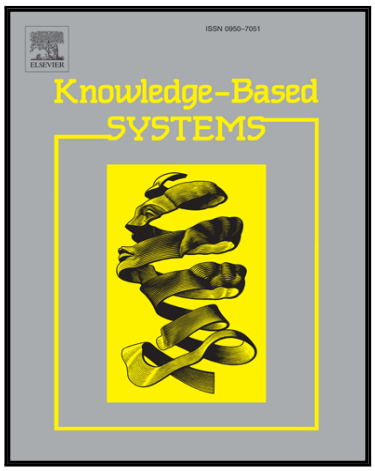

Granular Fuzzy Possibilistic C-Means Clustering Approach to DNA Microarray Problem

Hung Quoc Truong, Long Thanh Ngo, Witold Pedrycz

PII: S0950-7051(17)30293-9
DOI: 10.1016/j.knosys.2017.06.019
Reference: KNOSYS 3947

To appear in: *Knowledge-Based Systems*

Received date: 4 December 2016
Revised date: 10 June 2017
Accepted date: 12 June 2017

Please cite this article as: Hung Quoc Truong, Long Thanh Ngo, Witold Pedrycz, Granular Fuzzy Possibilistic C-Means Clustering Approach to DNA Microarray Problem, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.06.019

Granular Fuzzy Possibilistic C-Means Clustering Approach to DNA Microarray Problem

Hung Quoc Truong[a], Long Thanh Ngo[a], and Witold Pedrycz[b,c,d]

[a]*Department of Information Systems, Le Quy Don Technical University, Hanoi, Vietnam*
[b]*Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB Canada*
[c] *Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, 21589, Saudi Arabia*
[d] *Systems Research Institute, Polish Academy of Sciences Warsaw, Poland*
e-mail: truongqhung@gmail.com, ngotlong@mta.edu.vn, wpedrycz@ualberta.ca

Corresponding Author: Long Thanh Ngo, Mobile: +84-914 364 064.

Abstract

Deoxyribonucleic acid (DNA) microarray is an important technology, which supports a simultaneous measurement of thousands of genes for biological analysis. With the rapid development of the gene expression data characterized by uncertainty and being of high dimensionality, there is a genuine need for advanced processing techniques. With this regard, Fuzzy Possibilistic C-Means Clustering (FPCM) and Granular Computing (GrC) are introduced with the aim to solve problems of feature selection and outlier detection. In this study, by taking advantage of the FPCM and GrC, an Advanced Fuzzy Possibilistic C-Means Clustering based on Granular Computing (GrFPCM) is proposed to select features as a preprocessing phase for clustering problems while the developed granular space is used to cope with uncertainty. Experiments were completed for various gene expression datasets and a comparative analysis is reported.

Keywords: fuzzy clustering, fuzzy possibilistic c-means clustering, granular computing, feature selection, microarray technology, DNA analysis, gene expression data.

1. Introduction

Deoxyribonucleic acid (DNA) microarray is an important technology which facilitates the measurement of thousands of genes coming from different samples [16]. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the gene expression data, which often involves millions of measurements.

Clustering is a technique widely used in data mining consisting of bioinformatic. Currently, clustering problems often deal with large and highly dimensional datasets. This also raises important issues to be addressed on how to retrieve useful information from such datasets [36]. In this section, a literature review is given on the addressed issues: 1) the usage of the clustering techniques in DNA microarray problem and 2) the advantage of possibilistic approach to fuzzy clustering and 3) feature selection approaches to highly dimensional DNA problem.

Many clustering algorithms have been adapted or directly applied to gene expression data to partition a given data set into groups based on specified features to reveal natural structures, which have drawn a great deal of attention in the bioinformatics community. Genes or samples with similar expression patterns can be clustered together with similar cellular functions. This approach may further foster understanding of the functions of many genes for which information has not been previously available [13]. The clustering methods which were used for gene expression data, are mainly hierarchical clustering and some other conventional clustering algorithms such as a mixture of multivariate Gaussians (FMG), K-Means and spectral clustering (SPC) [13]. Eisen et al. [17] presented cluster analysis for genome-wide expression data which uses standard statistical algorithms to arrange genes according to similarity in patterns of gene expression, which clustering methods have been shown the usefulness in analysis of gene expression data. The drawbacks of various clustering methods for performing a large-scale dimensionality of gene expression data were also shown by Souto et al [13] where the applications of seven different clustering methods were studied. Those methods included finite mixture of Gaussians (FMG), K-Means and hierarchical methods to analyze 35 cancer gene expression datasets where the FMG exhibited the best performance, followed closely by K-Means.

Besides, Wang et al. [12] proposed a modified K-Means algorithm for human genetic research and other biomedical applications. Chen [8] proposed a neighbour-based method for gene assessment which is used for enhancing the discovery of interesting clusters. Mukhopadhyay et al. [4] proposed a way to improve fuzzy clustering by combining it with support vector machine (SVM) classifier for gene expression data. Sun et al. [3] proposed a new clustering method for gene expression datasets which is the combination of K-Means algorithm and a modified version of Quantum-behaved Particle Swarm Optimization (QPSO) algorithm, known as the Multi-Elitist QPSO (MEQPSO) model. Their results showed a promising research direction for gene clustering but still exhibits some restrictions especially when encoding the particle. Hastie et al. [9] presented a statistical method called "gene shaving" which defines subsets of genes through the coherent expression patterns and large variation across conditions. Gene shaving differs from the other widely used methods for gene expression analysis in which genes may belong to more than a single cluster, and the clustering may be controlled by some outcome measures. /The gene shaving method was used to analyze gene expression data with diffuse large B-cell lymphoma by determining a small cluster of genes whose expression is highly predictive of survival. However, the shaving process requires repeated computation of the largest principal component of a large set of variables. Thus, the gene shaving method is usually used for feature selection problem of the gene expression data which will be covered in more detail later.

Although these algorithms have exhibited usefulness for identifying biologically relevant groups of genes and samples, they do not work efficiently and produce sound results when coping with noisy, incomplete and uncertain data. Addressing the above challenges, fuzzy clustering algorithms were designed to deal with uncertain or vague data. Fuzzy C-Means (FCM) was considered as one of the most widely used fuzzy clustering which allows a data point to belong to more than one cluster with different membership grades [33]. The FCM algorithm assigns a pattern to a cluster on the basis of the inverse distance between them. In case the distances of a pattern to two centroids are approximately equal, confusion appears when assigning the pattern to clusters, which is considered as the noise sensitivity of fuzzy clustering [31]. To overcome this problem,

a certain version of fuzzy clustering is based on possibilistic approach which was first proposed by Krishnapuram et al. [32]. This algorithm determines a possibilistic partition in which a possibilistic membership is used to quantify a degree of typicality of a point belonging to a certain cluster. The larger the distance between an object to a centroid (prototype) is, the lower the possibilistic membership grade is, and the lower the impact of the particular object on the centroid is. Therefore, methods of outlier detection or noise removal are of interest.

However, in the possibilistic approach some drawbacks still exist, especially when it comes to choosing suitable values of the parameters of the clustering method. Pal et al. [31] proposed a method called Fuzzy Possibilistic C-Means which uses the membership values [33] as well as the typicality values of the PCM [32] to produce a better clustering algorithm. The constraint stating that the sum of all the typicality values of all data to a cluster must be equal to one causes problems; in particular for big data [30]. In order to handle this problem, Zhang et al. [30] proposed a combination of Fuzzy C-Means and Possibilistic C-Means, called Fuzzy Possibilistic C-Means (FPCM), to address some shortcomings associated with the possibilistic approach such as the noise sensitivity of FCM, resolve the coincident clusters of the possibilistic approach and eliminates the sum constraints of FPCM. Ferraro et al [44] focused on robust analysis of non-precise data on the basis of a fuzzy and possibilistic clustering method in which parameters were chosen by minimizing the Xie and Beni validity index.

Meanwhile, clustering techniques are commonly used in gene expression data. They also demonstrate some shortcomings when coping with highly dimensional data. Feature selection is one of the broadly used techniques to reduce the data dimensionality. It aims to select a subset of the relevant features according to a certain evaluation criterion so that the selected features fully represent the dataset to solve the problem [35, 36]. Many feature selection methods were proposed to analyze gene expression data. However, feature selection methods, which were proposed for clustering as filter, wrapper and hybrid models, are usually designed based on the greedy approach following a given evaluation criterion. This makes the methods time-consuming and of low efficiency when facing with very highly dimensional data. In such cases, forming relevant features is unclear and has to be carefully addressed.

Several studies related to DNA microarray problems have mentioned feature selection as an elementary tool for processing highly dimensional data. L.Shen et.al [7] used the penalized logistic regression combined feature reduction methods to cancer classification using microarray data. Zhu et al. [28] proposed a novel Markov blanket embedded genetic algorithm (MBEGA) for gene selection problem. The embedded Markov blanket based memetic operators are able to add or delete features (or genes) from a genetic algorithm (GA) solution so as to quickly improve the solution and fine-tune the search. Jaziri et al. [1] presented an efficient parallelization method for speeding up the complete backtranslation in generating all possible nucleic acid sequences for functional DNA microarrays. Kim et al. [10] also presented a meta-classifiers for high-dimensionality with the farthest-first clustering algorithm. Chen et al. [2] proposed a kernel-based clustering method for gene selection which was formed based on the best weights of genes by a process of kernel clustering. Vimaladev et al. [6] proposed Back Propagation Neural networks (BPN) and fast Genetic Algorithms (GA) to estimate the feature selection in gene expression data. Kah et al. [11] proposed a combined method of Gram-Schmidt orthogonal forward selection (OFS) and FunCluster to search for high-dimensional data in microarray data. Li et al. [5] study

the problem of building multiclass classifiers for tissue classification based on gene expression. The process of building multiclass classifiers is divided into two components: selection of the features (i.e. genes) to be used for training and testing and selection of the classification method.

Recently, Granular Computing (GrC) has emerged as a powerful vehicle to construct and process information granules. Information granules are formed by grouping similar objects, based on their similarity, closeness or proximity. It is used for handling complex problems, coping with massive data, capturing uncertainty, representing data of high dimensionality [36, 40]. In [39], W. Pedrycz synthesizes and reviews the granular fuzzy models which were built from the fuzzy data analysis and fuzzy regression. This paper also exhibited the direction of promising research on the fuzzy model based on GrC. Qian et. al [37] introduced the fuzzy granular structure distance to discriminate the difference between any two fuzzy granular structures which can be used to establish a generalized axiomatic constraint for fuzzy information granularity. Thus, this distance is a basis for granular clustering applications. In addition, GrC was applied to support vector machine (SVM) forming Granular support vector machine (GSVM) [34]. In this application, GSVM can improves the generalization ability and learning efficiency to a large extent when comparing with the traditional SVM and points out the research and development prospects. GrC can be used to solve the big data problems by hierarchical attribute reduction algorithms [38]. Beside, GrC can be combined with a clustering method to utilize feature selection for clustering to alleviate the negative impact of high dimensionality of the problem [40]. Sun et al. designed a feature selection method based on rough entropy [35] and GrC [40]. However, these feature selection methods are similar to the classification methods, which need labeled samples as training samples to select the necessary features, these applications were only focused on the classification or decision system problems.

From the above, we can see that the combination of clustering techniques and GrC is a promising way to apply clustering techniques to gene expression data problem while still dealing with the feature selection problem by GrC.

Thus, in this study, an advanced Fuzzy Possibilistic C-Means Clustering is proposed on a basis of a combination of FPCM algorithm [30] and Granular Computing [40] with an ultimate objective to handle the noise removal or outlier detection and feature selection for dealing with highly dimensional data. The proposed method not only takes advantage of the FPCM ability to handle noise, but also uses the concepts of GrC to assess the significance of the features, thus leading to the elimination of the effects of irrelevant features and noise. Namely, GrC is used to remove the irrelevant features and to form the granules which can handle the uncertainties to improve the efficiency of clustering methods. Thus, this algorithm potentially enhances the clustering results when working with gene expression data. Experiments are reported by using several publicly available gene expression data.

This paper is organized as follows: Section 2 briefly introduces some background concerning Fuzzy Possibilistic C-Means Clustering and Granular Computing; Section 3 proposes the advanced Fuzzy Possibilistic C-Means Clustering Based on Granular Computing; Section 4 offers some experimental results and section 5 covers conclusions and future research directions.

2. Preliminaries

2.1. Fuzzy Possibilistic C-Means Clustering Algorithm

Fuzzy Possibilistic C-Means Clustering Algorithm (FPCM) was proposed by Zhang et al. [30]. FPCM produces two types of membership grades: 1) A possibilistic membership that expresses the absolute degree of typicality of a point to any particular cluster, and 2) a membership that relates to the relative degree of sharing of the point among the clusters.

The objective function for FPCM is formed as follows:

$$J_{FPCM}(T,U,V;X,\gamma) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} t_{ik}^{p} d_{ik}^{2},\ 1 \le m,p \le \infty \tag{1}$$

$$+ \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} u_{ik}^{m} (1-t_{ik})^{p} \tag{2}$$

in which $d_{ik} = \| x_k - v_i \|$ is the Euclidean distance, c is the number of clusters, n stands for the number of objects, p is a weighting exponent (fuzzification coefficient) of the possibilistic membership ($p > 1$) and fuzzifier m ($m > 1$).

The scale parameter γ_i standing in (2) is to incorporate the possibilistic membership degrees and membership ones:

$$\gamma_i = K \frac{\sum_{k=1}^{n} t_{ik}^{p} u_{ik}^{m} d_{ik}^{2}}{\sum_{k=1}^{n} t_{ik}^{p} u_{ik}^{m}}, K > 0 \tag{3}$$

where K is a certain constant.

t_{ik} denotes the possibilistic membership degree of x_k belonging to the i^{th} cluster and u_{ik} stands for the degree of membership. They are determined as follows:

$$t_{ik} = \frac{1}{1 + \left(\frac{d_{ik}^{2}}{\gamma_i}\right)^{\frac{1}{p-1}}}, \forall i,k \tag{4}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{t_{ik}^{(p-1)/2}\, d_{ik}}{t_{jk}^{(p-1)/2}\, d_{jk}} \right)^{\frac{2}{m-1}}} \tag{5}$$

in which $i = 1, 2, ..., c$; $k = 1, 2, ..., n$.

The centroids (prototypes) are computed in the same way as in case of the FCM algorithm [33]:

$$v_i = \frac{\sum_{k=1}^{n} t_{ik}^{p} u_{ik}^{m} x_k}{\sum_{k=1}^{n} t_{ik}^{p} u_{ik}^{m}}, \forall i \tag{6}$$

$i = 1, 2, ..., c$.

Defuzzification (decoding) realized in the FPCM is realized in the following way: if $u_{ik} > u_{jk}$ for $j = 1, 2, ..., c$ and $j \neq i$ then x_k is assigned to the i^{th} cluster.

This algorithm is concisely described as follows:

Algorithm 1 Fuzzy Possibilistic C-Means Clustering algorithm

1 Input: A dataset $X = \{x_i, x_i \in R^d\}, i = 1, 2, ..., n$, the number of clusters $c\,(1 < c < n)$, weighting exponents $p, m(1 < p, m < +\infty)$ and error ε.

2 Output: The possibilistic membership matrix T, membership matrix U and the centroid matrix V.

3 Step 1:

3.1 The number of iterations is set to $l = 0$.

3.2 Execute a Fuzzy C-Means Clustering algorithm to find an initial $U^{(l)}$ and $V^{(l)}$.

3.3 Compute $\gamma_1, \gamma_2, ..., \gamma_c$ based on the $U^{(l)}$ and $V^{(l)}$ as follows: $\gamma_i = \frac{\sum_{k=1}^{n} u_{ik}^m d_{ik}^2}{\sum_{k=1}^{n} u_{ik}^m}$

4 Step 2:

repeat :

4.1 $l = l + 1$.

4.2 Update the possibilistic membership matrix $T^{(l)}$ by using (4).

4.3 Update the membership matrix $U^{(l)}$ by using (5).

4.4 Update the centroid matrix $V^{(l)} = \left[v_1^{(l)}, v_2^{(l)}, ..., v_c^{(l)}\right]$ by using (6).

4.5 Apply (3) to compute $\gamma_1, \gamma_2, ..., \gamma_c$ based on the $T^{(l)}$,$U^{(l)}$ and $V^{(l)}$.

until :

$$Max\left(||U^{(l+1)} - U^{(l)}||\right) \leq \varepsilon$$

5 Assign data x_k to i^{th} cluster if $u_{ik} > u_{jk}$, $j = 1, 2, ..., c$ and $j \neq i$.

2.2. Granular Computing

The framework of granular computing was proposed by Zadeh [42]. GrC is a computing paradigm of processing information [41]. When using granular computing in clustering, a granule is formed by a set of elements which are drawn together by indistinguishability, similarity, proximity or functionality.

Considering a clustering system $S = (X, A, V, f)$ denoted as $S(X, A)$ with $X = \{x_1, x_2, ..., x_n\}$ being a non-empty finite set of objects; $A = \{a_1, a_2, ..., a_d\}$ is a non-empty finite set of features; $V = \bigcup_{a \in A} V_a$ with V_a is called the value domain of the feature a, f is the information function of the system, $f : X \times A \to V$.

Some definitions [43] were introduced to granulate the clustering system. An indiscernibility relation on X is used to form a granule based on selecting the subsets of features.

Definition 2.1. *For each subset of features $B \subseteq A$, the non-empty set determines an indiscernibility relation on X as follows:*

$R_B = \{(x_i, x_j) \in X \times X | f_a(x_i) = f_a(x_j), \forall a \in B\}$

R_B is an equivalence relation on X, and it forms a partition of X, denoted by $X/R_B = \{[x_i]_B | x_i \in X\}$ where $[x_i]_B = \{x_j \in X | (x_i, x_j) \in R_B\}$ is called an equivalence class of x_i with respect to B.

A granule used for clustering system is defined as follows:

Definition 2.2. *Let $S = (X, A)$ be a clustering system. An information granule is defined as $gr_k = (\varphi_k, m(\varphi_k))$, where φ_k refers to the intention of information granule, and $m(\varphi_k)$ represents the extension of information granule.*

Suppose that $B = \{a_1, a_2, ..., a_{d'}\} \in A$ then there must exist $\varphi_k = \{I_1, I_2, ..., I_{d'}\}$ such that $I_j \in V_{a_j}$ is a set of feature values corresponding to B. Then, the intention of an information granule can be denoted by $\varphi_k = \{I_1, I_2, ..., I_{d'}\}$, and the extension can be denoted by $m(\varphi_k) = \{x \in X | f(x, a_1) = I_1 \wedge f(x, a_2) = I_2 \wedge ... \wedge f(x, a_{d'}) = I_{d'}, a_j \in B\}$, $j \in \{1, 2, ..., d^{'}\}$. Here, $m(\varphi_k)$ describes the internal structure of the information granule.

A granularity of system of features set B, denoted $GK(B)$, which is defined for examining the maintenance of clustering system.

Definition 2.3. *Let $S = (X, A)$ be a clustering system, the concept Granularity of System of features set B based on the Granules set $Gr = \{gr_k\}$ denoted $GK(B)$, $B \subseteq A$, is determined as follows:*

$$GK(B) = \sum_{k=1}^{|Gr/B|} |m(\varphi_k)|^2 / |X|^2, m(\varphi_k) \in gr_k \tag{7}$$

For example, the dataset $X = \{x_1, x_2, x_3, x_4\}$, $x_i \in R^3$, the set of features $A = \{a_1, a_2, a_3\}$ and $B = \{a_1, a_2\}$, where $x_1 = (1, 2, 3), x_2 = (1, 2, 1), x_3 = (2, 3, 1)$ and $x_4 = (1, 2, 2)$. Suppose $I_j = f(x_i, a_j) = x_i^{(j)}$ then we obtain the set of granules $Gr/B = \{gr_1, gr_2\}$, in which $gr_1 = (\varphi_1, m(\varphi_1))$, $\varphi_1 = (1, 2)$, $m(\varphi_1) = \{x_1, x_2, x_4\}$, and $gr_2 = (\varphi_2, m(\varphi_2))$, $\varphi_2 = (2, 3)$, $m(\varphi_2) = \{x_3\}$. Resulting in $GK(B) = (3^2/4^2) + (1^2/4^2) = 10/16$.

3. Advanced Fuzzy Possibilistic C-Means Clustering based on Granular Computing

3.1. Feature reduction based on Granular Computing

According to the underlying concepts of Granular Computing, the significance of a set of features in clustering system was proposed [43]. Given a clustering system $S = (X, A)$, there is a feature in A, denoted $a \in A$, so that we can express the degree of importance through the quantity of the granularity of A when the feature a is removed.

Definition 3.1. *The significance degree of feature $a \in A$, denoted $Sig_{A-\{a\}}(a)$, is defined as follows:*

$$Sig_{A-\{a\}}(a) = GK(A - a) - GK(A) \tag{8}$$

Note that the larger degree $Sig_{A-\{a\}}(a)$ takes, the more important the feature a is.

Definition 3.2. *Given an information system $S = (X, A)$ and feature $a \in A$, the feature a is called redundant feature to A if the value of $GK(A - a)$ is equal to $GK(A)$. Otherwise, the feature a is called necessary feature to A. The set of all the necessary features is the core of A, denoted $Core(A)$.*

Definition 3.3. *Given an information system $S = (X, A)$ and a set of features $C : C \subseteq A$. Set C is called a reduction of A if C is independent. All the reduction of A is denoted by Red(A).*

The reduction algorithm is described as follows:

Algorithm 2 Feature reduction based on Granular Computing

1 Input: A granular information system S=(X,A) where $X \neq \varnothing$ is the universe and $A \neq \varnothing$ is the set of features. The granularity of A is denoted as $GK(A)$.

2 Output: C is as the minimum reduction of A.

3 Step 1. Determine the core of features $Core(A)$ as follow: Calculate the significance degree of each feature $a \in A$, denoted $Sig_{A-\{a\}}(a)$, if $Sig_{A-\{a\}}(a) \neq 0$ then select feature a into $Core(A)$.

4 Step 2.

4.1 Assign $C := Core(A)$.

4.2 If $GK(C) = GK(A)$ then terminal criteria is meet.

4.3 **repeat** :

4.3.1 For each feature $a \in A - C$ to C, calculate its significance degree to $C \cup \{a\}$: $Sig_C(a)$.

4.3.2 Find the feature a so that its significance degree to C reaches the maximal value, i.e. $Sig_C(a) = \max\limits_{a' \in A-C} \left(Sig_C\left(a'\right)\right)$.

4.3.3 Add feature a to the core, i.e. $C := C \cup \{a\}$.

until : $GK(C) = GK(A)$

For example, the dataset $X = \{x_1, x_2, x_3, x_4\}$, $x_i \in R^4$, the set of features $A = \{a_1, a_2, a_3, a_4\}$, where $x_1 = (1, 1, 2, 1), x_2 = (2, 2, 1, 1), x_3 = (2, 2, 3, 1)$ and $x_4 = (3, 1, 2, 1)$.

Step 1:

Using Def.2.1, we have

$X/A = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$ and $|X/A| = 4, X_i = \{x_i\}, i = 1..4$

Using Eq.7, $GK(A) = \sum\limits_{i=1}^{|X/A|} |X_i|^2/|X|^2 = (1^2 + 1^2 + 1^2 + 1^2)/4^{\,2} = 1/4$

Step 2:

Using Def.2.1, we have

$X/(A - \{a_1\}) = \{\{x_1, x_4\}, \{x_2\}, \{x_3\}\}$

$X/(A - \{a_2\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$

$X/(A - \{a_3\}) = \{\{x_1\}, \{x_2, x_3\}, \{x_4\}\}$

$X/(A - \{a_4\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$

Using Eq.7, $GK(A - \{a_i\}) = \sum\limits_{i=1}^{|X/(A-\{a_i\})|} |X_i|^2/|X|^2$, we have

$GK(A - \{a_1\}) = 3/8$

$GK(A - \{a_2\}) = 1/4$

$GK(A - \{a_3\}) = 3/8$

$GK(A - \{a_4\}) = 1/4$

Calculate the significance degree of feature $a_i \in A$ using (8):

$Sig_{A-\{a_1\}}(a_1) = GK(A - \{a_1\}) - GK\{A\} = 3/8 - 1/4 = 1/8$

$Sig_{A-\{a_2\}}(a_2) = GK(A - \{a_2\}) - GK\{A\} = 1/4 - 1/4 = 0$

$Sig_{A-\{a_3\}}(a_3) = GK(A - \{a_3\}) - GK\{A\} = 3/8 - 1/4 = 1/8$

$Sig_{A-\{a_4\}}(a_4) = GK(A - \{a_4\}) - GK\{A\} = 1/4 - 1/4 = 0$

So $Core(A) = \{a_i \in A | Sig_{A-a_i}(a_i) > 0\} = \{a_1, a_3\}$, $GK(Core(A)) = GK(a_1, a_3) = 1/4, GK(Core(A)) = GK(A)$. Thus, $Core(A)$ is the minimum reduction of A.

3.2. Granular space construction and feature selection

Let consider a clustering system $S = (X, A)$ where $X = \{x_1, x_2, ..., x_n\}$ and $A = \{a_1, a_2, ..., a_d\}$. We construct a granular space as follows:

First, the objects $X = \{x_1, x_2, ..., x_n\}$ are clustered into c clusters on each j^{th} feature by FPCM algorithm, $j \in A$. On each j^{th} feature, the clusters are labeled by numbering them in ascending order starting from 1.

Secondly, a cluster label matrix, denoted F, is formed from $f(i, j)$ which is the label of the i^{th} object on the j^{th} feature, $1 \leq f(i, j) \leq c$, i.e. $F = [f(i, j)]_{(n\times d)}$.

Finally, from the values $\{f_1, f_2, ..., f_d\}$ of a row in the cluster label matrix F, we can construct a granule $gr_k = \{\varphi_k, m(\varphi_k)\}$ where $\varphi_k = \{f_1, f_2, ..., f_d\}$, $m(\varphi_k) = \{x_i \in X : f(i, 1) = f_1 \wedge f(i, 2) = f_2 \wedge ... \wedge f(i, d) = f_d\}$. So a granular space, denoted G, is formed from the set of granules, i.e. $G = \{gr_k\}, k = 1, 2, ..., n_g$ with n_g is the number of the granules, $1 \leq n_g \leq n$, denoted $n_g = |G|$.

Definition 3.4. *Consider a granular clustering system $S = (G, A)$, granular space $G = \{gr_k\}, k = 1, 2, ..., n_g$ and $n_g = |G|$. A non-conflict granular space with respect to A, denoted $GrSP$, is formed by $GrSP = \{gr_{k_1}\}$, in which $gr_{k_1} = \{\varphi_{k_1}, m(\varphi_{k_1})\}$ where $\varphi_{k_1} = \{f_1, f_2, ..., f_d\}$ and $f_1 = f_2 = ...f_d$. Otherwise, a conflict granular space with respect to A, denoted $GrSN$, is formed by $GrSN = \{gr_{k_2}\}$, in which $gr_{k_2} = \{\varphi_{k_2}, m(\varphi_{k_2})\}$, $\varphi_{k_2} = \{f_1, f_2, ..., f_d\}$ and $\exists f_p \neq f_q$*

Remark: The significance of a feature only affect the $GrSN$, thus the feature selection method can be only applied to the $GrSN$.

In the FPCM algorithm, the outlier or noisy object x_k can be removed, $X := X - \{x_k\}$ if x_k satisfies the following conditions:

$$t_{ik}^{(j)} < \theta \; with \; \forall i = 1, 2, ..., c \; and \; j = 1, 2, ..., d \tag{9}$$

where $t_{ik}^{(j)}$ is the possibilistic membership degree of x_k on the j^{th} feature in cluster i and θ is a noisy parameter.

Furthermore, the noisy feature a_j, $a_j \in A$ can be also removed, if $f(1, j) = f(2, j) = ... = f(n', j)$, where n' is the number of object in X after removing the outlier features.

$$A := A - \{a_j\} \tag{10}$$

The granular space construction and feature selection method can be briefly characterized as follows:

Algorithm 3 Granular construction and feature selection

1 Input: A dataset $X = \{x_i\}, i = 1..n$, $A = a_1, a_2, ..., a_d$, c is the number of cluster and θ is a noise filter parameter.

2 Output: The feature set C is the minimum reduction of A and the granular space G=$GrSN \cup GrSP$

3 Step 1:

3.1 Execute Algorithm 1 for each feature $a_j \in A$ to form a cluster label matrix $F = [f(i, j)]_{(n\times d)}$ where $f(i, j)$ is the cluster label of the i^{th} object on the j^{th} feature.

3.2 Remove outlier objects and features by using (9) and (10), respectively.

4 Step 2: Construct granular space
4.1 Initialize $GrSP = \varnothing, GrSN = \varnothing, r = 0, ID = \{1, 2, ..., n\}, k = 0$, where r is the index of row of the matrix F, ID is the index set and k is the number of granules.
4.2 **repeat**
4.2.1 $k = k + 1$
4.2.2 **repeat**
$r = r + 1$
until $r \in ID$
4.2.3 Set φ_k to set of values of r^{th} row in the matrix F: $\varphi_k = f(r, 1), f(r, 2), ..., f(r, d^{'})$, where $d^{'}$ is the number of features in A after removing the outlier features.
4.2.4 Find $m(\varphi_k) = \{x_i \in X : f(i, 1) = f(r, 1) \wedge f(i, 2) = f(r, 2) \wedge ... \wedge f(i, d^{'}) = f(r, d^{'})\}$.
if $|m(\varphi_k)| > 0$ **then**
4.2.4.1 **for** each $x_i \in m(\varphi_k)$:
$X = X - \{x_i\}, ID = ID - \{i\}$
4.2.4.2 $gr_k = (\varphi_k, m(\varphi_k))$
4.2.4.3 **if** $f(r, 1) = f(r, 2) = ... = f(r, d^{'})$ **then**
$GrSP = GrSP \cup \{gr_k\}$
else
$GrSN = GrSN \cup \{gr_k\}$
until $ID = \varnothing$
5 Step 3: Apply Algorithm 2 on the the granular set GrSN to reach the minimum reduction C of A.

3.3. Advanced FPCM based on Granular Computing

Consider a granular clustering system $S = (G, A)$, granular space $G = \{gr_k\}$, $k = 1, 2, ..., n$ and $n = |G|$.

The valued interval of the j^{th} feature of an input granule $gr_k = (\varphi_k, m_k(\varphi_k))$ is denoted $I_j^{(k)} = [a_j\ b_j]$ where a_j and b_j are defined as follows:

$$a_j = min(x_i^{(j)}), \forall x_i \in m_k(\varphi_k) \tag{11}$$

$$b_j = max(x_i^{(j)}), \forall x_i \in m_k(\varphi_k) \tag{12}$$

in which $x_i^{(j)}$ is the value of the object x_i on the j^{th} feature.

The new distance between a granule gr_k and the centroid $v_i = \{v_{i1}, v_{i2}..., v_{id}\}$, $d = |A|$, $i = 1, 2, ..., c$ is defined as follows:

$$\|gr_k - v_i\| = \sqrt{\sum_{j=1}^{d} \left(\left\| I_j^{(k)} - v_{ij} \right\| \right)^2} \tag{13}$$

where

$$||I_j^{(k)} - v_{ij}|| \stackrel{def}{=} \begin{cases} 0, \text{ if } v_{ij} \in [a_j, b_j] \\ \min\left(|a_j - v_{ij}|, |b_j - v_{ij}|\right) \end{cases} \tag{14}$$

The distance (13) is used to compute the possibilistic membership function and membership function as follows:

t_{ik} is the possibilistic membership degree of the granule gr_k in the i^{th} cluster and u_{ik} is the membership degree. They are determined in a similar way as in the FPCM algorithm:

$$t_{ik} = \frac{1}{1 + \left(\frac{d_{ik}^2}{\gamma_i}\right)^{\frac{1}{p-1}}}, \forall i, k \tag{15}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{t_{ik}^{(p-1)/2} d_{ik}}{t_{jk}^{(p-1)/2} d_{jk}}\right)^{\frac{2}{m-1}}} \tag{16}$$

in which $i = 1, 2, ..., c$, $k = 1, 2, ..., n$.

d_{ik} is calculated by using (13), if the distance between granule gr_k and v_i equals 0 then the membership u_{ik} is assigned value 1.

Cluster centroids are computed in the same way of FPCM as follows:

$$v_i = \frac{\sum_{k=1}^{n} t_{ik}^p u_{ik}^m \sum_{t=1}^{|m_k(\varphi_k)|} x_t | x_t \in m_k(\varphi_k)}{\sum_{k=1}^{n} t_{ik}^p u_{ik}^m}, \forall i \tag{17}$$

in which $i = 1, 2, ..., c$.

The GrFPCM algorithm comes in form:

Algorithm 4 Advanced FPCM based on Granular Computing

1 Input:

A clustering system $S(X, A)$ where a dataset $X = \{x_1, x_2, ..., x_n\}$, a set of features $A = a_1, a_2, ..., a_d$, the number of cluster c, error ε and noisy parameter θ.

2 Output:

The possibilistic membership matrix T, membership matrix U and the centroid matrix V.

3 Step 1: Apply Algorithm 3 on the clustering system $S(X, A)$ to obtain the feature set C which is the minimum reduction of A and the granular space G.

4 Step 2:

Apply Algorithm 1 on the clustering system $S = (G, C)$ as follows:

4.1 The number of iterations is set to $l = 0$.

4.2 **repeat** :

4.2.1 $l = l + 1$.

4.2.2 Update the possibilistic membership matrix $T^{(l)}$ by using (15).

4.2.3 Remove the outlier or noisy granular

$gr_{t_{ik} \geq \theta} = \{gr_k \in G : \max(t_{ik}) \geq \theta, \forall i = 1, 2, ..., c\}$.

4.2.4 Update the membership matrix $U^{(l)}$ by using (16).

4.2.5 Update the centroids $V^{(l)} = \left[v_1^{(l)}, v_2^{(l)}, ..., v_c^{(l)}\right]$ by using (17).

4.2.6 Apply (3) to compute $\gamma_1, \gamma_2, ..., \gamma_c$ based on the $T^{(l)}$, $U^{(l)}$ and $V^{(l)}$.

until :

$$Max\left(||U^{(l+1)} - U^{(l)}||\right) \leq \varepsilon$$

5 Assign data gr_k to the i^{th} cluster if $u_{ik} > u_{jk}$, $j = 1, 2, ..., c$ and $j \neq i$.

The diagram of algorithm 4 is described in Fig.1 below:

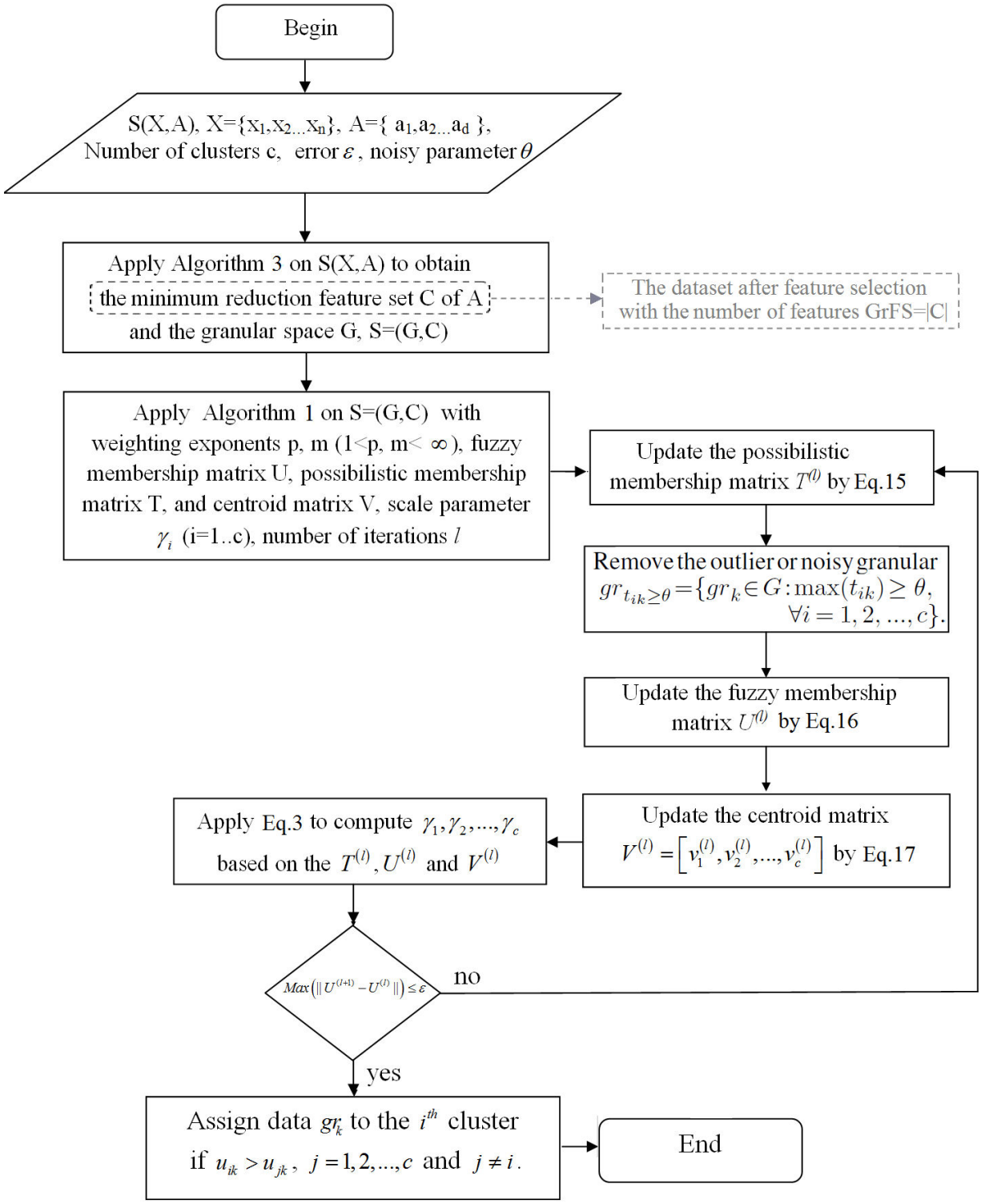

Figure 1: The diagram of algorithm 4

4. Experimental studies

4.1. Cluster analysis for Gene expression data

A gene expression data set from a microarray experiment can be represented by a real valued expression matrix $S = \{s_{ij}|1 \leq i \leq n, 1 \leq j \leq m\}$ where rows represent n genes, columns represent m different samples, and numbers in each cell represent the expression level of the particular gene i in the particular sample j. We consider that the samples as the objects and the genes as the features. The distinction of the sample based on clustering is to cluster the gene expression data into c clusters (c subtypes) where c is prior known number.

4.2. Results

In this section, twenty public gene expression datasets (benchmark data sets) which are described in Tab. 1 with the pre-defined number of clusters (classes) were used in the experiments. We also offer a comparative analysis of the clustering results among various clustering methods involved: FCM [33], FPCM [30], K-Means (KM), Mixture of multivariate Gaussians (FMG), spectral clustering (SPC), Shared nearest neighbor-based clustering (SNN), Hierarchical clustering with single linkage (SL), complete linkage (CL) and average linkage (AL) [13] and GrFPCM (the proposed method).

Table 1: The public gene expression datasets with ordinary number (O.N.), dataset names, number of samples (N), number of genes (M), number of classes (C), distribution of samples within the classes (Dist. Classes)

O.N.	Datatsets	N	M	C	Dist. Classes
1	Leukemia-V1 [18]	72	12582	2	24 ALL, 48 MLL
2	Leukemia-V2[18]	72	12582	3	24 ALL, 20 MLL, 28 AML
3	Leukemia-2c [28]	72	7129	2	47 ALL, 25 AML
4	Leukemia-3c [28]	72	7129	3	38 B-Cell, 9 T-Cell, 25 AML
5	Leukemia-4c[28]	72	7129	4	38 B-Cell, 9 T-Cell,21 BM, 4 PB
6	Lung Cancers-V1 [19]	203	12600	5	139 AD,17 NL,6 SCLC, 21 SD, 20 COID
7	Lung Cancers-V2 [20]	181	12533	2	31 MPM, 150 AD
8	Human Liver Cancers [22]	179	22699	2	104 HCC, 75 Liver
9	Breast, Colon Cancers [21]	104	22283	2	62 B, 42 C
10	Breast Cancers [29]	97	24482	2	46 Relapse, 51 Non-relapse
11	Colon Cancers [23]	37	22883	2	8 SCRC , 29 CCRC
12	Prostate Cancers -V1 [24]	110	42640	4	11 PT1, 39 PT2, 19 PT3, 41 Normal
13	Prostate Cancers -V2 [25]	104	20000	5	27 EPI, 20 MET, 32 PCA, 13 PIN, 12 STROMA
14	Bone marrow-V1 [27]	248	12625	2	43 T-ALL, 205 B-ALL
15	Bone marrow-v2 [27]	248	12625	6	15 T-ALL, 27 E2A-PBX1, 64 BCR-ABL, 20 TEL-AML1, 79 MLL, 43 Hyperdiploid >50
16	Ovarian [29]	253	15154	2	162 Cancers, 91 Normal
17	Lymmopha [29]	66	4026	3	46 DLBCL, 9 FL,11 CLL
18	CNS [29]	60	7129	2	21 Y, 39 N
19	SRBCT [29]	83	2308	4	29 EWS, 11 BL, 18 NB, 25 RMS
20	Bladder Cancers [26]	40	7129	3	9 T2+, 20 Ta, 11 T1

Through the adjustments in the experiments, the clustering results are stable with parameters which are set as follows:

Exponential parameters m and p are set to 2, the noise parameter $\theta = 0.1$, error $\varepsilon = 0.00001$, the adjustment γ in FPCM and GrFPCM methods is calculated with $K = 1$. The clustering algorithms such as FCM, Kmeans and SPC were done 30 times for each configuration and the best IC and ARI were selected.

Based on our proposed algorithm (GrFPCM), we performed gene expression data clustering in two main stages:

Stage 1: We have done the feature selection by step 1 of the GrFPCM on the experimental datasets to get the informative genes (subset of the relevant features). The comparing clustering algorithms such as K-Means, FCM, FPCM can be performed on the dataset after feature selection.

Stage 2: After performing feature selection of the gene expression datasets, we also have built up granules for the GrFPCM clustering method.

A given dataset S of n samples, and two groups (e.g. clusters) of these samples, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_r\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$, shown in Tab. 2.

Table 2: The contingency table

X /Y	Y_1	Y_2	$\ldots$	Y_r	sums
X_1	n_{11}	n_{12}	$\ldots$	n_{1r}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2r}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rr}	a_r
sums	b_1	b_2	$\ldots$	b_r	

The performance of the clustering was evaluated with incorrectly clustered instances (IC) and adjust rand index (ARI) [13] which are defined by the following expression:

$$\text{IC} = \frac{n - \sum n_{ii}}{n} \tag{18}$$

where n is the number of samples and n_{ii} is the value taken from Table 2

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{19}$$

where n_{ij},a_i, b_j are values from the Tab.2 and the notation $\binom{a}{b}$ is the binomial coefficient $\frac{a!}{b!(a-b)!}$.

The ARI index [15] is a version of the Rand index [14]. Though the Rand Index may only take a value between 0 and +1, the ARI can take values from -1 to 1, with 1 indicating a perfect agreement between the partitions. That means that the higher ARI index is equivalent to the better clustering results and vice versa.

Firstly, we made the gene expression datasets clustering from 20 datasets by running K-Means, FCM, FPCM and GrFPCM. Then clustering results were compared with the defined classes in the datasets to calculate IC values followed the formula (18). The results are listed in Tab.3.

Table 3: Clustering results with IC values of the experimental datasets without feature selection (N.o is the number of incorrectly clustered instances)

O.N.	Datatsets	Incorrectly clustered instances (%)							
		K-Means		FCM		FPCM		GrFPCM	
		N.o	%	N.o	%	N.o	%	N.o	%
1	Leukemia-V1 [18]	22	30.5556	20	27.7777	18	25	2	2.7778
2	Leukemia-V2[18]	21	29.1667	21	29.1667	15	20.8333	0	0
3	Leukemia-2c [28]	21	29.1667	20	27.7777	17	23.6111	2	2.7778
4	Leukemia-3c [28]	34	47.2222	18	25	13	18.0555	1	1.3889
5	Leukemia-4c[28]	42	58.3333	22	30.5556	22	30.5556	15	20.8333
6	Lung Cancers-V1 [19]	96	47.2906	95	46.798	61	30.0493	35	17.2413
7	Lung Cancers-V2 [20]	30	16.5746	2	1.105	2	1.105	0	0
8	Human Liver Cancers [22]	80	44.6927	89	49.7207	80	44.6927	22	12.2905
9	Breast, Colon Cancers [21]	44	42.3077	15	14.423	8	7.6923	3	2.8846
10	Breast Cancers [29]	45	46.3918	37	38.1443	29	29.8969	18	18.5567
11	Colon Cancers [23]	14	37.8378	13	35.1351	13	35.1351	11	29.7297
12	Prostate Cancers -V1 [24]	51	46.3636	63	57.2727	56	50.909	31	28.1818
13	Prostate Cancers -V2 [25]	55	52.8846	40	38.4615	65	62.5	29	27.8846
14	Bone marrow-V1 [27]	88	35.4839	87	35.0806	50	20.1613	6	2.4194
15	Bone marrow-v2 [27]	169	68.1452	107	43.1452	170	68.5484	73	29.4354
16	Ovarian [29]	112	44.2688	86	33.992	75	29.6442	2	0.7905
17	Lymmopha [29]	22	33.3333	20	30.303	20	30.303	10	15.1515
18	CNS [29]	29	48.3333	26	43.3333	19	31.6666	15	25
19	SRBCT [29]	52	62.6506	27	32.5301	22	26.506	5	6.0241
20	Bladder Cancers [26]	18	45	18	45	8	20	5	12.5

Lower IC index values point at the better clustering results. Thus, in Tab.3, the clustering results show that the GrFPCM algorithm shows its superiority over all 20 datasets, particularly, the IC values equal 0 with two datasets (2^{nd} and 7^{th}) which mean that GrFPCM reaches the absolute accuracy with these datasets. Next, Fig.2 visualizes the clustering results (IC index values) computed on the basis of the K-Means, FCM, FPCM and GrFPCM in Tab.3. Obviously, the proposed algorithm (GrFPCM) obtained the best results (exhibiting the smallest IC index values) in all experimental datasets.

Secondly, the K-Means, FCM and FPCM methods were done on the datasets with feature selection by the GrFPCM method. It means that the compared clustering algorithms were done with the datasets which their features were reduced by step 1 of the GrFPCM algorithm. Then, IC index values were also calculated to assess the clustering results, which are shown in Tab.4.

Note that: N.o is the number of incorrectly clustered instances; GrFS is the number of features after performing feature selection by step 1 of the GrFPCM algorithm.

In Tab.4, the clustering results reveal that GrFPCM also exhibited the best performance with

Table 4: Clustering results with IC values of the datasets after performing feature selection (the K-Means, FCM and FPCM methods were done on the datasets with feature selection by GrFPCM method)

O.N.	Datatsets	GrFS	Incorrectly clustered instances (%)							
			K-Means		FCM		FPCM		GrFPCM	
			N.o	%	N.o	%	N.o	%	N.o	%
1	Leukemia-V1 [18]	34	7	9.7222	7	9.7222	7	9.7222	2	2.7778
2	Leukemia-V2[18]	150	10	13.889	10	13.889	5	6.9444	0	0
3	Leukemia-2c [28]	81	8	11.1111	7	9.7222	5	6.9444	2	2.7778
4	Leukemia-3c [28]	104	6	8.3333	6	8.3333	5	6.9444	1	1.3889
5	Leukemia-4c[28]	126	21	29.1667	21	29.1667	16	22.2222	15	20.8333
6	Lung Cancers-V1 [19]	512	42	20.6896	40	19.7044	40	19.7044	35	17.2413
7	Lung Cancers-V2 [20]	93	5	2.7624	2	1.105	0	0	0	0
8	Human Liver Cancers [22]	80	76	42.4581	80	44.6927	72	40.2235	22	12.2905
9	Breast, Colon Cancers [21]	22	8	7.6923	5	4.8077	5	4.8077	3	2.8846
10	Breast Cancers [29]	1054	22	22.6804	20	20.619	20	20.619	18	18.5567
11	Colon Cancers [23]	51	11	29.7297	11	29.7297	11	29.7297	11	29.7297
12	Prostate Cancers -V1 [24]	68	35	31.8182	34	30.9091	34	30.9091	31	28.1818
13	Prostate Cancers -V2 [25]	138	40	38.4615	40	38.4615	63	60.5769	29	27.8846
14	Bone marrow-V1 [27]	216	44	17.7419	35	14.1129	6	2.4194	6	2.4194
15	Bone marrow-v2 [27]	186	99	39.9194	81	32.6613	107	43.1452	73	29.4355
16	Ovarian [29]	35	12	4.7431	7	2.7668	7	2.7668	2	0.7905
17	Lymmopha [29]	272	18	27.2727	18	27.2727	11	16.6667	10	15.1515
18	CNS [29]	32	23	38.3333	23	38.3333	19	31.6667	15	25
19	SRBCT [29]	162	27	32.5301	27	32.5301	14	16.8675	5	6.0241
20	Bladder Cancers [26]	79	12	30	12	30	8	20	5	12.5

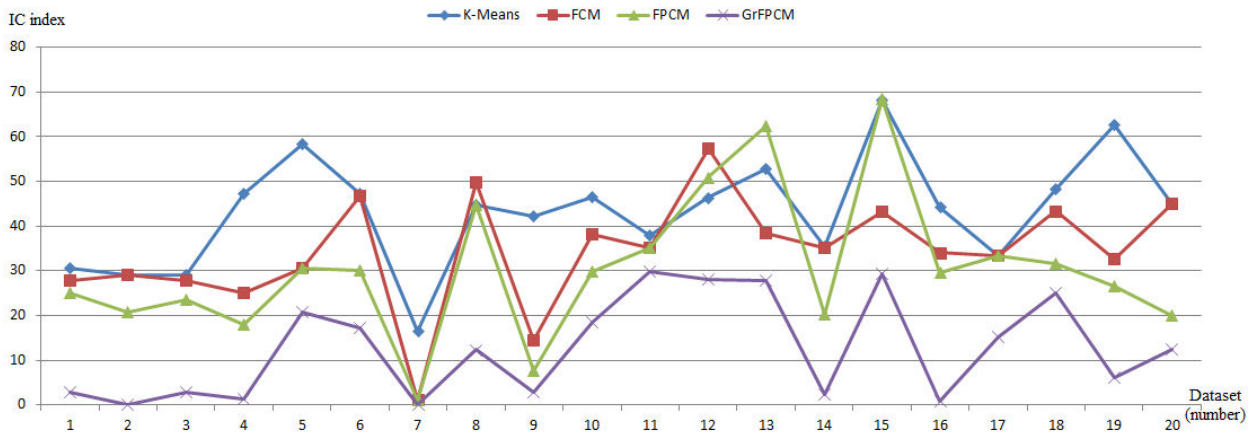

Figure 2: IC index values for K-Means, FCM, FPCM, GrFPCM

the smallest IC values over all twenty datasets. However, K-Means, FCM and FPCM have achieved much better results than theirs when performing on the datasets without feature selection which shown in Tab.3. Meanwhile, Fig.3 shows the clustering results (IC index values) computed from the K-Means, FCM, FPCM and GrFPCM in Tab.4. Clearly, GrFPCM leads to the best results (smallest IC index values) in all experimental datasets.

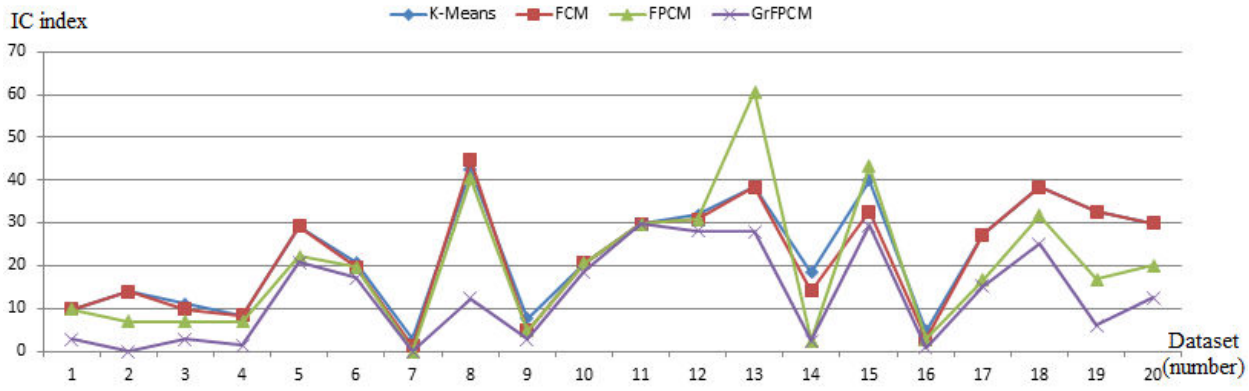

Figure 3: IC index values with feature selection for K-Means, FCM, FPCM, GrFPCM

Fig. 4 shows us a comparison of the clustering results (IC index values) produced from the K-Means, K-Means (GrFS), FCM, FCM (GrFS), FPCM, FPCM (GrFS) and GrFPCM methods, where K-Means (GrFS), FCM (GrFS), FPCM (GrFS) methods are completed on the datasets with feature selection by GrFPCM methods. Clearly, the clustering results with feature selection are much more outstanding than those without feature selection.

Finally, methods of the K-Means, FMG, SNN, SL, CL, AL, SPC, FCM [33], and FPCM [30] were done on the datasets with the different feature selection algorithms which were referenced from [13] such as Removing features with low variance (Lung Cancers [20], Prostage Cancers [24, 27]), Univariate feature selection (Bone marrow [27]), Recursive feature elimination (Leukemia [18], Breast, Colon Cancers [21], Colon Cancers [23]), Feature selection using a back-

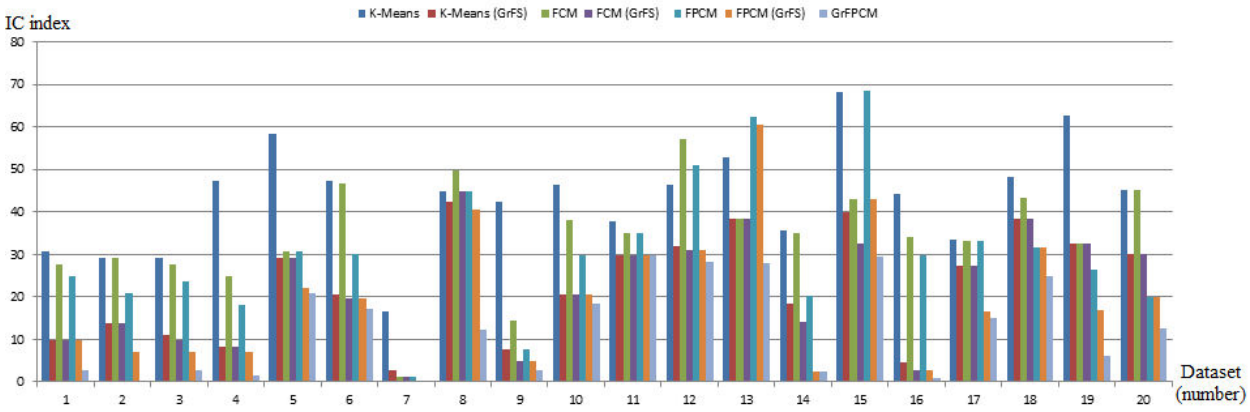

Figure 4: IC index values with and without feature selection for K-Means, FCM, FPCM, GrFPCM

ward removal process (Human Liver Cancers [22]), Tree-based feature selection (Lung Cancers [19]), Signal-to-noise ratio (SNR) ranking (Bladder Cancers [26]).

It means that the datasets with different feature selection algorithms are compared with the datasets which their features were selected by the proposed algorithm. Next, the ARI (also called Cr in reference [13]) values of K-Means, FMG, SNN, SL, CL, AL and SPC methods are referenced from [13] and ARI values are also calculated followed the formula (19) from 12 datasets with FCM, FPCM and GrFPCM which were listed in Tab.5 and Tab.8.

Note that: FS is the number of features on a dataset with the different feature selection algorithms which were referenced from [13].

Table 5: Clustering results with ARI values of the datasets after performing feature selection [13]

N.O.	Datatsets	FS [13]	SL	AL	CL	FMG	SPC	SNN	K-Means
			ARI	ARI	ARI	ARI	ARI	ARI	ARI
1	Leukemia-V1 [18]	1081	-0.01	0.21	0.18	0.27	**0.78**	0.29	0.27
2	Leukemia-V2[18]	2194	-0.01	0.54	0.49	**0.88**	**0.88**	0.77	0.37
3	Lung Cancers-V1 [19]	1543	-0.01	0.33	0.33	0.26	0.27	0.35	**0.42**
4	Lung Cancers-V2 [20]	1626	-0.01	-0.04	**0.92**	-0.05	0.05	0.72	0.85
5	Human Liver Cancers [22]	85	0.00	0.00	-0.01	**0.73**	0.04	0.47	0.42
6	Breast, Colon Cancers [21]	182	0.02	0.78	**0.92**	0.07	**0.92**	0.78	0.42
7	Colon Cancers [23]	2202	-0.04	0.08	-0.02	**0.46**	0.02	0.10	0.24
8	Prostate Cancers -V1 [24]	1288	0.01	0.04	0.23	0.26	0.18	0.09	**0.4**
9	Prostate Cancers -V2 [25]	2315	0.01	0.01	0.32	0.36	0.07	0.26	**0.48**
10	Bone marrow-V1 [27]	2526	-0.01	-0.01	-0.08	**0.96**	0.21	0.35	0.52
11	Bone marrow-v2 [27]	2526	0.00	0.19	0.27	0.36	0.23	0.20	0.37
12	Bladder Cancers [26]	1203	-0.06	0.11	0.11	**0.65**	0.40	0.25	0.15
Mean			-0.01	0.19	0.30	0.43	0.34	0.39	0.41
STD			0.02	0.25	0.33	0.31	0.34	0.25	0.17

We performed an ANOVA analysis for Tab.5 as follows:

Table 6: Summary of Anova: Singer Factor for Tab.5

Groups	Count	Sum	Average	Variance
SL	12	-0.11	-0.00917	0.00048
AL	12	2.24	0.18667	0.06299
CL	12	3.66	0.305	0.10955
FMG	12	5.21	0.43417	0.09779
SPC	12	4.05	0.3375	0.11218
SNN	12	4.63	0.38583	0.06104
K-Means	12	4.91	0.40917	0.03003

Table 7: Anova Analysis for Tab.5

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.76131	6	0.29355	4.33453	0.00081	2.21882
Within Groups	5.21476	77	0.06772			
Total	6.97607	83				

Conclusion: In Tab.7, if $F > Fcrit$, we reject the null hypothesis. This is the case $4.335 > 2.219$. Therefore, we reject the null hypothesis. The means of the seven populations are not all equal. At least one of the means is different.

Table 8: Clustering results with ARI values of the CL, FMG, SPC, K-Means, FCM and FPCM performed on the datasets with feature selection in [13] and GrFPCM performed on the original datasets

N.O.	Datatsets	FS [13]	CL	FMG	SPC	K-Means	FCM	FPCM	GrFPCM	
			ARI	ARI	ARI	ARI	ARI	ARI	GrFS	ARI
1	Leukemia-V1 [18]	1081	0.18	0.27	0.78	0.27	0.32	0.38	34	**0.89**
2	Leukemia-V2[18]	2194	0.49	0.88	0.88	0.37	0.37	0.54	150	**1**
3	Lung Cancers-V1 [19]	1543	0.33	0.26	0.27	0.42	0.25	0.34	512	**0.45**
4	Lung Cancers-V2 [20]	1626	0.92	-0.05	0.05	0.85	0.95	0.95	93	**1**
5	Human Liver Cancers [22]	85	-0.01	**0.73**	0.04	0.42	0.4	0.42	80	0.59
6	Breast, Colon Cancers [21]	182	**0.92**	0.07	**0.92**	0.42	0.53	0.71	22	0.89
7	Colon Cancers [23]	2202	-0.02	**0.46**	0.02	0.24	0.17	0.25	11	0.37
8	Prostate Cancers -V1 [24]	1288	0.23	0.26	0.18	0.4	0.32	0.38	60	**0.52**
9	Prostate Cancers -V2 [25]	2315	0.32	0.36	0.07	0.48	0.51	0.31	216	**0.62**
10	Bone marrow-V1 [27]	2526	-0.08	**0.96**	0.21	0.52	0.53	0.61	216	0.88
11	Bone marrow-v2 [27]	2526	0.27	0.36	0.23	0.37	0.41	0.36	186	**0.63**
12	Bladder Cancers [26]	1203	0.11	**0.65**	0.40	0.15	0.36	0.45	79	0.63
Mean			0.30	0.43	0.34	0.41	0.43	0.48	0.71	
STD			0.33	0.31	0.34	0.17	0.20	0.20	0.22	

We performed an ANOVA analysis for Tab.8 as follows:

Conclusion: In Tab.10, if $F > Fcrit$, we reject the null hypothesis. This is the case $2.99 > 2.22$. Therefore, we reject the null hypothesis. The means of the seven populations are not all equal. At least one of the means is different.

Table 9: Summary of Anova: Singer Factor for Tab.8

Groups	Count	Sum	Average	Variance
CL	12	3.66	0.305	0.10955
FMG	12	5.21	0.43417	0.09779
SPC	12	4.05	0.3375	0.11218
K-Means	12	4.91	0.40917	0.03003
FCM	12	5.12	0.42667	0.03915
FPCM	12	5.7	0.475	0.03948
GrFPCM	12	8.47	0.70583	0.04694

Table 10: Anova Analysis for Tab.8

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.22113	6	0.20352	2.99848	0.01097	2.21882
Within Groups	5.22637	77	0.06787			
Total	6.44750	83				

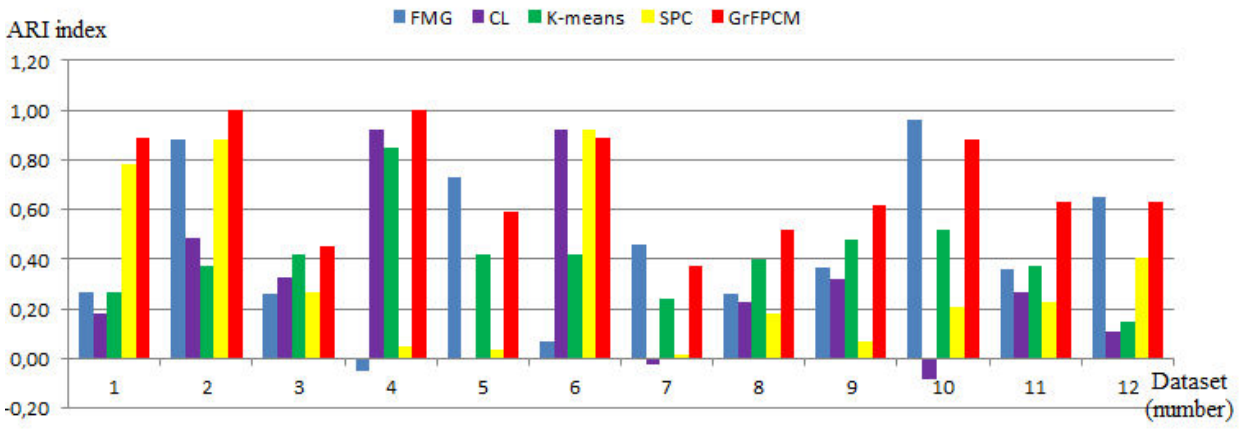

Figure 5: ARI values with feature selection for FMG, CL, K-Means, SPC, GrFPCM

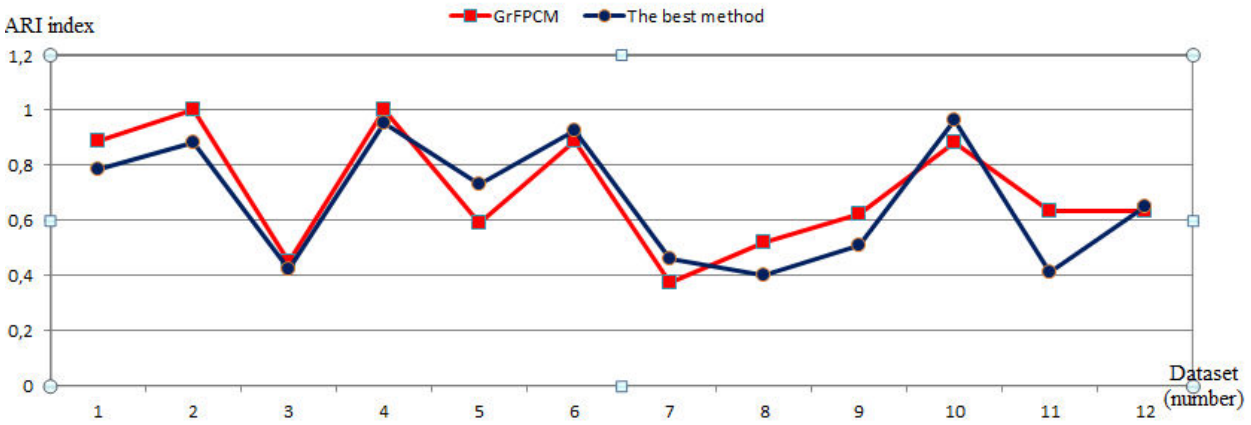

Figure 6: ARI values for the best comparison algorithms and GrFPCM

Fig.5 visually shows the ARI values in Tab.5 and Tab.8 among algorithms (with the largest ARI indexes) including FMG, K-Means, SPC, CL and GrFPCM. In Fig.5, GrFPCM still shows superior to the rest algorithms with the highest ARI values in 6 over 12 datasets. Fig.6 clearly shows the dominance of GrFPCM when compared to the best values of the remaining algorithms.

In Tab.5 and Tab.8, we noticed that the proposed algorithm (GrFPCM) outperformed the other algorithms with the highest ARI values. It even has the absolute ARI values which reach to 1 in some cases. Namely, Tab. 5 shows ARI values of seven algorithms for all twelve datasets. Although the results are different among datasets, the FMG, K-Means, SPC and CL produce the highest ARI values when running on 5, 3, 3 and 2 datasets respectively. Also, these best algorithms are selected for comparison presented in Tab.8.

In Tab.8, the GrFPCM obtains the best ARI values when running on 7 datasets, followed FMG with the largest ARI values when running on 4 datasets, among five considered algorithms. In addition, the mean of ARI values produced by the GrFPCM is 0.71 while by the FMG is only 0.43. Fig.7 visually represents the results coming from FMG and GrFPCM algorithms over 12 datasets.

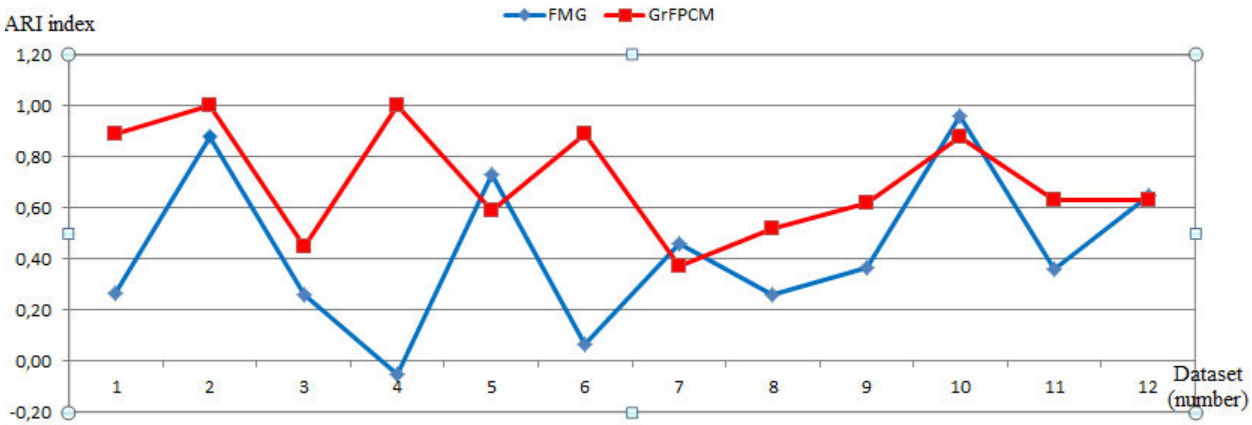

Figure 7: ARI values for FMG and GrFPCM algorithms

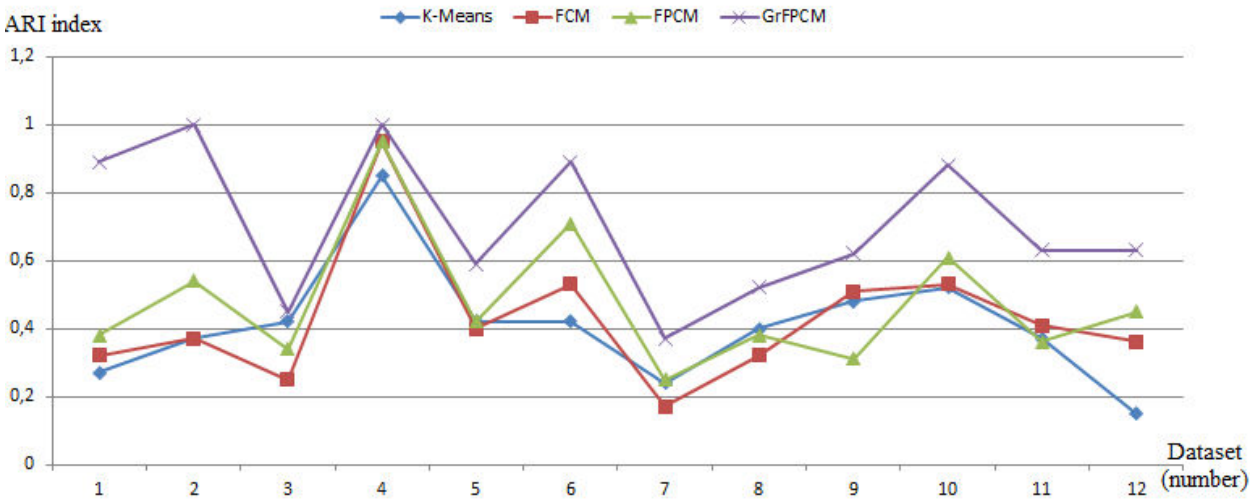

Figure 8: ARI index values for K-Means, FCM, FPCM and GrFPCM

Fig.5, Fig.6, Fig.7 and Fig.8 plotted the clustering results (ARI index values) obtained from the K-Means, FMG, CL, SPC, FCM, FPCM and GrFPCM in Tab.5 and Tab.8. The ARI values were calculated based on the clustering results coming from the K-Means, FMG, SPC, CL, SNN, FCM, FPCM and GrFPCM, where GrFPCM was done on the original datasets and others were done on the datasets with feature selection [13]. The proposed algorithm (GrFPCM) attained the best results (highest ARI index values) in almost experimental datasets.

5. Conclusions

In this study, we have presented an advanced Fuzzy Possibilistic C-Means clustering method based on concepts of Granular Computing, which can reduce feature space to produce a set of essential features, while eliminating those of marginal relevance. The proposed method takes advantage of the fuzzy possibilistic memberships in which a possibilistic membership is used to quantify a degree of typicality of a point belonging to a certain cluster and a membership is used to deal with the vague values. In addition, GrPFCM also endowed with ideas of GrC to becomes beneficial when coping with the uncertainty factors and to utilize feature selection for clustering to alleviate the negative impact of high dimensionality of the problems. The experiments completed for a number of well-known datasets demonstrate that the proposed method shows the better clustering results than other compared methods such as FMG, FCM, FPCM, K-Means, CL and SPC through two indexes IC and ARI.

In terms of future developments, it would be advantageous to involve more advanced methods (say, evolutionary optimization) to optimize the parameters of the clustering method. Besides, one may focus on using the concepts of Granular Computing to develop an advanced type-2 Fuzzy Possibilistic C-Means clustering method. The complexity of type-2 membership functions can be handled by information granules. Thus, this method can be used to increase performance of the traditional type-2 clustering algorithms by reducing the computational complexity to solve the real applications with high level of uncertainty.

Acknowledgements

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2016.09.

References

References

[1] F. Jaziri, E. Peyretaillade, P. Peyret, D.R.C. Hill, High performance computing of oligopeptides complete back-translation applied to DNA microarray probe design, Concurrency and Computation-Practice and Experience, vol.28(7), pp.2073-2091, 2016.

[2] H. Chen,Y. Zhang,I. Gutman, A kernel-based clustering method for gene selection with gene expression data, Journal of Biomedical Informatics, vol.62, pp.12-20, 2016.

[3] J. Sun, W. Chen, W. Fang, X. Wun, W. Xu, Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization, Engineering Applications of Artificial Intelligence, vol.25, pp.376-391, 2012.

[4] A. Mukhopadhyay, U. Maulik, Towards improving fuzzy clustering using support vector machine: Application to gene expression data, Pattern Recognition, vol.42, pp.2744-2763, 2009.

[5] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics, vol.20(15), pp.2429-2437, 2004.

[6] M. Vimaladev, B. Kalaavathi, A microarray gene expression data classification using hybrid back propagation neural network, Genetika, vol.46(3), pp.1013-1026, 2014.

[7] L. Shen, E.C. Tan, Dimension Reduction Based Penalized Logistic Regression For Cancer Classification Using Microarray Data, IEEE/ACM Trans. Computational Biology and Bioinformatics, vol.2(2), pp.166-175, 2005.

[8] C.H. Chen, Comparing batch update with randomized update for identifying salient genes applied to cancer gene expression clustering, Journal of information science, vol.40(6), pp.835-845, 2014.

[9] T. Hastie, R. Tibshirani,M. B Eisen, A. Alizadeh, R. Levy, L. Staudt, W. Chan, D. Botstein, and P. Brown, Gene Shaving as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns, Genome Biology, vol. 1 (2), pp. 1-21, 2000.

[10] K.J. Kim, S.B. Cho, Meta-classifiers for high-dimensional, small sample classification for gene expression analysis, Pattern analysis and applications, vol.18(3), pp.553-569, 2015.

[11] W. S. Kah, K. Moorthy, M.S. Mohamad, S. Kasim, S.Deris, S.Omatu, M. Yoshioka, Biological analysis of microarray data using orthogonal forward selection with a clustering approach, Journal of biological systems, vol.23(2), pp. 275-288, 2015.

[12] T. Wang, T.J. Li, G. F. Shao, S.X. Wu, An improved K-means clustering method for cDNA microarray image segmentation, Genetics and molecular research,vol.14(3), pp.7771-7781,2015.

[13] M. Souto, I. G Costa, D. Araujo ,T. B Ludermir and A. Schliep, Clustering cancer gene expression data: a comparative study, BMC Bioinformatics, vol. 9(1), pp. 1-14, 2008.

[14] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, vol.66 (336), pp. 846850, 1971.

[15] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification, vol.2, pp.193218, 1985.

[16] D. Jiang, C. Tang and A.Zhang, Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, vol. 16(11), pp. 1370-1386, 2004.

[17] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, IProc. Natl. Acad. Sci. USA, vol. 95, pp. 14863-14868, 1998.

[18] SA Armstrong et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nature Genetics, vol. 30, pp. 41-47, 2002.

[19] A Bhattacharjee et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc Natl Acad Sci USA, vol. 98(24), pp. 13790-13795, 2001.

[20] GJ Gordon et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Res, vol. 62(17), pp.4963-4967, 2002.
[21] D Chowdary et al., Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative, J Mol Diagn, vol.8, pp. 31-39, 2006.
[22] X Chen et al., Gene Expression Patterns in Human Liver Cancers, Mol Biol Cell, vol.13(6), pp.1929-1939, 2002.
[23] P Laiho et al., Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis, Oncogene, vol. 26(2), pp. 312-320, 2007.
[24] J Lapointe et al., Gene expression profiling identifies clinically relevant subtypes of prostate cancer, Proc Natl Acad Sci USA, vol. 101(3), pp.811-816, 2004.
[25] SA Tomlins et al., Integrative molecular concept modeling of prostate cancer progression, Nature Genetics, vol.3, pp. 41-51, 2007.
[26] L Dyrskjot et al., Identifying distinct classes of bladder carcinoma using microarrays, Nature Genetics, vol. 33, pp. 90-96, 2003.
[27] EJ Yeoh et al., Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell, vol. 1(2), pp.133-143, 2002.
[28] Z. Zhu, Y. S. Ong and M. Dash, Markov Blanket-Embedded Genetic Algorithm for Gene Selection, Pattern Recognition, vol. 49(11), pp. 3236-3248, 2007.
[29] T. Li, C. Zhang and M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics, vol.20, pp.2429-2437, 2004.
[30] J.-S Zhang and Y.-W. Leung, Improved Possibilistic C-Means Clustering Algorithms, IEEE Trans. on Fuzzy Systems, vol. 12(2),pp.209-217, 2004.
[31] N.R. Pal, K. Pal, and J.C. Bezdek, A mixed c-means clustering model, in Proc. IEEE Int. Conf. Fuzzy Systems, pp. 11-21, 1997.
[32] R. Krishnapuram and J. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst., vol. 1, pp. 98-110, 1993.
[33] J. C. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms, New York: Academic, 1981.
[34] S. Ding, H. Huang, J. Yu, Research on the hybrid models of granular computing and support vector machine, Artificial Intelligence Review, vol.43(4), pp.565-577, 2015.
[35] L. Sun, J. C. Xu, and Y. Tian, Feature Selection Using Rough Entropy-Based Uncertainty Measures in Incomplete Decision Systems, Knowledge Based Systems, vol.36, pp.206-216, 2012.
[36] Q. H. Hu, J. F. Liu, and D. R. Yu, Mixed Feature Selection Based on Granulation and Approximation, Knowledge-Based System, vol.21, pp.294-304, 2008.
[37] Y. Qian, Y. Li, J. Liang, Fuzzy Granular Structure Distance, IEEE Trans. on Fuzzy Systems, vol.23(6), pp. 2245-2259,2015.
[38] J. Qian, L. Ping, X. Yue, C. Liu, Hierarchical attribute reduction algorithms for big data using Map Reduce, Knowledge-based Systems, vol.73, pp.18-31, 2015.
[39] W. Pedrycz, From fuzzy data analysis and fuzzy regression to granular fuzzy data analysis, Fuzzy Sets and Systems, vol.274, pp.12-17,2015.
[40] L. Sun, J. Xu, Y. Hu, and L. Du Granular Space-Based Feature Selection and Its Applications, Journal of Software, vol. 8(4),pp.817-826,2013.
[41] A.Bargiela, W.Pedrycz, Granular Computing. An introduction, Kluwer Academic Publishers, 2003.
[42] L.A. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logics, Fuzzy Sets and Systems, vol.19, pp. 111-127, 1997.
[43] H. Runxin and H. Nian, The Reduction of Facial Feature Based on Granular Computing, Electronics and Signal Processing, LNEE 97, pp. 1015-1021, 2011.
[44] M. B. Ferraro, P. Giordani, Possibilistic and fuzzy clustering methods for robust analysis of non-precise data, International Journal of Approximate Reasoning, vol. 88, 2338, 2017.