DOI: https://doi.org/10.59294/HIUJS.VOL.6.2024.632

# A new approach for data clustering based on granular computing

**Truong Quoc Hung[*], Nguyen Huy Liem, Vu Minh Hoang,**
**Tran Thi Hai Anh and Nguyen Thi Lan**
*Institute of Simulation Technology, Le Quy Don University, Vietnam*

**ABSTRACT**

*This paper introduces a new clustering technique based on granular computing. In traditional clustering algorithms, the integration of the high shaping capability of the existing datasets becomes fussy which in turn results in inferior functioning. Furthermore, the laid-out technique will be able to avoid these challenges through the use of granular computing to bring in a more accurate and prompt clustering process. The creation of a novel algorithm hinges on utilizing granules, which are the information chunks that reveal a natural structure as part of the data and also help with natural clustering. A testing of the algorithm's features is carried out by using state-of-the-art datasets and then an algorithm's effectiveness is compared to the other clustering methods. The results of the experiment show significant improvement in clustering accuracy and reduction in data analysis time, thus testifying how granular computing is efficient in data analysis. This quest is not only going to serve as a reinforcement in data clustering, but it will also probably be an input in the broader area of unsupervised learning, reinforcing positions for scalable and interpretable solutions for data-driven decision-making.*

***Keywords:*** *data clustering, clustering of information, granular Computing, information granule, unsupervised learning, accuracy of the algorithm*

## 1. INTRODUCTION

Fuzzy clustering algorithms were developed to handle uncertain or imprecise information. The Fuzzy Possibilistic C-means (FPCM) method can be utilized to identify outliers or eliminate noise [1]. However, clustering problems often involve large and high-dimensional datasets, which present challenges in extracting useful information from these datasets [2]. Most clustering algorithms, including the FPCM algorithm, are generally sensitive to large amounts of data.

Data clustering is one of the major areas that has gained a lot because of the huge progress being noticed in the areas of granular computing, whereby granular computing is the current frontier in clustering development [3]. A concept of segmental computing, where information blocks in its structure give a different paradigm for structuring and analyzing data as compared to the traditional one. The existing research, as the systematic studies so to speak, has developed a vast background for microorganisms and has been helpful for different types of tasks. In addition, the complicated ties arising from the limitations of the conducted research which are, first, the indefinite nature of scalability, and secondly, the uncertainty towards the final results have been revealed; however, the proposed study is purposed to address all these.

Many heuristic algorithms deal with high-dimensional datasets by removing noise and redundant features (also known as feature selection). However, these algorithms need labeled samples as training samples to select the necessary features. Therefore, they are not suitable for clustering problems. Granular computing (GrC) is a general computation theory for effectively using granules (such as classes, clusters, subsets, groups, and intervals) to construct an efficient computational model for complex applications with vast amounts of data,

*Corresponding author: Dr. Truong Quoc Hung*
*Email: truongqhung@gmail.com*

information, and knowledge. GrC is also one of the ways to deal with feature selection problems. In addition, GrC may be used to construct a granular space containing a granule set that is smaller than the original one but continues to represent the original dataset. Thus, the size of the dataset is reduced, so that clustering problems with large and high-dimensional datasets can be solved more effectively. Therefore, hybrid models between GrC and fuzzy clustering can improve the clustering results [4-6]. Recently, the idea of granular gravitational forces (GGF) to group data points into granules and then process clusters on a granular space has been proposed [7, 8]. In this method, the size and noise of the original dataset are reduced, and the initial cluster centroids are determined.

Datasets have become both tangled in complexity and massive in size, so the major motive for current work is to develop clustering algorithms that can avoid the colossal impact of this. The authors present the new theories and methods that bring rare viewpoints to existing schemes and which as a result allows one to see more clearly the clustering methods. This is most significant, especially, in big data or machine learning which are where analysis of data and comprehensibility is taken as of the highest priority.

The organized outlook of this paper is well-crafted as it begins with theoretical foundations of how granular computing will help you in data clustering (in section 2). After the previous two sections, the method that combines the justified granularity principle into a new clustering algorithm is presented and discussed in the implementation and evaluation (Section 3). The next segments will cover all phases of the current experiment and the final outcomes deduced from the research conducted.

## 2. PROBLEM

### 2.1. Theoretical Foundations

The proposal of the clustering method foundations on granular computing (GC), which tool enables data analysis by creating information granules instead of data volumes. GC does its job using data encapsulation, which helps to decouple the data from the physical form it has and allows operations on it to be more inexpensive to perform.

### *2.1.1. Assumption*

This research is in that granules, movements of natural systems which are apparent on national scales but manifest only in the analysis of local data clusters, are sufficient to understand the natural systems in question. These sphere-shaped kernels are supposed to be achieved with an even granulation throughout the entire archival data so it can provide a uniform measurement of granularity throughout the dataset.

### *2.1.2. Formulas*

Granularity Coefficient (GC): The Grain Size Coefficient is a quantitative measurement that specifies how varying the granularity is within a dataset. The granularity is computed by dividing the sum of the granules by the data point's total number of data points. Formula: *GC = g/n*, where g is the number of granules, and n is the number of samples.

Bigger GC specification shows that data was divided into smaller-grained groups with a higher number of their rations, which also means that clustering was done more accurately. On the contrary, the low GC value is a hint that the distribution has a wider granularity and a bigger size, which means that the groups are fewer and larger in those clusters [9]. The Granulation Radius is the area occupied by a granule in the whole set. It indicates the dimension of the granules when the spatial relationships of the data points are considered as another main factor.

*a. The process of Generating Granular Space*

Some definitions which were proposed in [10], are introduced to granulate the clustering system as follows:

*Definition 1: Granular Space Computing*

Given dataset $X = \{x_i, x_i \in R^d\}, i = 1, 2, ..., n,$ where $n$ is the number of samples on $X$. The granular space $G = \{G1, G2, ..., G_g\}$ is used to cover and represent the data set $X$. The $G_j$ coverage degree is determined as $\sum_{j=1}^{g}(|G_j|)/n$, where $|G_j|$ is the number of samples in $G_j$.

The basic model of granular space coverage can be expressed as:

$$min\ \beta_1 * \frac{n}{\sum_{j=1}^{g}\frac{|G_j|}{n}} + \beta_2 * g \quad (1)$$

When other factors remain unchanged, the higher the coverage, the less the sample information is lost, and the more the number of granules, the characterization is more accurate. Therefore, the minimum number of granules should be considered to obtain the maximum coverage degree when generating granules. In most cases, $\beta_1$ and $\beta_2$ are set to 1 by default.

*Definition 2:* The Process of Generating Granular Space

- For each $G_j$ , the $\theta_j$ is the center of $G_j$ and $r_j$ is the radius of $G_j$ . The definitions of $\theta_j$ and $r_j$ are as follows:

$$\boldsymbol{\theta_j = (1/|G_j|)\sum_{i=1}^{|G_j|} x_i} \quad (2)$$

$$\boldsymbol{r_j = max(\|x_i - \theta_j\|)} \quad (3)$$

- Distribution Measure $DM_j$ is defined as follows:

$$\boldsymbol{DM_j = \frac{s_j}{|G_j|}} \quad (4)$$

where $\boldsymbol{s_j = \sum_{i=1}^{|Gj|} \| x_i - \theta_j \|}$ is the sum radius in $G_j$.

- We treat the whole dataset as a granular space *O*. Suppose that $O_k$ are sub-granules of *O* and both $DM_{O1}$ and $DM_{O2}$ are smaller than $DM_{O1}$ then *O* was split into $O_1$ and $O_2$. The $DM_W$ (weighted *DM* value) can better adapt to noisy data. It is defined as follows:

$$\boldsymbol{DM_w = \frac{|O_1|}{|O|} DM_{O_1} + \frac{|O_2|}{|O|} DM_{O_2}} \quad (5)$$

- Removing granules with a radius that is too large: if $r_j > 2 * max(mean(r), median(r))$ then $G_j$ is split.

The Generation of the Granular-Space algorithm can be briefly described as follows:

**Algorithm 1** *Generation of Granular-Space.*

```
Input: A dataset $X = \{x_i , x_i \in R^d\}$, $i = 1, 2, ..., n$ the number of clusters $c$ ($1 < c < n$)
Output: The granular space $G$
1- For each granule $G_j$ in $X$ do
2- Calculate $DM_O$, $DM_W$ by using (1), (2), (3), (4)
3- If $DM_W \geq DM_O$ Then Split $G_j$
4- If the number of $G$ not changing Then break;
5- End For
6- For each granule $G_j$ in $X$ do
7- calculate $(mean(r), median(r),$
8- If $r_j \geq 2 * max(mean(r), median(r))$ Then Split $G_j$
9- If the number of $G$ is not changing Then break;
10- End For
11- Return $G$
```

*b. Clustering FPCM based on Granular-Space (FPCM-GS).*

First, execute Algorithm 1 to obtain the granular space G, then apply the Fuzzy Possibilistic C-Means Clustering Algorithm [1] on the granular space *G* (FPCM-GS).

The objective function for FPCM-GS was built as follows:

$$\boldsymbol{J = \sum_{i=1}^{c} \sum_{k=1}^{g} u_{ik}^m \, t_{ik}^p \, d_{ik}^2 + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} u_{ik}^m \, (1 - t_{ik})} \quad (6)$$

where *g* is the number of granules on *G*, *c* is the number of clusters, $d_{ik}$ is the distance between the centroid $v_i$ and the $\theta_k$ which is the center of $G_k$; *p* and *m* are weighting exponents (possibilistic membership and fuzzifier), $\gamma_i$ is the scale parameter is determined as follows:

$$\boldsymbol{\gamma_i = K \frac{\sum_{k=1}^{g} t_{ik}^p u_{ik}^m d_{ik}^2}{\sum_{k=1}^{g} t_{ik}^p u_{ik}^m}, K > 0} \quad (7)$$

$t_{ik}$ is the possibilistic membership degree and $u_{ik}$ is the degree of fuzzy membership. They are computed as follows:

$$\boldsymbol{t_{ik} = \frac{1}{1 + (\frac{d_{ik}^2}{\gamma_i})^{\frac{1}{p-1}}}} \quad (8)$$

$$\boldsymbol{u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{t_{ik}^{(p-1)/2} d_{ik}}{t_{jk}^{(p-1)/2} d_{jk}})^{\frac{2}{m-1}}}} \quad (9)$$

The centroids of cluster $v_i$ are determined in the same way of FPCM as follows:

$$\boldsymbol{v_i = \frac{\sum_{k=1}^{g} t_{ik}^p u_{ik}^m \theta_k}{\sum_{k=1}^{g} t_{ik}^p u_{ik}^m}} \quad (10)$$

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

The FPCM-GS algorithm can be briefly described as follows:

**Algorithm 2** *Advanced FPCM based on Granular-Space.*

```
Input: A dataset $X = \{x_i , x_i \in R^d\}$, $i = 1, 2, ..., n$, the number of clusters $c$ ($1 < c < n$), error $\varepsilon$.
Output: T (the possibilistic membership matrix), U (the fuzzy membership matrix), and V (the centroid matrix).
1-Execute Algorithm 1 to obtain the granular space G
2- $l$ = 0
3- Repeat:
4- $l = l + 1$
5- Update $T^{(l)}$ by using (8)
6- Update $U^{(l)}$ by using (9)
7- Update $V^{(l)}$ by using (10)
8- Apply (7) to compute $\gamma_1, \gamma_2, ..., \gamma_c$
9- Until:
10- $Max(\|U^{(l+1)} - U^{(l)}\|) \leq \varepsilon$
11- Return $T, U, V$
```

## 2.2. Experiment Preparation

The proposed clustering algorithm was experimentally validated by carefully implementing and executing a set of experiments that were conducted according to a well-defined set of guidelines.

Some well-known available datasets are used in the experiments. We also offer a comparative analysis of the clustering results between some clustering algorithms (FCM, PCM, FPCM) and GrFPCM. Through the process of experiment, the clustering results are stable with parameters: m = p = 2; $\varepsilon = 0.00001$; $K = 1$.

### *2.2.1. Instrumentation*

The clustering results are evaluated by determining indices: False Positive Rate (FPR) and TPR (True Positive Rate). They are defined as follows:

$$FPR = \frac{FP}{TN + FP}; \; TPR = \frac{TP}{TP + FN} \qquad (11)$$

in which:
- *FP:* the number of incorrectly classified data.
- *TN:* the number of correctly misclassified data.
- *TP:* the number of correctly classified data.
- *FN:* the number of incorrectly misclassified data.

### *2.2.2. Experimental Materials*

The algorithms are implemented in the VC++ program and run on Intel Core i7-3517U CPU 1.90GHz - 2.40GHz, 8.0 GB RAM.

## 3. RESULTS AND DISCUSSION

### 3.1. Input Data

The data that formed the input were a heterogeneous set of multilayered, multidimensional data sets that had different levels of complexity and considerable size. These datasets were retrieved from adequately performing information strings and treated to make them consistent and valid for experimentation.

Specifically, in this case, the well-known datasets are WDBC, DNA, Madelon, Global Cancer Map, and Colon are considered. The datasets are shown in Table 1.

**Table 1**. Datasets are used to illustrate the proposed method

| Datasets | Number of samples | Number of clusters |
|---|---|---|
| WDBC | 569 | 2 |
| DNA | 106 | 2 |
| Madelon | 4400 | 2 |
| Global Cancer Map | 190 | 14 |
| Colon | 62 | 2 |

### 3.2. Simulation results and comments

### *3.2.1. Clustering results*

The results of the experiment are reported in terms of indices TPR and FPR, which are shown in Table 2 and graphically shown in Figures 1 and 2. These results also show the quality of classification when performing the clustering by each method. Table 2 shows the results of the clustering, in which the lower the FTR value and the higher the TPR value, the better the method is. The FPCM-GS algorithm obtained the smallest FPR and the highest TPR on all five datasets.

**Table 2**. The results of the experiment in terms of indices TPR and FPR

| Datasets | FCM | | FPCM | | FPCM-GS | |
|---|---|---|---|---|---|---|
| | FPR | TPR | FPR | TPR | FPR | TPR |
| WDBC | 4.5% | 89.5% | 2.8% | 92.7% | **1.8%** | **95.8%** |
| DNA | 6.7% | 85.6% | 3.1% | 91.4% | **1.7%** | **96.1%** |
| Madelon | 5.9% | 86.1% | 3.3% | 90.8% | **2.0%** | **94.9%** |
| Global Cancer Map | 4.8% | 89.6% | 5.5% | 90.2% | **1.2%** | **96.8%** |
| Colon | 7.9% | 79.1% | 9.5% | 80.9% | **1.6%** | **92.2%** |

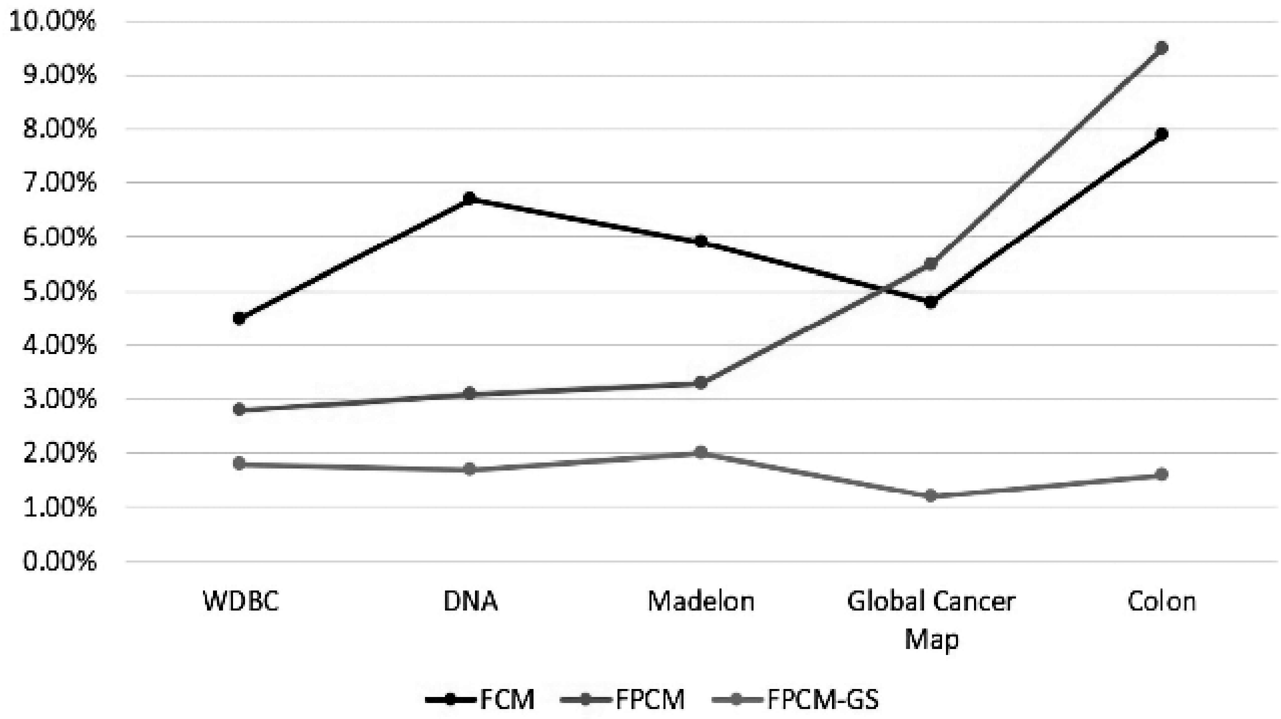
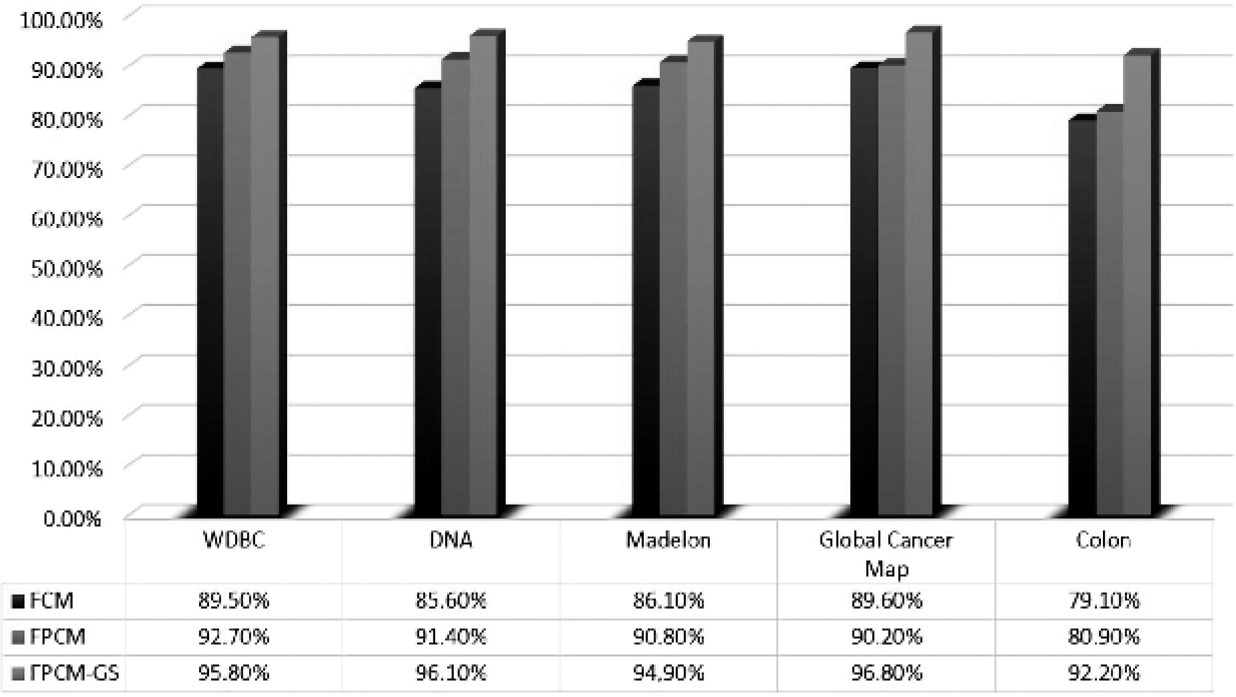


**Figure 1.** The FPR values of clustering results



| | WDBC | DNA | Madelon | Global Cancer Map | Colon |
|---|---|---|---|---|---|
| ■ FCM | 89.50% | 85.60% | 86.10% | 89.60% | 79.10% |
| ■ FPCM | 92.70% | 91.40% | 90.80% | 90.20% | 80.90% |
| ■ FPCM-GS | 95.80% | 96.10% | 94.90% | 96.80% | 92.20% |

**Figure 2.** The TPR values of clustering results

From the results of the experiment in terms of indices TPR and FPR, the TPR values obtained by running FPCM-GS on five datasets are greater than 92% and obviously higher than the ones obtained from other algorithms. In addition, the FPR values are also smaller than the ones reached by other methods. Therefore, we can conclude that by forming the granular space for experimental datasets, the quality of the clustering results has been improved.

The computed results of the analytical techniques demonstrated that the clustering approach of granular computing provided outcomes that were more accurate than the ones of the old-fashioned methods.

The findings reveal that the micro-style of clustering exhibits much more refined and even more detailed clustering of the data points than the macro approach, with the former being particularly more efficient in cases of irregular datasets [11].

The outcomes authenticate the very ideas of the granular computing concept, which states that a higher level of arbitrary computing is able to be more productive. The practical applications of the algorithm made it possible to confirm the algorithm's potential for real-world usage.

While other clustering algorithms were previously used, the proposed method includes the principle of justifiable granularity that combines detail and abstraction not observed earlier.

The applied advantages of this study are diverse. The Algorithm's capability of quickly and accurately grouping large datasets is a priceless tool for data-driven decision-making in different areas of influence, such as finance, healthcare, and social media analytics.

The innovative clustering algorithm that we suggest, based on the principle of justifiable granularity, cuts off from similar techniques the previous works used. While the principle of granularity is the primary aspect of granular computing, it provides the ability to make information granules of a specific amount of detail at both the necessary and sufficient levels for the purpose intended. Traditional clustering methods often, on the one hand, employ simplistic approaches that disregard data details to achieve a neat result or go ahead with complicated models that are extremely difficult to understand and employ. Nevertheless, the algorithm developed here walks the tightrope, offering just the desired level of granularization that is based on the requirements of the data and the purpose of the analysis.

It along with the interconnection of the parts of this matter is not a limitless picture but has practical implications with a validity of most areas of life. The capability of grouping data with corresponding decision capacity is of utmost value in finance as risk assessment and identification of fraud can be done with a higher level of accuracy. It can be used in the healthcare system to better identify groups or stratify patients and provide treatment plans tailored to individual needs in healthcare. Apart from social media analytics, it represents a more accurate picture of user behavior and desires towards the product.

Another key advantage is the algorithm's adaptability enabling it to work with data of varying sizes and complexities and hence making it a versatile tool for data scientists and analysts. Its efficiency in clustering large datasets quickly and accurately addresses one of the key challenges in the field of big data: busting the myths of who runs the web today, we offer various technological solutions to attain speed without compromising quality.

Besides, the scalability of this algorithm is one of the reasons why it is perceived as a practical solution. As data is supposed to increase in volume and variety, the worth of algorithms that keep up with the growth becomes even more prevalent. The architecture of the algorithm relies on the foundational granular computing principle which makes it scalable so it can effectively handle large volumes of data without having an added burden of a proportional increase in computational complexities.

Also, it offers fresh fields for research and advances in technology. For these fields, it can be referred to as a prototype, that principle of intuitiveness of the technology looked into areas of machine learning and artificial intelligence, and could potentially revolutionize how machines operate tackling complex problems, instantly as humans used to do. The previous algorithm will be amended by including real-time data analysis in comparison with the contribution to the applications where quick decision-making is primarily important.

However, the proposed clustering algorithm is a revolutionary leap in the method used before for implementation. The approach adds precision by going to the details. The final effect is the greater significance and practicality of the process. Now, we can witness the unparalleled rise of data in all spheres of decision-making. This unquestionably confirms the impactful need for such an algorithm. Such development is considered as the new stage in the progress towards a more intelligent, adaptive, and effective information processing machine.

## 4. CONCLUSION

The research done has provided a lot of useful results with theories and methods considered. The findings showed that the use of granular computing for data clustering offered a unique way that increase the interpretability and speed of the clustering processes. The cost of this result is the fact that it is a new dimension that is not only a continuation of the existing achievements but also includes the principle of justifiable granularity.

The flexibility of the indicated method makes it possible to implement it in bioinformatics, financial analysis, and other fields where data clustering plays the main role. A big data era is characterized by a lot of data, with its complex structures. Therefore, this approach is particularly effective in this kind of data.

Nevertheless, there are still many aspects that deserve more attention. As far as the scalability of the algorithm with regard to exponentially growing data sizes, the incorporation of the method into real-time data streams, and the adjustment to different types of data anomalies, these areas need to be further studied.

The enhancement of algorithms' robustness and the exploration of this algorithm's applicability in other AI domains, like deep learning and cognitive computing, are the main directions of this research problem's development. Also, the prospects of incorporating granular computing with other unsupervised learning practices like deep learning are to be considered to generate more sophisticated data analytical tools.

In the future, research could investigate the use of alternative granular computing methods to enhance clustering quality, especially when working with large datasets. Additionally, it would be beneficial to incorporate more advanced methods, such as evolutionary optimization, to fine-tune the parameters of the clustering method.

## REFERENCES

[1] 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.

[2] 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.

[3] Tang, Y., Gao, J., Pedrycz, W., Hu, X., Xi, L., Ren, F., & Hu, M. (2024), "Modeling and Clustering of Parabolic Granular Data", *IEEE Transactions on Artificial Intelligence*, 1(01), 1-15.

[4] Ding S., Huang H., Yu J. (2015), "Research on the hybrid models of granular computing and support vector machine", *Artificial Intelligence Review*, 43(4), pp.565-577.

[5] Dong H., Li T., Ding R., Sun J. (2018*),* "A novel hybrid genetic algorithm with granular information forfeature selection and optimization", *Applied Soft Computing*, 65, 33-46.

[6] Fu C., Lu W., Pedrycz W., Yang J. (2019), *"Fuzzy granular classification based on the principle of justifiable granularity"*, *Knowledge-Based Systems*,170, pp.89-101.

[7] Alswaitti M., Ishak M. K., Isa N. A. M. (2018), "Optimized gravitational based data clustering algorithm", *Engineering Applications of Artificial Intelligence*, 73, pp. 126148.

[8] Sanchez M.A., Castillo O., Castro J.R., Melin P. (2014), *"Fuzzy granular gravitational clustering algorithm for multivariate data"*, *Information Sciences*, 279, pp. 498511.

[9] X. Hu, Y. Tang, W. Pedrycz, K. Di, J. Jiang and Y. Jiang, "Fuzzy Clustering With Knowledge Extraction and Granulation," *in IEEE Transactions on Fuzzy Systems*, vol. 31, no. 4, pp. 1098-1112, April 2023, doi: 10.1109/TFUZZ.2022.3195033.

[10] J. Xie, W. Kong, S. Xia, G. Wang and X. Gao, "An Efficient Spectral Clustering Algorithm Based on Granular-Ball", *in IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9743-9753, 1 Sept. 2023, doi: 10.1109/TKDE.2023.3249475.

[11] Truong, H. Q., Ngo, L. T., & Pedrycz, W. (2016, October), *"Advanced fuzzy possibilistic C-means clustering based on granular computing", In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 002576-002581). IEEE.

# Nghiên cứu phương pháp phân cụm dữ liệu tính dựa trên tính toán hạt

**Trương Quốc Hùng, Nguyễn Huy Liêm, Vũ Minh Hoàng**
**Trần Thị Hải Anh và Nguyễn Thị Lan**

**TÓM TẮT**

*Bài báo này giới thiệu một kỹ thuật phân cụm mới dựa trên các khái niệm của tính toán hạt. Trong các thuật toán phân cụm truyền thống, việc tích hợp khả năng định hình cao của các bộ dữ liệu gốc khá phức tạp, từ đó dẫn đến kém hiệu quả. Kỹ thuật đề xuất giải quyết những hạn chế đó thông qua việc sử dụng tính toán hạt để giúp quá trình phân cụm được nhanh chóng và chính xác hơn. Thuật toán phân cụm mới trở nên tự nhiên hơn khi nó thực hiện trên không gian hạt giống như là các khối thông tin mang cấu trúc tự nhiên của dữ liệu gốc. Quá trình thử nghiệm thuật toán mới được thực hiện trên các bộ dữ liệu hiện đại và so sánh với các phương pháp phân cụm khác. Các kết quả cho thấy thuật toán phân cụm được đề xuất đã cải thiện đáng kể về độ chính xác cũng như giảm được thời gian xử lý dữ liệu. Qua đó chứng minh rằng việc tích hợp của tính toán hạt mạng lại hiệu quả đáng kể trong phân cụm dữ liệu. Kết quả nghiên cứu này không chỉ giúp cải thiện hiệu quả trong phân cụm dữ liệu mà có thể được sử dụng trong bước tiến xử lý của kỹ thuật học không giám sát và cải thiện giải pháp cho các bài toán hỗ trợ ra quyết định.*

***Từ khóa:*** *phân cụm dữ liệu, phân cụm thông tin, tính toán hạt, hạt thông tin, học không giám sát, độ chính xác của thuật toán*

Received: 15/05/2024
Revised: 09/06/2024
Accepted for publication: 10/06/2024