# Introduction

**1. Motivation.**

Noise elimination plays a significant role in solving problems of clustering. Fuzzy possibilistic clustering may be applied to outlier detection or noise removal. However, one of the disadvantages of this method is sensitivity to datasets that are large or high-dimensional or both. Recently, many researchers are interested in this issue, but there are no effective methods.

**2. Study Objectives**

The study objectives of the thesis include: A solution for the fuzzy possibilistic clustering to deal with the high dimensional and noisy datasets; A solution for the fuzzy possibilistic clustering to reduce the noise factor and the uncertainty of the massive data, increase the quality, and decrease the clustering time; And a solution for the fuzzy possibilistic clustering to improve the clustering efficiency and deal with the local optimization situation.

**3. Research subjects**

The research subjects consist of following problems: Several algorithms for fuzzy clustering and fuzzy possibilistic clustering; several methods of granular computing (GrC), granular gravitational forces (GGF), particle swarm optimization (PSO), and their application to clustering; the extensions of fuzzy possibilistic clustering algorithms by GrC, GGF, and PSO.

**4. Research scope**

Firstly, this research focuses mainly on fuzzy clustering methods, particularly the fuzzy possibilistic C-means (FPCM) and interval type-2 FPCM (IT2FPCM) methods, and then investigates methods of GrC, GGF, and PSO, to improve these clustering methods.

Secondly, the combination of the clustering methods with other methods (GrC, GGF, and PSO); the primary goal of this is to improve the quality of clustering results.

Finally, experimental results are obtained to demonstrate the performance of the proposed algorithms.

**5. Contributions**

This thesis proposes four main algorithms: 1) Propose the GrFPCM algorithm to cope with the uncertainty factors, address noises, and alleviate the negative impact of the high dimensionality. 2) Propose the GIT2FPCM algorithm to reduce the noise factor and uncertainty of the data and decrease the execution time. 3) Propose the method of combining the GIT2FPCM with the PSO (PGIT2FPCM) to optimize the objective function. 4) Propose the AGrIT2FPCM algorithm to improve the distance measurement between a granule and a centroid of the cluster.

**6. Thesis Outline**

This thesis is organized into four chapters, as follows: Chapter 1 discusses the major issues and theoretical background. Chapter 2 presents in detail the FPCM algorithm based on GrC. Chapter 3 presents in detail the IT2FPCM clustering based on GGF, or both GGF and PSO. Chapter 4 states the conclusions which present the contributions and some recommendation for future research.

# Chapter 1. Overview

## 1.1. Related works

### 1.1.1. Fuzzy clustering

### 1.1.2. Fuzzy Possibilistic C-means clustering

The possibilistic C-means (PCM) method determines a possibilistic membership which is used to quantify a degree of typicality of a point belonging to a specific cluster. FPCM uses the membership values as well as the typicality values of the PCM to produce a better clustering algorithm.

### 1.1.3. Type-2 Fuzzy Possibilistic C-means clustering

The FPCM algorithm is similar to most other type-1 fuzzy clustering algorithms, which do not address well the uncertainty of input data. Thus, this algorithm has been improved using type-2 fuzzy sets to develop the type-2 fuzzy possibilistic C-means algorithm.

### 1.1.4. Granular Computing

GrC has emerged as a powerful vehicle to construct and process information granules, which are formed by grouping objects based on their similarity, closeness, or proximity. It is used to handle complex problems, cope with massive amounts of data, capture uncertainty, and represent data with high dimensionality. GrC can solve problems in computational intelligence, and it is a basis for feature selection methods.

### 1.1.5. Particle Swarm Optimization

PSO is a resident-based optimization tool that can be used and applied easily to solve the problem of optimizing transformational functions. Therefore, PSO algorithms are widely used and are constantly being improved to further enhance the efficiency of the algorithm. More recently, researchers approached have used PSO to improve fuzzy clustering algorithms.

## 1.2. The limitations of related works and study objectives

The fuzzy possibilistic clustering may be applied to outlier detection or noise removal. However, one of the disadvantages of this method is sensitive to large or high dimensional datasets or both. Meanwhile, GrC is a powerful tool to study granulation for handling complex problems, uncertain information, big data, and high dimensional data. So GrC may be applied to create preprocessing steps for clustering. However, it has not been specifically applied to fuzzy possibilistic clustering algorithms. From these brief analyzes, the study objectives of the thesis include: 1) Improving the FPCM clustering based on GrC. It is also considered as an appropriate solution to clustering problem that deals with the high dimensional and noisy datasets. 2) Improving the IT2FPCM clustering based on the GGF. This solution reduces the noise factor and the uncertainty of the data, increases the quality of clustering, and decreases the clustering time. 3) Integrating PSO with the GIT2FPCM method to improve the clustering efficiency. Therefore, this direction is an appropriate solution for the clustering problem that deals with large and noisy datasets.

## 1.3. Theoretical Background

### 1.3.1. Similarity Measurement

The most commonly used similarity measures for continuous data: Euclidean distance, Manhattan distance, Minkowski distance, Mahalanobis distance. This thesis uses Euclidean measurement.

### 1.3.2. Cluster Evaluation

The evaluation indicators are usually classified into three types: the internal examination, external assessment, and a relative test. Both internal and external criteria are used in this thesis.

***1.3.3. Fuzzy Clustering***

***1.3.3.1. Fuzzy C-Means Clustering***

***1.3.3.2. Possibilistic C-Means Clustering***

***1.3.3.3. Fuzzy Possibilistic C-Means Clustering (Pal et al.)***

The objective function of this algorithm is defined as follows:

|  |  |
| --- | --- |
|   | (1.9) |

where ; and .

|  |  |
| --- | --- |
|   | (1.10) |
|   | (1.11) |

where .

|  |  |
| --- | --- |
|  | (1.12) |

**Algorithm 1.3** FPCM algorithm *( Pal et al.)*

1 Input: ,

2 Output: , , and .

3 Initialize , calculate and by using Eq.1.10, Eq.1.11.

4 **repeat:** Update , , and by using Eq.1.12, Eq.1.10, and Eq.1.11

**until:** ,

5 Assign data to the cluster if

### 1.3.3.4. Fuzzy Possibilistic C-Means Clustering (Zhang et al.)

The objective function of FPCM is formed as follows:

|  |  |
| --- | --- |
|   | (1.13) |
|   |  (1.14) |
| where  | (1.15) |
|   | (1.16) |
|   | (1.17) |
|   | (1.18) |

**Algorithm 1.4** FPCM algorithm (*Zhang et al.*)

1 Input: , , and error .

2 Output: , , and .

3 Execute FCM to find an initial and .

|  |
| --- |
| Compute as follows:  |

4 **repeat:** Update , , and by using Eq.1.16, Eq.1.17, and Eq.1.18.

Apply Eq.1.15 to compute .

**until:**

5 Assign data to the cluster if and .

### 1.3.3.5. Interval Type-2 Fuzzy Possibilistic C-Means Clustering

### The IT2FPCM is an extension of FPCM (Pal et al.): and; is in the range ; is in the range ; is in the interval , where:

|  |  |
| --- | --- |
|  | (1.19) |
|  | (1.20) |
|  | (1.21) |
|  | (1.22) |
|   | (1.23) |
|   | (1.24) |
| where |   |  |

**Algorithm 1.5** The interval type-2 fuzzy possibilistic C-Means clustering

1 **Input:** , , and *ε*.

2 **Outputs:** Cluster centroid matrix

3 Initialize the cluster centroids by random.

4 **Repeat:** Update , by using Eq.1.19-1.22.

Update , by using Eq.1.23, and Eq.1.24;

Reduce the type: Define by using Eq.1.9.

**Until** .

## 1.4. Summary

This chapter presented an overview of the fuzzy clustering methods and related research results, including fuzzy clustering, fuzzy possibilistic clustering, and interval type-2 fuzzy possibilistic clustering. Moreover, the techniques used to develop the hybrid fuzzy clustering methods have been introduced, including GrC, GGF, and PSO. The final section summarised the theoretical background of this thesis.

# Chapter 2. Fuzzy Possibilistic C-Means Clustering Based on Granular Computing

This chapter presents the main contents, including the GrC theory, the feature reduction method based on GrC, the granular space construction and feature selection method, and the FPCM algorithm based on GrC. The results in this chapter have been published in [II] and [III].

## 2.1. Granular Computing

Def.2.1 determines an indiscernibility relation among objects of X of a clustering system on a subset of features.

**Definition 2.1**. *A clustering system , ,where*

 *is called the value domain of feature . is the information function of the system, A subset of features, determines an indiscernibility relation .*

*Based on , is divided in to equivalence classes as .*

Suppose that . , such that , which is a set of feature values corresponding to *B*. The intention of an information granule , and the extension

A granule for a clustering system is defined as Def.2.2.

**Definition 2.2.** *Let be a clustering system. An information granule is defined as where refers to the intention of the information granule and represents the extension of the information granule.*

## The system granularity of *B*, denoted *,* defines the granularity of the clustering system on a subset of features.

|  |  |
| --- | --- |
|  | (2.1) |

## 2.2. Feature reduction based on granular computing

Def.2.4, Def.2.5, and Def.2.6 determine a reduction set of features:

**Definition 2.4.** *The significance degree of feature is defined as follows:*

|  |  |
| --- | --- |
|  | (2.2) |

*A larger value of takes, the indicates that “” is more important.*

**Definition 2.5.** *Given , the feature is called a redundant feature to A if the value of is equal to . The set of all necessary features is called the core of , denoted .*

**Definition 2.6.** *Given a clustering system and a set of features set is called a reduction. The set of all reductions of A is denoted by .*

**Algorithm 2.1** Feature reduction based on GrC

1. Input: .
2. Output: Set of features that is the minimum reduction of.
3. Step 1: Determine the core of features as follows:

 For each if then select into .

1. Step 2: Assign

 If then the termination criterion is satisfied.

 **repeat:** For each , calculate .

 Find the so that

 Add feature to the core:

 **until:** .

## 2.3. Granular space construction and feature selection

In this research, we propose a method of granule formation. The objects are clustered into clusters by the FPCM on each feature. The clusters are labelled by numbering them in ascending order (1, 2..). Therefore, a cluster label matrix is formed from the label of the object on the feature . The new information function

is denoted as = . From the values of a row in the , with , we can construct a granule , in which .

**Definition 2.7*.*** *A non-conflict granular space is formed by , in which , , and . Conversely, a conflict granular space is formed by .*

The significance of a feature only affects the .

**Algorithm 2.2** Granular space construction and feature selection

1. Input: , , , and .
2. Output: The minimum reduction of and
3. Step 1: Execute Algorithm 1.4 on each to form .
4. Step 2: Construct granular space

4.1 Initialize: ( is the index of the row of the , is the index set, and is the number of granules).

4.2 Remove the outlier objects: and if . Then, remove if .

4.3 **repeat**

4.3.1 ; **repeat**  **until**

4.3.2 Set to the set of values of the row

4.3.3 Find

 **if then**

4.3.4.1 **for** each :

4.3.4.2

4.3.4.3 **if then**

 **else**

**until**

5 Step 3: Apply Algorithm 2.1 on GrSN to reach the minimum reduction *C*.

## 2.4. Fuzzy possibilistic C-means clustering based on GrC

We consider with . The value interval of the feature of is :

|  |  |
| --- | --- |
|  | (2.5) |
|  | (2.6) |

where is the value of object on the feature.

The distance between and the centroid :

|  |  |
| --- | --- |
|  | (2.7) |

where

|  |  |
| --- | --- |
|  | (2.8) |
|  | (2.9) |
|  | (2.10) |
|  | (2.11) |

where = and = with .

**Algorithm 2.3** Fuzzy possibilistic C-meansclustering based on GrC

1. Input: , , , and noise filter parameter .
2. Output: , , and .
3. Step 1: Apply Algorithm 2.2 on to obtain , .
4. Step 2: Apply Algorithm 1.4 on , as follows:

4.1 The number of iterations: .

4.2 **repeat:** .

 Update by using Eq. 2.9.

 Remove the outlier or noisy granules:

.

 Update by using Eq. 2.10.

 Update by using Eq. 2.11.

 Apply Eq. 1.15 to compute .

**until:** ,

5 Assign to the cluster if and .

## 2.5. Time complexity

The complexity of the GrFPCM depends on the FPCM and the granular space construction and feature selection: Form : construct a granular space: , apply Algorithm 2.1: . In addition, FPCM is applied on : . So the time complexity of GrFPCM is .

## 2.6. Experimental studies

### 2.6.1. Experiment 1

The well-known datasets WDBC, E. coli promoter gene sequences (DNA), and Madelon1 were used. The clustering results were evaluated by determining TPR and FPR, in Table 2.6.

Table 2.6: Clustering results for experiment 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | FCM | FPCM | GrFPCM |
| FS | TPR | FPR | FS | TPR | FPR | FS | TPR | FPR |
| WDBC | 30 | 89.5% | 4.5% | 30 | 92.7% | 2.8% | 4 | **95.4%** | **1.9%** |
| DNA | 57 | 85.6% | 6.7% | 57 | 91.4% | 3.1% | 2 | **96.1%** | **1.7%** |
| Madelon | 500 | 86.1% | 5.9% | 500 | 90.8% | 3.3% | 12 | **94.8%** | **2.1%** |

### 2.6.2. Experiment 2

Five public datasets (Lymphoma, Leukaemia, Global Cancer Map (GCM), Embryonal Tumours (ET), and Colon2) were used to illustrate the application of the proposed method to high-dimensional datasets.

Table 2.8: Clustering results for experiment 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | FCM | FCM(FS) | FPCM | FPCM(FS) | GrFPCM |
| TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR |
| Lymphoma | 89.2% | 4.6% | 89.9% | 4.2% | 89.8% | 3.1% | 93.2% | 1.8% | **96.1%** | **1.7%** |
| Leukaemia | 72.1% | 9.5% | 82.1% | 7.2% | 81.4% | 7.3% | 89.4% | 4.2% | **93.6%** | **1.4%** |
| GCM | 89.6% | 4.8% | 90.4% | 3.2% | 90.2% | 5.5% | 93.2% | 2.5% | **96.8%** | **1.2%** |
| ET | 80.1% | 9.1% | 87.6% | 6.3% | 88.1% | 7.6% | 91.1% | 4.6% | **95.3%** | **1.9%** |
| Colon1http://www.ics.uci.edu/ mlearn/mlrepository.html2http://www.upo.es/eps/bigs/datasets.html | 79.1% | 7.9% | 81.7% | 6.9% | 80.9% | 9.5% | 86.8% | 4.9% | **92.8%** | **3.4%** |

## Table 2.8 shows that the TPR values obtained by executing GrFPCM on the five datasets are greater than 92% and obviously higher than those obtained by the other methods. In addition, the FPR values are smaller than those achieved by the other algorithms. For FCM and FPCM, the TPR and FPR values obtained after reducing features are better than those obtained without reducing features.

## 2.7. Granular fuzzy possibilistic C-means clustering approach to DNA microarray problem

### 2.7.1. Cluster analysis for gene expression data

A gene expression dataset from a microarray experiment can be represented by a real-valued expression matrix , where rows represent genes, columns represent different samples, and the numbers in each cell represent the expression level of gene in sample . We consider the samples to be objects, and the genes to be features.

### 2.7.2. Results

Table 2.11: Clustering results with IC index values of the experimental datasets

|  |  |  |
| --- | --- | --- |
| No | Datasets | Incorrectly clustered instances (%) |
| K-means | FCM | FPCM | GrFPCM |
|  |  | N.I | % | N.I | % | N.I | % | N.I | % |
| 1 | Leukaemia-V1 | 22 | 30.5556 | 20 | 27.7777 | 18 | 25 | 2 | **2.7778** |
| 2 | Leukaemia-V2 | 21 | 29.1667 | 21 | 29.1667 | 15 | 20.8333 | 0 | **0** |
| 3 | Leukaemia-2c | 21 | 29.1667 | 20 | 27.7777 | 17 | 23.6111 | 2 | **2.7778** |
| 4 | Leukaemia-3c | 34 | 47.2222 | 18 | 25 | 13 | 18.0555 | 1 | **1.3889** |
| 5 | Leukaemia-4c | 42 | 58.3333 | 22 | 30.5556 | 22 | 30.5556 | 15 | **20.8333** |
| 6 | Lung Cancers-V1 | 96 | 47.2906 | 95 | 46.798 | 61 | 30.0493 | 35 | **17.2413** |
| 7 | Lung Cancers-V2 | 30 | 16.5746 | 2 | 1.105 | 2 | 1.105 | 0 | **0** |
| 8 | Human Liver Cancers | 80 | 44.6927 | 89 | 49.7207 | 80 | 44.6927 | 22 | **12.2905** |
| 9 | Breast, Colon Cancers | 44 | 42.3077 | 15 | 14.431 | 8 | 7.6923 | 3 | **2.8846** |
| 10 | Breast Cancers | 45 | 46.3918 | 37 | 38.1443 | 29 | 29.8969 | 18 | **18.5567** |
| 11 | Colon Cancers | 14 | 37.8378 | 13 | 35.1351 | 13 | 35.1351 | 11 | **29.7297** |
| 12 | Prostate Cancers -V1 | 51 | 46.3636 | 63 | 57.2727 | 56 | 50.909 | 31 | **28.1818** |
| 13 | Prostate Cancers -V2 | 55 | 52.8846 | 40 | 38.4615 | 65 | 62.5 | 29 | **27.8846** |
| 14 | Bone marrow-V1 | 88 | 35.4839 | 87 | 35.0806 | 50 | 20.1613 | 6 | **2.4194** |
| 15 | Bone marrow-v2 | 169 | 68.1452 | 107 | 43.1452 | 170 | 68.5484 | 73 | **29.4354** |
| 16 | Ovarian | 112 | 44.2688 | 86 | 33.992 | 75 | 29.6442 | 2 | **0.7905** |
| 17 | Lymphoma | 22 | 33.3333 | 20 | 30.303 | 20 | 30.303 | 10 | **15.1515** |
| 18 | CNS | 29 | 48.3333 | 26 | 43.3333 | 19 | 31.6666 | 15 | **25** |
| 19 | SRBCT | 52 | 62.6506 | 27 | 32.5301 | 22 | 26.506 | 5 | **6.0241** |
| 20 | Bladder Cancers | 18 | 45 | 18 | 45 | 8 | 20 | 5 | **12.5** |

We performed clustering on the 20 gene expression datasets in Table 2.11 by executing K-means, FCM, FPCM, and GrFPCM.

The clustering results show that the GrFPCM is superior over all 20 datasets; in particular, the IC index values equa l0 with two datasets.

Additional, the datasets with different feature selection algorithms were compared to the datasets in which features were selected by the proposed algorithm. The ARI values of K-means, FMG, SNN, SL, CL, AL, and SPC methods were taken from reference, and ARI values were calculated for 12 datasets with FCM, FPCM, and GrFPCM.

Table 2.14: Clustering results with ARI values of the CL, FMG, SPC, K-means (KM), FCM, FPCM, and GrFPCM methods datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Datasets | FS | CL | FMG | SPC | KM | FCM | FPCM | GrFPCM |
| ARI | ARI | ARI | ARI | ARI | ARI | GrFS | ARI |
| 1 | Leukaemia-V1 | 1081 | 0.18 | 0.27 | 0.78 | 0.27 | 0.32 | 0.38 | 34 | **0.89** |
| 2 | Leukaemia-V2 | 2194 | 0.49 | 0.88 | 0.88 | 0.37 | 0.37 | 0.54 | 150 | **1** |
| 3 | Lung Cancers-V1 | 1543 | 0.33 | 0.26 | 0.27 | 0.42 | 0.25 | 0.34 | 512 | **0.45** |
| 4 | Lung Cancers-V2 | 1626 | 0.92 | -0.05 | 0.05 | 0.85 | 0.95 | 0.95 | 93 | **1** |
| 5 | Human Liver Cancers | 85 | -0.01 | **0.73** | 0.04 | 0.42 | 0.4 | 0.42 | 80 | 0.59 |
| 6 | Breast, Colon Cancers | 182 | **0.92** | 0.07 | **0.92** | 0.42 | 0.53 | 0.71 | 22 | 0.89 |
| 7 | Colon Cancers | 2202 | -0.02 | **0.46** | 0.02 | 0.24 | 0.17 | 0.25 | 11 | 0.37 |
| 8 | Prostate Cancers -V1 | 1288 | 0.23 | 0.26 | 0.18 | 0.4 | 0.32 | 0.38 | 60 | **0.52** |
| 9 | Prostate Cancers -V2 | 2315 | 0.32 | 0.36 | 0.07 | 0.48 | 0.51 | 0.31 | 216 | **0.62** |
| 10 | Bone marrow-V1 | 2526 | -0.08 | **0.96** | 0.21 | 0.52 | 0.53 | 0.61 | 216 | 0.88 |
| 11 | Bone marrow-v2 | 2526 | 0.27 | 0.36 | 0.23 | 0.37 | 0.41 | 0.36 | 186 | **0.63** |
| 12 | Bladder Cancers | 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 |
| Standard deviation | 0.33 | 0.31 | 0.34 | 0.17 | 0.20 | 0.20 | 0.22 |

Fig. 2.5: ARI values with feature selection for FMG, CL, K-means, SPC, and GrFPCM.



**2.8. Summary**

This chapter presented an advanced FPCM method based on GrC (GrFPCM), which can reduce the feature space to produce a set of essential features, and eliminate irrelevant features and noise objects. GrFPCM takes advantage of the fuzzy possibilistic memberships to deal with vague values. In addition, GrFPCM handles uncertainty factors efficiently and alleviates the negative impacts of the high dimensionality of problems. The experiments demonstrated that GrFPCM obtains better clustering results than other methods; in particular, this algorithm potentially enhances the clustering results when working with gene expression data.

# Chapter 3. Interval Type-2 Fuzzy Possibilistic C-Means Clustering Based on Granular Gravitational Forces and PSO

Chapter 2 presented some results of an improved FPCM algorithm based on feature noise reduction using GrC. This chapter presents some results, in the direction of reducing noise and uncertainty of data objects, by extending the IT2FPCM algorithm using GGF and PSO. The results in this chapter have been published in [I] and [IV].

## 3.1. Interval type-2 fuzzy possibilistic C-means clustering based on granular gravitational forces

### 3.1.1. Granular gravitational model

The law of universal gravitation states that two points interact with each other by a gravity force. This is the main idea of clustering algorithms based on granular gravity.

### 3.1.2. Gravitational granular space construction

# We present a method of advanced gravitational granular space construction, in which the granule grouping steps are improved. The IT2FPCM method is performed on the resulting granule set.

# The information granule is represented by (position), (mass), and (gravitational density of the granule)

**Algorithm 3.1** Group granules and initialize centroids

1 **Input**: , , and .

2 **Output:** , .

3 Step 1: 3.1 Assign number of initial granules .

 3.2 Initialize granules: .

4 Step 2: **repeat:**

4.1 Calculate all gravitational forces:

4.2 Calculate gravitational density: with .

4.3 Sort granules according to :.

4.4 Finding , nearest granule from :

4.4.1 Update the mass of : .

4.4.2 Determine gravitational centroid:

 .

4.4.3 Determine factor : .

4.4.4 Update position: .

4.4.5 and , .

 **until:** .

5 Step 3: Determine initial centroids:

5.1 Initializing granule set: with , .

5.2 **repeat:** Group the granule set according to step 2

 **until:** .

5.3 Determine with , .

***3.1.3. Interval type-2 fuzzy possibilistic C-means clustering based on granular gravitational forces***

GIT2FPCM performs the IT2FPCM algorithm on the granular space with input dataset .

|  |  |
| --- | --- |
|  | (3.9) |
|  | (3.10) |

where .

|  |  |
| --- | --- |
|  | (3.11) |
|  | (3.12) |
|  | (3.13) |
|  | (3.14) |

where and .

The type reduction is performed as follows:

|  |  |
| --- | --- |
|  | (3.15)  |
|  | (3.16) |
|  | (3.17) |

**Algorithm 3.2** IT2FPCM clustering based on GGF.

1 **Input**: and .

2 **Output**:

3 Step 1: Apply Algorithm 3.1 to obtain and

4 Step 2:

4.1 Update and using Eq. 3.9 and Eq. 3.10.

4.2 Update and using Eq. 3.11 and Eq. 3.12.

4.3 Update and according to Eq. 3.13 and Eq. 3.14.

4.4 Reduce type of , , and using Eq. 3.15-3.17.

4.5 Calculate the objective function as follows:

|  |  |
| --- | --- |
|  | (3.18) |

5 Step 3: Repeat step 2 until .

## 3.2. Interval type-2 fuzzy possibilistic C-means clustering based on granular gravitational forces and PSO

***3.2.1. Particle swarm optimization***

***3.2.2. Interval type-2 fuzzy possibilistic C-means clustering based on granular gravitational forcesand PSO***

From the set of initial clustering centroids, obtained from Algorithm 3.1, we randomly generate sets of initial cluster centroids. These sets of initial cluster centroids are considered as particles of the swarm. The best position of the particle corresponds to the set of cluster centroids for which the objective function value of particle is the smallest. Similarly, the best position of the swarm corresponds to the set of cluster centroids for which the objective function value of the swarm is the smallest.

**Algorithm 3.3** IT2FPCM clustering based on GGF and PSO.

1 **Input**: , , , and ; PSO parameters.

2 **Output**:

3 Step 1: Apply Algorithm 3.1 to obtain and .

 Set up a swarm of *N* particles: .

4 Step 2:

4.1 Determine the fuzzy partition matrix for each particle.

4.2 Determine the possibilistic partition matrix for each particle.

4.3 Determine for each particle based on the objective function.

4.4 Determine for the swarm based on the objective function.

4.5 Update the velocity matrix of each particle.

4.6 Update the position matrix of each particle.

4.7 If PSO termination criteria are not satisfied, return to step 2.

5 Step 3:

5.1 Apply GIT2FPCM (step 2, step 3) for each particle where the initial cluster centroid is .

5.2 Update for each particle.

5.3 Determine for the swarm.

6 Step 4: If PGIT2FPCM termination criteria are not satisfied, return to step 2.

**3.3. Interval type-2 fuzzy possibilistic C-means clustering based on advanced granular computing**

In this method, GrC is used to create granules for dimensionality reduction, and then the method of GGF is used to determine the centroids of granules, to improve the measurement of the distance between the granules and the cluster centroids.

### 3.3.1. Determine the centroids of granules based on granular gravitational forces

Firstly, the grouping process is repeated until only one point remains in each granule. Secondly, the position of the remaining point is considered to be the corresponding granule centroid.

**Algorithm 3.4** Determine the centroids of granules

1 Input: with .

2 Output: Set of centroids of granules .

3 **For to**

4 Step 1:

4.1 Assign number of initial points .

4.2 Initialize points in granule: and .

5 Step 2:

 **repeat:**

5.1 Calculate all gravitational forces in the granule:

 with ,

5.2 Calculate gravitational density for each point: .

5.3 Sort points in the granule according to in descending order.

5.4 Find , the nearest point from , and group points and .

5.4.1 Update mass of : .

5.4.2 Determine gravitational centroid:

.

5.4.3 Determine factor : .

5.4.4 Update position .

5.4.5 Update number of points in granule:

 **until:** .

6 Step 3: Centroid of granule: .

7 **Next.**

### 3.3.2. Interval type-2 fuzzy possibilistic C-means clustering based on advanced granular computing

In this section, we improve the granular space, which is the result of Algorithm 2.2. We then apply the IT2FPCM algorithm to clustering on the advanced granular space.

In the advanced granular space (AGS), the distance between a granule and the centroid of cluster is determined by Eq. 3.28.

|  |  |
| --- | --- |
|  | (3.28) |

in which is the result of Algorithm 3.4.

**Algorithm 3.5** AGrIT2FPCM

1 **Input**: , , , , error , and noise filter parameter .

2 **Output**: , , .

3 Step 1: Apply Algorithm 2.2 on to obtain the feature set , which is the minimum reduction of and the granular space .

4 Step 2: Apply Algorithm 3.4 on to obtain

5 Step 3: Apply the IT2FPCM on , as follows:

5.1 The number of iterations is set to

5.2 **repeat:**

5.2.1

5.2.2 Update .

5.2.3 Remove the outlier or noisy granules

5.2.4 Update .

5.2.5 Update .

 **until:**

6 Assign to the cluster if and .

**3.4. Time complexity**

The time complexity of calculating the granule grouping and centroid initialization: ; IT2FPCM: , and PSO: *.* Therefore, the time complexity of both GIT2FPCM and PGIT2FPCM are .

In addition, the time complexity of calculating the granular space construction and feature selection: , determining the centroids of granules: and IT2FPCM: . Thus, the time complexity of AGrIT2FPCM is

**3.5. Experiments**

### 3.5.1. Experiments with the GIT2FPCM and PGIT2FPCM

For a comparative analysis, several clustering methods were used, including FPCM, IT2FPCM, and GIT2FPCM, PGIT2FPCM which are the proposed algorithms. Validity indexes were used to evaluate the clustering results, including PC-I, CE-I, and XB-I. The execution time of the algorithms was also evaluated. The well-known datasets were considered as listed in Table 3.1.

Table 3.1: Characteristic of datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Number of samples | Number of features | Number of classes |
| Iris | 150 | 4 | 3 |
| Haberman’s Survival | 306 | 3 | 2 |
| Banknote Authentication | 1372 | 4 | 2 |
| HTRU2 | 17898 | 8 | 2 |

A larger PC-I index, or a smaller CE-I or XB-I index, indicates a better clustering result. Thus, from the results in Table 3.2, GIT2FPCM and PGIT2FPCM obtained better results than FPCM and IT2FPCM. The PGIT2FPCM algorithm achieved the best PC-I and CE-I indices on all datasets, and the GIT2FPCM algorithm achieved the best XB-I index on the majority of the datasets.

The synthesis results in Table 3.2 indicate that the execution time of the proposed algorithms was less, and hence they are more efficient than the existing algorithms. The average execution time of the GIT2FPCM algorithm was less than FPCM or IT2FPCM by a factor of several hundred; the execution speed of the PGIT2FPCM algorithm was also several times faster than FPCM or IT2FPCM. Moreover, the larger the dataset, the more efficient they were. By reducing the number of elements in the granular space, the proposed algorithms showed significantly better performance on large datasets.

Table 3.2: Validity indices and times from four clustering algorithms

|  |  |  |
| --- | --- | --- |
| Dataset | Index | Algorithm |
| FPCM | IT2FPCM | GIT2FPCM | PGIT2FPCM |
| Iris (150;4;3) | PC-I | 0.783167 | 0.783551 | 0.785561 | **0.785563** |
| CE-I | 0.395949 | 0.395634 | 0.393457 | **0.393455** |
| XB-I | 0.137148 | 0.134695 | **0.127773** | 0.127775 |
| T | 18.454 | 13.51 | **0.046** | 10.936 |
| Haberman’s Survival (306;3;2) | PC-I | 0.739771 | 0.742654 | 0.746108 | **0.747386** |
| CE-I | 0.413675 | 0.409478 | 0.404494 | **0.401771** |
| XB-I | 0.222922 | 0.210895 | **0.193094** | 0.194350 |
| T | 33.665 | 24.398 | **0.109** | 21.591 |
| Banknote Authentication (1372;4;2) | PC-I | 0.737558 | 0.737870 | 0.750121 | **0.750159** |
| CE-I | 0.418323 | 0.418043 | 0.400196 | **0.400193** |
| XB-I | 0.178647 | 0.178848 | 0.146360 | **0.146339** |
| T | 280.833 | 110.886 | **0.936** | 31.247 |
| HTRU2(17898;8;2) | PC-I | 0.812079 | 0.814396 | 0.872349 | **0.872517** |
| CE-I | 0.311227 | 0.308114 | 0.222784 | **0.222720** |
| XB-I | 0.168908 | 0.162629 | **0.070581** | 0.070638 |
| T | 4746.24 | 1971.81 | **82.962** | 393.029 |

Fig. 3.1: Validity indices and execution times obtained by four clustering algorithms.



### 3.5.2. Experiments with the AGrIT2FPCM algorithm

In this section, we also offer a comparative analysis of the clustering results using various clustering methods: FPCM, GrFPCM (FPCM perform on the granular space from Algorithm 2.2), and AGrIT2FPCM. Of these three, AGrIT2FPCM is the proposed method. The performance of the clustering algorithms was evaluated by TPR and FPR, which are defined in Eq. 2.12. While FPCM performs clustering on the original datasets with all features, GrFPCM perform clustering on the granular space which are the output of Algorithm 2.2, and AGrIT2FPCM perform clustering on the advanded granular space.

Table 3.5: Characteristics of datasets and feature selection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Number of samples | Number of features | Class | Feature Selection |
| WDBC1 | 569 | 30 | 2 | 4 |
| DNA1 | 106 | 57 | 2 | 2 |
| Madelon1 | 4400 | 500 | 2 | 12 |
| Lymphoma2 | 45 | 4026 | 2 | 15 |
| Leukaemia2 | 38 | 7129 | 2 | 6 |
| Global Cancer Map(GCM)2 | 190 | 16063 | 14 | 16 |
| Embryonal Tumours2 | 60 | 7129 | 2 | 8 |
| Colon2 | 62 | 2000 | 2 | 9 |

Table 3.6: TPR and FPR for FPCM, GrFPCM, and AGrIT2FPCM.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | FPCM | GrFPCM | AGrIT2FPCM |
| FS | TPR | FPR | FS | TPR | FPR | FS | TPR | FPR |
| WDBC | 30 | 92.6% | 2.8% | 4 | 95.4% | 1.9% | 4 | **96.1%** | **1.6%** |
| DNA | 57 | 91.5% | 2.80% | 2 | 96.20% | **1.90%** | 2 | **97.20%** | **1.90%** |
| Madelon | 500 | 90.8% | 3.30% | 12 | 94.80% | 2.10% | 12 | **95.80%** | **1.90%** |
| Lymphoma | 4026 | 88.9% | 2.20% | 15 | **95.60%** | **2.20%** | 15 | **95.60%** | **2.20%** |
| Leukaemia | 7129 | 81.6% | 7.90% | 6 | 94.70% | **2.60%** | 6 | **97.40%** | **2.60%** |
| Global Cancer Map | 16063 | 90.0% | 5.30% | 16 | 96.80% | **1.10%** | 16 | **97.90%** | **1.10%** |
| Embryonal Tumours | 7129 | 88.3% | 8.30% | 8 | 95.00% | **1.70%** | 8 | **96.70%** | **1.70%** |
| Colon1http://www.ics.uci.edu/ mlearn/mlrepository.html2http://www.upo.es/eps/bigs/datasets.html | 2000 | 80.6% | 9.70% | 9 | 93.50% | **3.20%** | 9 | **95.20%** | **3.20%** |

## The clustering results (the classification quality) are reported in terms of the TPR and FPR indices, and are shown in Table 3.6. A higher TPR value and lower FPR value indicate a better algorithm. Table 3.6 shows that the TPR values obtained by AGrIT2FPCM are greater than 95% and obviously higher than those obtained by other methods. Additionally, the FPR values are smaller than those achieved by other algorithms.

## 3.6. Summary

This chapter presented advanced IT2FPCM clustering based on GGF and PSO. Granules are constructed from the original data objects, by grouping based on GGF, to create a dataset of granules. We then proposed the GIT2FPCM algorithm, which performs IT2FPCM clustering on the set of granules. This method also reduces the noise factor and uncertainty of the data, thereby increasing the quality of the clustering. Further, the clustering time decreases significantly, as a consequence of the reduced dataset size. Moreover, we proposed a method of combining the GIT2FPCM algorithm with the PSO algorithm to optimize the objective function and improve the quality of clustering. Additionally, the AGrIT2FPCM algorithm was presented in this chapter. Based on GGF, this algorithm determines the centroids of granules to improve the measurement of the distance between the granules and the centroids of the cluster. This algorithm also utilizes the advantages of IT2FPCM for processing uncertainty and noisy data.

**Chapter 4**

# 4 Conclusions

The main contributions of this thesis are summarized in this section.

1. We proposed GrFPCM, an algorithm for the FPCM based on GrC. This algorithm constructs the granular space to eliminate the effects of irrelevant features and noise objects. FPCM is then executed on the granular space. Therefore, the GrFPCM copes with the uncertainty factors and alleviate the negative impact of the high dimensionality of problems. The DNA microarray problem was presented, as an application of the GrFPCM. The results demonstrated that GrFPCM achieves better results than some other existing clustering methods.

2. We proposed GIT2FPCM, an algorithm for the IT2FPCM clustering based on GGF. This algorithm is to construct granules from the original data objects by grouping based on GGF. The IT2FPCM clustering method is executed on the set of granules and the initial cluster centroids. This method reduces the noise factor and uncertainty of the data, thereby increasing the quality of the clustering. Furthermore, the clustering execution time decreases significantly, as a consequence of the reduced dataset size.

3. We proposed a method of combining GIT2FPCM with the PSO, to optimize the objective function and enhance clustering efficiency. This method considers sets of initial cluster centroids, obtained from GIT2FPCM, as the particles of the swarm. A combination of updating the positions of the centroids and updating the positions of the particles is performed, to determine the best positions of the centroids.

4. We proposed AGrIT2FPCM, an algorithm for the IT2FPCM clustering based on advanced GrC. Based on GGF, this algorithm determines the centroids of the granules that are created by the GrFPCM, by improving the measurement of distance between the granules and centroids of the clusters. Further, this algorithm utilizes the advantages of IT2FPCM in processing uncertainty and noisy datasets.

# Publications

[I]. Trương Quốc Hùng, Ngô Thành Long, Phạm Thế Long; Phân cụm C-Means khả năng mờ loại hai khoảng dựa trên hạt giảm chiều cải tiến; Tạp chí Nghiên cứu Khoa học và Công nghệ Quân sự; Số 59; 2019.

[II]. Hung Quoc Truong, Long Thanh Ngo, Witold Pedrycz; Advanced Fuzzy Possibilistic C-Means Clustering Based on Granular Computing; IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2016 (DOI: [10.1109/SMC.2016.7844627](https://doi.org/10.1109/SMC.2016.7844627)).

[III]. Hung Quoc Truong, Long Thanh Ngo, Witold Pedrycz; Granular Fuzzy Possibilistic C-Means Clustering Approach to DNA Microarray Problem; Knowledge-Based Systems; 133; pp.53-65; 2017 (SCI - Q1 ranking DOI:10.1016/j.knosys.2017.06.019 - IF: 4.396).

[IV]. Hung Quoc Truong, Long Thanh Ngo, Long The Pham; Interval Type-2 Fuzzy Possibilistic C-Means Clustering Based on Granular Gravitational Forces and Particle Swarm Optimization; Journal of Advanced Computational Intelligence and Intelligent Informatics; 23(3); pp.592-601; 2019 (Scopus - Q3 ranking).