) TaqMan(v_0,v_1,v_5) = YES
(5) TaqMan(v_0,v_2,v_3) = YES
(6) TaqMan(v_0,v_2,v_4) = NO
(7) TaqMan(v_0,v_2,v_5) = YES
(8) TaqMan(v_0,v_3,v_4) = NO
(9) TaqMan(v_0,v_3,v_5) = NO
(10) TaqMan(v_0,v_4,v_5) = YES

The real-time PCR reaction involves primers (Sigma Genosys, Japan), TaqMan probes (Sigma Genosys, Japan), and QuantiTect Probe PCR Kit (QIAGEN, Japan). Ten separate real-time PCR reactions were performed in parallel, in order to implement the first stage of the proposed HPP readout. After the initial activation step at 95ºC for 15 minutes, the amplification consists of 45 cycles of denaturation and annealing/extension, performed at 94ºC for 15 s and 60ºC for 60 s, respectively. The resulting real-time PCR amplification plots are illustrated in Figure 3.

All real-time PCR reactions are completed, the *in vitro* output is subjected to an algorithm for *in silico* information processing, producing the satisfying Hamiltonian path of the HPP instance in $O(n^2)$ TIME (here, n denotes vertex number). The next step is to use all the information from 10 TaqMan reactions to allocate the each nodes of the Hamiltonian path. This can be done by applying the *in silico* algorithm below.

Input: A[0…|V|-1]=2 // A[2, 2, 2, 2, 2, 2, 2]
A[0]=1, A[|V|-1]=|V| // A[1, 2, 2, 2, 2, 2, 7]

```
for k=1 to |V|-3
    for l=2 to |V|-2
        while l>k
            if TaqMan(v0, vk, vl) = YES
                A[l] = A[l]+1
            else A[k] = A[k]+1
            endif
        endwhile
    endfor
endfor
```

Output : A = {1, 2, 4, 6, 3, 5, 7}

3. Implementation of clustering algorithms

From the output graph of real-time PCR, an unsupervised learning such as clustering algorithm can be implemented for automatic classification of a data [4]. Clustering algorithm, K-means for example is easy to implement, since the number of groups for classification of TaqMan reactions are already known. However, hard partitioning or non-fuzzy clustering algorithm requires an exhaustive search in a huge space because some variables can only take two values 0 and 1. Meanwhile fuzzy clustering is more computational tractability than hard partitioning algorithm [5]. Hence, all the variables are continuous, so that derivatives can be computed to find the right direction for the search [6].

Fuzzy C-means (FCM) has become the well known and powerful method in cluster analysis, and has been applied in many fields. However FCM clustering algorithm cannot handles noise and

outliers data. Thus, AFCM clustering algorithm has been developed by Wu and Yang, to enhance thre robustness of FCM to noise and outliers as well as tolerate with unequal sized clusters [7].

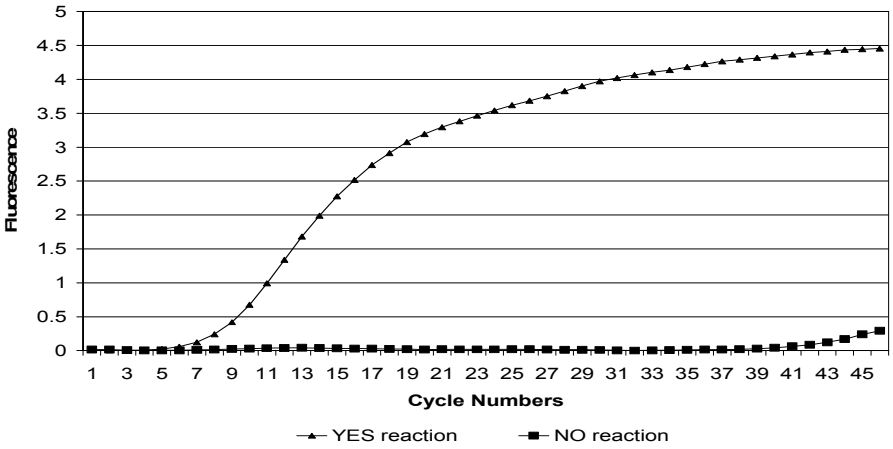

Figure 2. An example of reaction plots corresponding to TaqMan(v_0, v_k, v_l) = YES(first condition) and TaqMan(v_0, v_k, v_l) = NO (second condition).

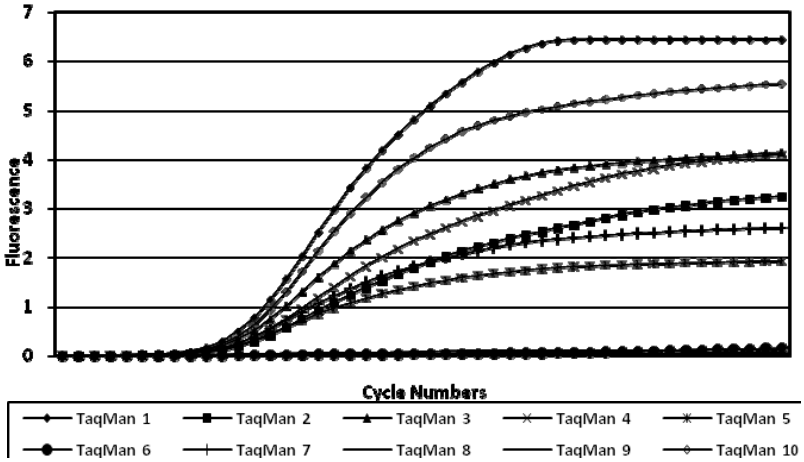

Figure 3. Output of real-time PCR. Reactions 1 to 10 indicate the $[(|V|-2)^2-(|V|-2)]/2$ TaqMan(v_0, v_k, v_l) reactions of the input instance, as defined in Section 3. Note that reactions 6, 8, and 9, which correspond to "NO", show virtually no amplification

3.1. FCM clustering algorithm

FCM is a data clustering technique based on the optimization of the objective function [8]:

$$J(U,V) = \sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^m \left\| x_j - v_i \right\|^2 \qquad (1)$$

where C is a number of clusters and N is a number of data. Every data point in the data set requires to belong to a cluster at a particular membership degree. The purpose of FCM is to group data points into different specific clusters. Let $X = \{x_1, x_2, ..., x_N\}$ be a collection of data. By minimizing (1), X is classified into C homogeneous clusters, where μ_{ij} is the membership degree of data x_j to a fuzzy cluster set $v_i, V = \{v_1, v_2, ..., v_C\}$ are the cluster centers. $U = (\mu_{ij})_{N*C}$ is a fuzzy partition matrix, and μ_{ij} indicates the membership degree of each data point in the data set to the cluster i. The value of U should satisfy the following conditions:

$$\mu_{ij} \in [0,1], \quad \forall i = 1,...C, \forall j = 1...N \qquad (2)$$

$$\sum_{i=1}^{C} \mu_{ij} = 1, \ \forall j = 1,...N \qquad (3)$$

The $\|x_j - v_i\|$ is the Euclidean distance between x_j and v_i. The parameter m is called fuzziness value index, which control the fuzziness value of membership of each datum. The cluster center can be calculated by using the following equation

$$v_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m x_j}{\sum_{j=1}^{N} (\mu_{ij})^m}, \ \forall i = 1,..., C \quad (4)$$

Then, clustering can be achieved by iteratively minimize the aggregate distance between each data point in the data set and cluster centers until no further minimization is possible. Then, the fuzzy partition matrix U is updated by using following equation

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}} \quad (5)$$

3.2. AFCM clustering algorithm

Wu and Yang have shown that Euclidean norm is not robust in a noisy environment [7]. Thus, they proposed the exponential distance metric to replace the Euclidean distance in the objective function. The exponential distance is given by:

$$d(x_j, v_i) = \left(1 - \exp\left(-\beta \|x_j - v_i\|^2\right)\right)^{1/2} \quad (6)$$

where β is a positive constant, defined by:

$$\beta = \left(\frac{\sum_{j=1}^{N} \|x_j - \bar{x}\|^2}{N} \right)^{-1} \quad (7)$$

and $\bar{x}$ is defined as

$$\bar{x} = \left(\frac{\sum_{j=1}^{N} x_j}{N} \right) \quad (8)$$

By replacing Euclidean distance with the exponential distance, the new objective function is formulated as:

$$J(U,V) = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij})^m \left(1 - \exp\left(-\beta \|x_j - v_i\|^2\right)\right) \quad (9)$$

which defines the formulation of AFCM clustering algorithm. The necessary condition for minimizing equation (9) is given in equation (10) and (11):

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{(1 - \exp(-\beta \|x_j - v_i\|^2))}{(1 - \exp(-\beta \|x_j - v_k\|^2))} \right)^{\frac{1}{m-1}}} \quad (10)$$

$$v_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m \exp(-\beta \|x_j - v_i\|^2) x_j}{\sum_{j=1}^{N} (\mu_{ij})^m \exp(-\beta \|x_j - v_i\|^2)} \quad (11)$$

A fixed-point iteration method can be implemented for calculating the v_i.

3.3. Classification of TaqMan reactions

In order to cluster the results of TaqMan reaction, namely "YES" and "NO" reaction, each graph of the reactions are represented as vector $\mathbf{x}_j = \{x_{j(1)}, x_{j(2)}, x_{j(3)}, \ldots, x_{j(45)}, x_{j(46)}\}$. The reactions are clustered into two groups, having their centre at $\mathbf{v}_1 = \{v_{1(1)}, v_{1(2)}, \ldots, v_{1(45)}, v_{1(46)}\}$ and $\mathbf{v}_2 = \{v_{2(1)}, v_{2(2)}, \ldots, v_{2(45)}, v_{2(46)}\}$. The two centres can be viewed as graphs that similar to the TaqMan reaction "YES" and "NO". It is noticed that the centre that is located in the amplification region always have greater value than the other center in the non-amplification region. The two centers are called as "YES" and "NO" center, where "YES" center is greater than "NO" center. This information is used to classify the TaqMan reactions "YES" and "NO" by comparing the fuzzy partition matrix U. If $\mathbf{v}_2$ represents the "YES" center and $\mathbf{v}_1$ represent the "NO" center (note that $\mathbf{v}_2$ is not always represent the "YES" center when the FCM algorithm is implemented), then $\mathbf{v}_2 > \mathbf{v}_1$. For example, if U_{11} and U_{12} equals to 0.6 and 0.4, respectively, the "YES" and "NO" reaction can be determined by following this rule

if ($\mathbf{v}_1 > \mathbf{v}_2$ and $U_{i1} > U_{i2}$) or ($\mathbf{v}_2 > \mathbf{v}_1$ and $U_{i2} > U_{i1}$))
    $\mathbf{x}_j$ ="YES"
else $\mathbf{x}_j$ ="NO"

Based on the proposed rule, $\mathbf{x}_j$ is classified as "NO" reaction since $U_{i1} > U_{i2}$ and $\mathbf{v}_1 < \mathbf{v}_2$. This rule is applied to the remaining "YES" and "NO" reactions. The whole classification process for FCM and AFCM can be described in the following steps

FCM

Step 1: Initialize the membership matrix U with random values, subject to (2) and (3)
Step 2: Calculate the cluster center V by using (4)
Step 3: Update fuzzy partition matrix U by using (5)
Step 4: Stop if $\|U(t+1)-U(t)\| < e$, otherwise go to step 2
Step 5: Determine "YES" and "NO" centers (either $\mathbf{v}_1 > \mathbf{v}_2$ or $\mathbf{v}_2 > \mathbf{v}_1$)

Step 6: Classify each TaqMan reactions by using the predefined rule

2 AFCM

Step 1: Initialize the membership matrix U with random values, subjected to conditions (2) and (3)

Step 2: Calculate the cluster center V by solving the equation (11), and calculate for each dimension of V.

Step 3: Update fuzzy partition matrix U based on (10)

Step 4: Stop if $\|U(t+1)-U(t)\| < e$, otherwise go to step 2

Step 5: Determine "YES" and "NO" centers (either $\mathbf{v}_1 > \mathbf{v}_2$ or $\mathbf{v}_2 > \mathbf{v}_1$)

Step 6: Classify each TaqMan reactions by using the predefined rule

4. Result and Discusion

FCM and AFCM are implemented to classify the TaqMan reaction produced by the DNA Engine Opticon 2. The clusterinf parameters used are e=0.00001, m=2, N=10, and C=2. The algorithms are implemented on Matlab 7.0. The results are shown in Figure 4, and Figure 5. The fuzzy partition membership values for those two different algorithms are listed in Table 1. Based on Table 1, two misclassifications occurred when the conventional FCM is applied. On the other hand, the AFCM has correctly classified all the TaqMan reactions. Based on the amplification output of TaqMan reactions, reactions 1 and 10 are clearly out of range from the "YES" group, reactions 6, 8 and 9 are closely grouped together to form "NO" group, and reactions 2,3,4,5 and 7 can be viewed as "YES" group. The misclassification, occurred for the FCM, is affected by reactions 1 and 10, as the FCM clustering algorithm is based on the "means" or "average". From this observation, reactions 1 and 10 can be classified as outliers or noises.

The misclassification problem can be corrected by using different distance measure such as exponential Euclidean distance, which is employed by AFCM. Table 1 shows that the AFCM has correctly classified all the TaqMan reaction, which proves that the exponential distance is more robust than Euclidean distance, in terms of handling of outliers and noises.

5. Conclusion

Based on the DNA computing readout approach based on the real-time PCR, the output of real-time PCR is clustered for automatically implementation of *in silico* information processing algorithm. The AFCM clustering has correctly classified the two reactions, namely "YES" and "NO".

6. Acknowledgement

This research is supported financially by the Ministry of Higher Education (MOHE) under Fundemantal Research Grant Scheme (FRGS) (Vo7 78225). This work also is partly supported by the Universiti Teknologi Malaysia under Initial Research Grant Scheme for Students (IRGS) (Vot 78285) and the Ministry of Science, Technology, and Innovation (MOSTI) under ScienceFund (Vot 79034). Muhammad Faiz Mohamed Saaid is indebted to Universiti Teknologi Malaysia for granting him a financial support and opportunity to do this research.

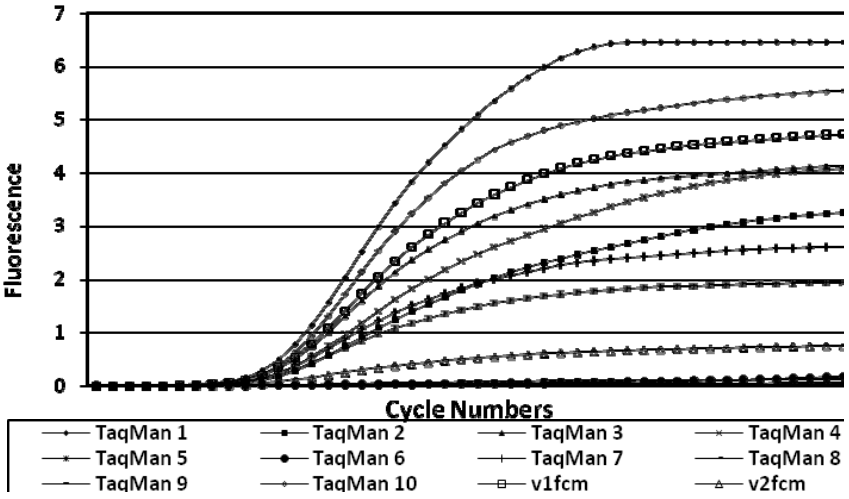

Figure. 4. Output of real-time PCR. with "YES" and "NO" centers calculated using FCM clustering algorithm

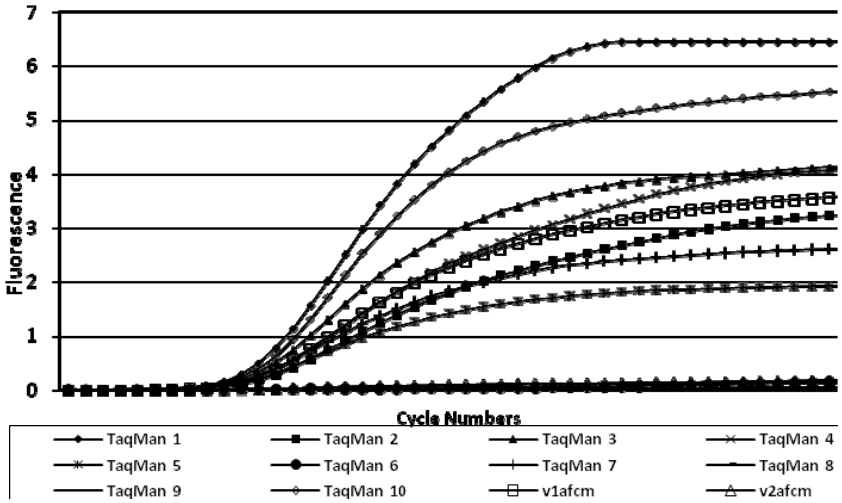

Figure. 5. Output of real-time PCR. with "YES" and "NO" centers calculated using AFCM clustering algorithm

Table 1. Fuzzy partition value for each TaqMan reactions

TaqMan	manual	FCM		Reaction	AFCM		Reaction
		U_{i1}	U_{i2}	($\mathbf{v}_1 > \mathbf{v}_2$)	U_{i1}	U_{i2}	($\mathbf{v}_1 > \mathbf{v}_2$)
1	YES	0.89454	0.10546	YES	0.52675	0.47325	YES
2	YES	0.60642	0.39358	YES	0.93949	0.060506	YES
3	YES	0.97209	0.027907	YES	0.90602	0.093975	YES
4	YES	0.90606	0.093941	YES	0.97078	0.029225	YES
5	YES	0.17222	0.82778	NO	0.58404	0.41596	YES
6	NO	0.018769	0.98123	NO	0.000716	0.99928	NO
7	YES	0.45078	0.54922	NO	0.8486	0.1514	YES
8	NO	0.021623	0.97838	NO	0.00265	0.99735	NO
9	NO	0.017352	0.98265	NO	0.000141	0.99986	NO
10	YES	0.96744	0.032556	YES	0.61813	0.38187	YES

7. References

[1] L. Overbergh, "The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression," *Journal of Biomolecular Techniques* vol. 14, pp. 557-559, 2003.

[2] N.J. Walker, "A technique whose time has come," *Science* vol. 296, pp. 557-559, 2002.

[3] Z. Ibrahim, J. A. Rose, Y. Tsuboi, O. Ono, and M. Khalid, "A New Readout Approach in DNA Computing Based on Real-Time PCR with TaqMan Probes," *Lecture Notes in Computer Science (LNCS), Springer-Verlag, C. Mao and T. Yokomori (Eds.)*, vol. 4287, pp. 350-359, 2006.

[4] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London: Arnold, 2001.

[5] J. C. Bezdek, "Cluster Validity with Fuzyy Sets", *J. Cybernatics*, vol. 3., pp. 58-73. 1974.

[6] L. X. Wang, *A Course of Fuzzy Systems and Control*. Prentice Hall, 1997.

[7] K.L. Wu, and M.S. Yang ,"Alternative C-Means Clustering Algorithm." *Pattern Recognition* vol 35., pp 2267-2278. 2000

[8] J. Bezdek, "*Pattern Recognition with Fuzzy Objective Function Algorithms*," Plenum Press, New York, 1981.