Feature Selection Approach Based on Improved Fuzzy C-Means with Principle of Refined Justifiable Granularity

Wentao Li, Shichao Zhai, Weihua Xu, Witold Pedrycz, *Life Fellow, IEEE*, Yuhua Qian, *Member, IEEE*, Weiping Ding, *Senior Member, IEEE*, and Tao Zhan

***Abstract*—Fuzzy C-Means (FCM) is a clustering algorithm based on partition of the universe. However, the partition generated by an equivalence relation is strict in practical application and exhibits relatively poor fault-tolerant mechanism. In this paper, a novel binary relation based on improved FCM with the principle of refined justifiable granularity is presented. Different expressions of the proposed binary relation under different values of weight parameter are discussed, and the changes of the properties of the binary relation under different parameter values are provided. By measuring the significance of attributes in the feature space, a feature selection method, called forward heuristic feature selection (FHFS), is designed to construct the low-dimension feature space based on maximizing the original data and information retention through the defined degrees of aggregation and dispersion. It is shown how the results of feature selection and classification performance vary when the values of the weight factor locate in different ranges. To illustrate the superiority and effectiveness of the proposed FHFS algorithm, nine high-dimensional datasets and eight image datasets from UCI repository are used and compared with other feature selection methods, respectively. The results of experimental evaluation and the significance test show that the proposed learning mechanism is a superior algorithm.**

***Index Terms*—Feature selection; Granular computing; Information granularity; Justifiable granularity**

I. Introduction

FEATURE selection is a process to select important features (or remove redundant and irrelevant features) from original target feature space to improve the accuracy and simplify learning tasks [1]–[3]. When the size of datasets is extraordinarily large, it can significantly save the storage space and computational overhead of data analysis from the feature selection procedure. The significance measures of features and selection criteria are important to feature selection, influencing the effectiveness and classification performance of the reduction sets [4]–[7].

The existence of uncertainty brings difficulties and challenges to data processing. From the viewpoint of granular computing (GrC), uncertainty associates with a kind of representation of information at different levels of information granularity [8]. Zadeh first explored the concept of GrC and coined an informal yet highly descriptive notion of an information granule [9]. In general, by information granule, one regards a collection of elements drawn together by their closeness (resemblance, proximity, functionality, etc.), articulated in terms of some useful spatial, temporal, or functional relationships. GrC is about to represent, construct, and process information granules. Many excellent achievements regard to GrC have been reported in references [10]–[13].

It is worth stressing that information granules, as encountered in natural language, are implicit in their nature, and information granules permeate human endeavors. No matter which problem is taken into consideration, we usually set it up in a certain conceptual framework composed of some generic and conceptual meaningful entity information granules, which we regard to be of relevance to the problem formulation, further problem solving, and a way in which the findings are communicated to the community. Information granules are formalized in many different ways [14]. Clustering delivers a natural mechanism to construct information granules in the presence of numeric data. As a matter of fact, a main agenda of clustering is to reveal a structure of data, namely, from a collection of clusters-information granules. There is a genuine diversity of clustering algorithms. Depending upon the method used, the results arise information granules expressed in terms of sets, fuzzy sets, rough sets, and so forth.

Clustering is to divide a dataset into different classes or clusters according to a specific standard (such as distance criterion), so that the similarity of objects in the same cluster is as large as possible, and the differences of objects not in the same cluster are as prominent as possible. Clustering usually does not need to use labelled training data for learning and unsupervised learning [15]–[19]. As a typical

This paper is supported by the National Natural Science Foundation of China (Nos. 12201518, 61976245, 61976120), the China Postdoctoral Science Foundation (2021M700432), the Science and Technology Research Program of Chongqing Education Commission (Nos. KJQN202100205, KJQN202100206), and the Natural Science Key Foundation of Jiangsu Education Department (No.21KJA510004). (*Corresponding author: Tao Zhan.*)

W. Li is with the College of Artificial Intelligence, Southwest University, Chongqing, 400715, P.R. China, and the School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, P.R. China. (E-mail: drliwentao@gmail.com)

S. Zhai and W. Xu are with the College of Artificial Intelligence, Southwest University, Chongqing, 400715, P.R. China. (E-mails: 13754808860@163.com, chxuwh@gmail.com)

W. Pedrycz is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6R 2V4, Canada. (E-mail: wpedrycz@ualberta.ca)

Y. Qian is with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, and the School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, P.R. China. (E-mail: jinchengqyh@126.com)

W. Ding is the School of Information Science and Technology, Nantong University, Nantong, 226019, P.R. China. (E-mail: dwp9988@163.com)

T. Zhan is with the School of Mathematics and Statistics, Southwest University, Chongqing, 400715, P.R. China. (E-mail: zhantao0817@163.com)

clustering method of unsupervised learning, Fuzzy C-Means (FCM) based clustering algorithm has been widely used in feature selection and image processing [20]–[25]. However, algorithms from the above literatures define clustering as the largest set of density connected points, searching clusters of arbitrary shapes in noisy spatial datasets, but the effect is not obvious when dealing with datasets with uneven density. The essential reason of this induced limitation is that these methods do not consider the requirements of intra-class compactness and between-class sparsity. It is found that after introducing the principle of justifiable granularity [8], [11]–[14], [26]–[28], this issue can be solved to a certain extent.

Based on the above analysis, in this paper, we establish a novel FCM-based feature selection approach with wider application range compared with traditional basic FCM. This novel method mainly relaxes the limitation of partition of universe, improves FCM algorithm for clustering, and then uses the principle of justifiable granularity to formulate feature selection rules. Meanwhile, we apply the proposed algorithm to the practical application scenario of the image feature extraction. Through the comparison of common evaluation indicators, one notes that the method presented in this paper exhibits better experimental results and application results of image features. The main contents and innovations of this paper are summarized as follows.

1) The binary relation based on improved FCM with the principle of refined justifiable granularity is defined to overcome the limitation of the traditional FCM-based partition. The proposed method exhibits better performance in the process of image segmentation. Meanwhile, the novel justifiable granularity based binary relation can change the properties by adjusting the parameter value, which has adaptive characteristics, and is suitable for wider circumstances compared with the traditional FCM-based binary relation.
2) A novel criterion based on the degrees of aggregation and dispersion for measuring the importance of attributes is developed. The newly defined significance measurement not only exhibits interpretability, but also reduces the amount of computation and effectively improves the operation efficiency of the heuristic attribute importance traversal algorithm. The proposed of attribute importance measurement criteria has guiding significance for establishing a set of mature feature selection criteria.
3) Rules of improved FCM-based forward heuristic feature selection (FHFS) with justifiable granularity are presented, and the related algorithms for achieving refinement of information granularity and FHFS are derived to improve the computational efficiency and recognition accuracy of big data processing in different scenarios.
4) Nine publicly available high-dimensional datasets are utilized for experimental evaluation, and the experimental results show the superiority of the proposed algorithm. Experiments related to image processing are also designed to show that the proposed method has significant effects and advantages in image feature selection after image segmentation.

This paper designs a feature selection algorithm based on improved FCM by using the principle of refined justifiable granularity, and verifies its performance through utilizing the data dimensionality reduction and image feature extraction. Fig. 1 describes the flow chart of the entire work. The paper is organized as follows. Some necessary and important concepts about FCM-based clustering and optimal information granularity with the literature survey are introduced in Section II, and the motivation of this paper is recalled. In Section III, the construction method of BRIG is defined and its important properties are further investigated. In Section IV, we mainly design the related algorithm to derive the feature selection approach based on improved FCM with principle of refined justifiable granularity. In Section V, the corresponding experimental testing is conducted by nine datasets from UCI datasets to test the advantage of FHFS, and the specific application in image processing is proposed and verified. Finally, Section VI covers some conclusions.

TABLE I: TERMINOLOGY NOTATION

Terminology	Explanation
$G(\overline{x}_i)$	The decision function of samples
$d_p(\overline{x}_1, \overline{x}_2)$	Minkowski distance
x_k	n-dimensional vectors
$\varphi_k \in \{1, 2, ..., c\}$	The index of class
ν_i	Centroid vector of i-th cluster
μ_{ik}	The membership degree
Ω_i	The i-th information granularity
ρ_i	Size of information granularity Ω_i
$cov(\Omega_i)$	Coverage of Ω_i
$spec(\Omega_i)$	Specificity of Ω_i
$H(\Omega_i)$	Homogeneity of Ω_i
$V(\Omega_i)$	Optimal information granularity index
BRIG	Binary relation based on information granularity
DA	Degree of aggregation
DD	Degree of dispersion
GDA_B	DA of the fuzzy decision system under B
DS_B	Separability of the fuzzy decision system under B
SIG	Significance of attributes
FHFS	Forward heuristic feature selection

II. Basic concepts and recent literature review

In this section, some basic concepts of clustering and the principle of justifiable granularity are briefly reviewed. Necessary symbolic notations are explained in Table I.

A fuzzy subset X of U is defined as a membership function assigning to each element x of U a certain degree of membership. The value $X(x) \in [0, 1]$ is referred to as the membership degree of x to the fuzzy set X.

Definition 2.1: An information system is a tuple (U, C, V, f), where $U = \{x_1, x_2, \cdots, x_n\}$ is a non-empty and finite set of objects; $C = \{a_1, a_2, \cdots, a_m\}$ is a non-empty and finite set of attributes; $f = \{f_l | U \to V_l, l \leq m\}$, f_l is the value of a_l on $x \in U$, V_l is the domain of a_l, $a_l \in C$.

A decision information system is $I = (U, C \cup D, V, f)$, where $C \cap D = \emptyset$, C and D are the condition and decision attribute set, respectively. A decision information system is called a fuzzy decision system, if all attribute values are fuzzy.

A. Clustering

The objective of clustering is to divide the dataset into several disjoint subsets. Given a m-dimensional sample dataset

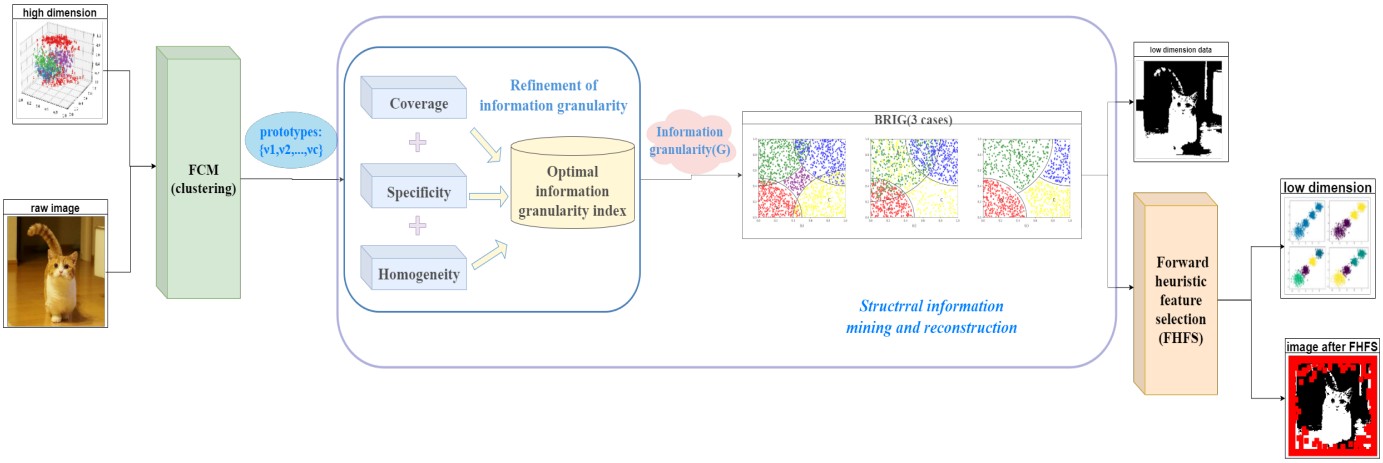

Fig. 1: The schematic flow chart.

$X = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_n\}$, and $\overline{x}_i \in \mathbf{R}^m$. Suppose that a criterion (not unique) can be found so that each sample can be associated with a specific group according to its special characteristics. The overall structure of the dataset $g_k = G(\overline{x}_i)$, $k = \{0, 1, 2, ..., t\}$, where g_k refers to the decision attribute value of each sample. $G(\overline{x}_i)$ indicates the decision function of samples about decision attributes. In general, each group is called a cluster, and the process of finding the function G is called clustering.

Definition 2.2: Let $X = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_n\}$ be a sample dataset, $\overline{x}_i \in \mathbf{R}^m$. The Minkowski distance of $d_p(\overline{x}_1, \overline{x}_2)$ is

$$d_p(\overline{x}_1, \overline{x}_2) = (\sum_j |\overline{x}_1^j - \overline{x}_2^j|^p)^{\frac{1}{p}}, \tag{1}$$

where p is any real number greater than or equal to 1. $d_p(\overline{x}_i, \overline{x}_j)$ is called the Manhattan distance if $p = 1$, the Euclidean distance if $p = 2$, and the Tchebychev distance if $p = \infty$. Most of the clustering algorithms use the Euclidean distance. When $d_p(\overline{x}_i, \overline{x}_j)$ is larger, it means that the distance between $\overline{x}_i$ and $\overline{x}_j$ is larger, indicating that the similarity between them is smaller.

The distance formula provides a specific calculation method for the measurement of membership in Fuzzy C-Means (FCM) algorithm and the measurement of coverage in the refinement of information granularity algorithm.

B. FCM

Let the dataset compose of a collection of ordered pairs (x_k, φ_k), where x_k $(k = 1, 2, ..., N)$ are n-dimensional vectors and $\varphi_k \in \{1, 2, ..., c\}$ denotes the index of the corresponding class of input instance x_k. The prototype set $\{\nu_1, \nu_2, ..., \nu_c\}$ denotes the centroid vector of all clusters and each cluster has a corresponding prototype.

The FCM (Algorithm 1) can be summarized as follows. First, the prototypes $\{\nu_1, \nu_2, ..., \nu_c\}$ of the clusters are initialized randomly. Then, the distance d_{ik} between instance x_k and prototype ν_i will be calculated as the Euclidean distance. For each instance x_k, it will be distributed to the specific cluster whose corresponding prototype ν_i shows the shortest distance to x_k. Next, the membership degrees of instance x_k to the i-th cluster μ_{ik} are updated as follows

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}}, \tag{2}$$

where $m(m > 1.0)$ denotes the fuzzification coefficient which has an influences on the geometry of the membership functions generated by the algorithm. The membership degree measures the proximity of the instance x_k to the i-th cluster. When the distance between the instance x_k and the i-th cluster is far, that is, when the d_{ik} is large, the membership degree of the instance x_k with respect to the i-th cluster is small.

At the same time, the prototypes are updated in an iterative manner through the minimization of the value of loss function J until the changes of J lane lower than some predefined value. The expression of J and updated ν_i are defined as

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^m d_{ik}^2 \tag{3}$$

and

$$\nu_i = \frac{\sum_{k=1}^{N} \mu_{ik}^m x_k}{\sum_{k=1}^{N} \mu_{ik}^m}. \tag{4}$$

When the membership degree of the instance x_k to the i-th cluster is large and the distance is large, that is, the loss function J is large at this time, which means that the cluster formed by the FCM algorithm still does not form an effective cluster structure, and the algorithm needs to be optimized. Meanwhile, the cluster center ν_i can be regarded as the weighted sum of each instance in the cluster, and the weight of each instance is proportional to the membership degree of the instance relative to the cluster. When the membership degree is large, the weight is large, indicating that the current instance provides a large proportion to the formation of cluster center.

C. Justifiable granularity

Information granularity exhibits a rigorous data representation with the clusters formed by FCM, and it is closely combined with data expression methods.

Algorithm 1 FCM

Input: A collection of ordered pairs $D = (x_k, \varphi_k)$.
Output: Clusters of data.

```
Initialize: Prototypes {ν_1, ν_2, ..., ν_c}.
t ⇐ 1.
J^(0) ⇐ 0.
ε ⇐ 0.001.
while true do
    for each x_k ∈ {x_1, x_2, ..., x_N}, ν_i ∈ {ν_1, ν_2, ..., ν_c} do
        Compute d_ik;
        Find the nearest ν_i from x_k, and distribute x_k to the corresponding cluster;
        Compute μ_ik;
        Update the prototypes {ν_1, ν_2, ..., ν_c};
    end for
    Compute loss function J^(t);
    if |J^(t) − J^(t−1)| ≤ ε then
        break;
    end if
    t ⇐ t + 1;
end while
return Clusters and new prototypes {ν_1, ν_2, ..., ν_c};
```

Definition 2.3: (Information granularity) Let $I = (U, C \cup D, V, f)$ be a decision information system, if $\forall P \subseteq C$ has a real number $G(P)$ correspondence and satisfies: (1) Nonnegativity: $G(P) \geq 0$; (2) Invariance: $\forall P, Q \subseteq C$, if $P \approx Q$, then $G(P) = G(Q)$; (3) Monotonicity: $\forall P, Q \subseteq C$, if $P \prec Q$, then $G(P) < G(Q)$, then G is called the information granularity of I.

Based on the principle of justifiable granularity, the refinement of each information granularity should fulfill some requirements. (1) *Integrity of information*. The information granularity should reflect the existing experimental data as much as possible. (2) *High specificity*. The information granularity should be as specific as possible to make sure that smaller and more distinguishable information granularity is formed. (3) *Homogeneity measure*. To improve the similarity and homogeneity of data in the same information granularity, the diversity of data instances covered by the information granularity is quantified in terms of the entropy criterion.

Obviously, criteria (1) and (2) are conflicting. To meet the granularity requirements in terms of information coverage, we introduce the basic concepts of entropy and granularity into the framework of information coverage.

Definition 2.4: (Coverage) The n-dimensional vector $x_k \in \{x_1, x_2, ..., x_N\}(k = 1, 2, ..., N)$. The $\nu_i \in \{\nu_1, \nu_2, ..., \nu_c\}$ denotes the prototype of the i-th clusters, μ_{ik} denotes the membership degrees of instance x_k to the i-th cluster, the coverage of information granularity Ω_i can be expressed as

$$cov(\Omega_i) = \frac{1}{N} \sum_{x_k : ||x_k - \nu_i||^2 \leq n\rho_i^2} \mu_{ik}, \tag{5}$$

where ρ_i denotes the size of information granularity Ω_i.

The $cov(\Omega_i)$ represents the reflection degree of the current particle size Ω_i on the existing experimental data. When the size ρ_i of the granularity is larger, the more instances the granularity Ω_i can accommodate, and the higher the degree of response of the granularity Ω_i to the original data information.

Definition 2.5: (Specificity) The specificity is concerned with the interpretation semantics of information granularity. It is expressed as follows

$$spec(\Omega_i) = 1 - \rho_i, \ \rho_i \in [0, 1]. \tag{6}$$

Since the size of ρ_i is proportional to coverage, the specificity of the above definition is inversely proportional to coverage.

Definition 2.6: Let N_{ik} be the number of instances belonging to class $k, k = 1, 2, ..., p$, covered by information granularity $\Omega_i, i = 1, 2, ..., c$, and the overall number of instances within the current information granularity is $N_i = \sum_{k=1}^{p} N_{ik}$. The entropy function defined over the distribution of instances in each information granularity is defined as follows

$$H(\Omega_i) = -\sum_{k=1, N_{ik} \neq 0}^{p} \frac{N_{ik}}{N_i} log(\frac{N_{ik}}{N_i}). \tag{7}$$

When the granularity Ω_i contains more categories of instances, it means that the homogeneity of the granularity is lower and the degree of confusion is higher, and the amount of information contained is greater.

Theorem 2.1: (1) When all the instances in the current information granularity fall in the same category, the value of entropy criterion is equal to 0. (2) The maximum value of entropy H_{max} is attained when there is a uniform distribution of categories of instances (which is equal to $1/p$), i.e.

$$H_{max} = -\sum_{k=1}^{p} log(\frac{1}{p})/p. \tag{8}$$

Proof. (1) It is easy to verify $H(\Omega_i) = (-1)\frac{N_i}{N_i} log(\frac{N_i}{N_i}) = (-1)log(1) = 0$.

(2) Let $\frac{N_{ik}}{N_i} = t_k$, then $H(\Omega_i) = -\sum_{k=1, N_{ik} \neq 0}^{p} t_k log(t_k)$, and $\sum_{k=1}^{p} t_k - 1 = 0$. From Lagrange multiplier, $G(t_1, t_2, ..., t_p, \lambda) = -\sum_{k=1}^{p} t_k ln(t_k) + \lambda(\sum_{k=1}^{p} pt_k - 1)$. Take the derivatives of t_k and λ respectively, then $\frac{\partial G}{\partial t_k} = -ln(t_k) - 1 + \lambda = 0$, and $\sum_{k=1}^{p} t_k - 1 = 0$. Therefore, $t_k = e^{\lambda - 1}$, which means t_k is independent of the value of k. So H_{max} is attained when there is a uniform distribution of categories of instances, and $H_{max} = -\sum_{k=1}^{p} log(\frac{1}{p})/p$. □

This means that when the instances included in the granularity Ω_i belong to different categories, the entropy of the granularity is the largest. To determine the optimal size of the information granularity, an objective function is constructed to seek the optimal size of information granularity by combining the above three mentioned criteria.

Definition 2.7: The optimal information granularity index of Ω_i is expressed as

$$V(\Omega_i) = cov(\Omega_i) \cdot spec(\Omega_i)^{\alpha} \cdot (1 - H(\Omega_i)/H_{max}), \tag{9}$$

and the optimal value of ρ of Ω_i is determined as

$$\rho_{i-opt} = argmax_{\rho_i} V(\Omega_i), \tag{10}$$

where the non-negative weight factor α ($\alpha \geq 0$) indicates the balance between the coverage and specificity criteria, and the weight of specificity criteria varies with the change of the value of α. With the increase of the values of α, the produced information granularity becomes more specific. ρ_{i-opt} represents the corresponding size when the granularity is optimal.

The optimal information granularity index reflects the level of coverage criterion, specificity criterion, entropy criterion at the same time, which makes information granularity more explanatory and justifiable. When $V(\Omega_i)$ reaches the maximum, the information granularity described by $V(\Omega_i)$ has high cohesiveness and homogeneity. It indicates that the information granularity can better satisfy the three criteria above in this case. At the same time, the corresponding size of information granularity is extremely optimal. The purpose of this index is to reach the optimization goal in refinement of information granularity algorithm.

D. Brief literature review

From a general point of view, basic clustering techniques can be categorized into hierarchical methods [2], [5], [13], [15]–[18] and partitional methods [3], [19], [25], [26]. The task for partitional clustering algorithms is to partition the dataset into several clusters so that the samples in one cluster will find the largest difference from samples in other clusters. The latest researches show that various clustering algorithms have been developed to solve practical problems encountered in pattern recognition [5], [16], [18], machine learning [12], [15], data mining [2], [3], and bio-informatics [17], [19], [24]. Among these algorithms, the K-means algorithm [19] is considered as a partitional clustering, and can be extended into FCM in case each sample performs as a member of multiple clusters which has membership value indicating a soft assignment.

FCM requires two basic parameters, namely the number of clusters and the flexible parameter of the algorithm. The outputs of the algorithm are prototypes and a fuzzy partition matrix. This matrix represents membership degree of object belonging to each cluster. Different FCM-based clustering algorithms have been put forward to meet different criteria and task intentions. For example, Lei et al. proposed an improved FCM algorithm based on morphological reconstruction and membership filtering that is significantly faster and more robust than the basic FCM [24]. Parker et al. introduced a geometric progressive FCM and minimum sample estimate random FCM to accelerate the basic FCM algorithm [25]. Nie et al. proposed an unsupervised linear method to construct anchor-based similarity matrix and then performs spectral analysis [19]. Based on the basic characteristics of FCM, the FCM-based clustering have been successfully applied to feature selection, which is closely related to the research of this paper.

GrC concentrates on processing information granules. With the aid of principle of justifiable granularity, scholars have comprehensively discussed the formation and visualization process of information granules. Among them, Zhu et al. elaborated on the detailed realization of hierarchically structured granular models [11], and designed a collection of easily interpretable ellipsoidal information granules by engaging the principle of justifiable granularity [14]. Nguyen et al. constructed interval membership values for each class prediction from the meta-data of observation by using information granule [12]. Hu et al. come up with some strategy to address the fuzzy rule-based model by utilizing the structural information granules [26]. Lu et al. developed the formation of input hyper-box information granules through performing the hyper-box iteration granulation algorithm governed by justifiable granularity [27]. Ju et al. proposed a Dempster-Shafer theory-based rough granular description model based on the justifiable granularity [28].

The method of classical FCM-based granularity only discusses the potential structure information of elements, which can not show the internal relationship between elements in a visual and intuitive modality. At the same time, the parameter α will directly affect the construction effect of information granularity, and the process of parameter debugging increases the workload of constructing justifiable granularity. Inspired by the FCM-based clustering and the principle of justifiable granularity, in this paper, we want to build a unified rule to describe the way of justifiable granularity construction in different situations, and provide the corresponding geometric representation, so as to further discover the potential regularity of constructing information granularity.

III. Construction of Novel Binary Relation Based on Information Granularity

In this section, we construct a novel binary relation based on improved information granularity principle. Algorithm 2 describes the improved construction method of justifiable granularity. In order to utilize the unified rules to describe the construction mode and regularity of information granularity in different cases, we propose a novel binary relation with the principle of justifiable granularity combined with Algorithm 2, and discuss the specific expression of the binary relation through the parameter α in different cases, so as to further reflect the relationship between samples in different cases by means of visualization.

In order to better characterize the structure and internal relationship between samples, we need to introduce a new binary relation based on information granularity (BRIG).

Definition 3.1: Let $x_k \in \{x_1, x_2, ..., x_N\}(k = 1, 2, ..., N)$ be n-dimensional vector, $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_c\}$ be the set of information granularity induced by the improved FCM. The BRIG is defined as $R_G = \{< x_i, x_j > | \exists x_k \, \mathrm{s.t.} (x_i \in \Omega_m \wedge x_k \in \Omega_m) \wedge (x_k \in \Omega_n \wedge x_j \in \Omega_n) or (x_i \in \Omega_m \wedge x_j \in \Omega_m)\}$.

Theorem 3.1: BRIG satisfies the property of symmetry.

Proof. $\forall x_i, x_j \in \{x_1, x_2, ...x_N\}$, if $< x_i, x_j > \in R_G$, then $\exists x_k((x_i \in \Omega_m \wedge x_k \in \Omega_m) \wedge (x_k \in \Omega_n \wedge x_j \in \Omega_n))$ or $x_i \in \Omega_m \wedge x_j \in \Omega_m$, which means $x_k((x_j \in \Omega_m \wedge x_k \in \Omega_m) \wedge (x_k \in \Omega_n \wedge x_i \in \Omega_n))$ or $x_j \in \Omega_m \wedge x_i \in \Omega_m$, then $< x_j, x_i > \in R_G$. □

Theorem 3.2: When α lies in a specific range (the range varies according to the datasets), BRIG satisfies the reflexivity, vice versa.

Algorithm 2 Refinement of information granularity

Input: Clusters of data.
Output: Justifiable granularity.

Initialize: $P = \{0.01, 0.02, ..., 1.00\}, V = \emptyset, \Omega = \emptyset$.
for each $\nu_i \in \{\nu_1, \nu_2, ..., \nu_c\}$ **do**
  $\Omega_i = \emptyset, max\_v = 0, temp = 0$;
  for each $\rho_t \in P$ **do**
    $\Omega_t = \emptyset$;
    for each $x_k \in \{x_1, x_2, ..., x_N\}$ **do**
      if $||x_k - \nu_i||^2 \leq n\rho_i^2$ **then**
        $\Omega_t \Leftarrow \Omega_t \cup \{x_k\}$
      end if
    end for
    Compute $cov(\Omega_t), spec(\Omega_t), H(\Omega_t)$;
    Compute $V(\Omega_t)$;
    if $V(\Omega_t) > max\_v$ **then**
      $max\_v = V(\Omega_t)$;
      $temp = t$;
    end if
  end for
  $\Omega_i = \Omega_{temp}$;
  $\Omega \Leftarrow \Omega \cup \Omega_i$;
end for
return Ω;

Proof. If α in a specific range, then x_i belongs to one granularity, i.e. $\forall x_i \in \{x_1, x_2, ..., x_N\}, x_i \in \Omega_m (i = 1, 2, ..., N, m = 1, 2, ..., M)$. It means $x_i \in \Omega_m \wedge x_i \in \Omega_m$, then $< x_i, x_i > \in R_G$. □

Theorem 3.3: The classical FCM-based binary relation is an equivalence relation.

Proof. It is easy to prove that this relation satisfies the reflexivity, symmetry and transitivity. □

The equivalence classes formed by FCM constitute a partition of universe. However, equivalence relation-based partition is so strict in practical application and possesses relatively poor fault-tolerant mechanism. For example, in image segmentation, because the pixel values are continuously changing values (non-discrete values), the recognition accuracy and robustness of the algorithm are rather limited. We need to overcome the defect of the poor fault tolerance caused by using equivalence relations through constructing non-equivalence relations (similarity relation or binary relation with only symmetry) to replace equivalence relation and improve the anti-interference ability of noise points in the process of sample clustering. It can exhibit a better fault-tolerant mechanism. The description of BRIG with different value of weight factor α is represented in Fig. 2. As shown in Fig. 2 (a)-(c), the BRIG from Definition 3.1 is more general, and better information granularity construction could be achieved by adjusting the parameter α.

IV. Feature selection based on measurement of attribute importance

In this section, the degree of aggregation (DA) and the degree of dispersion (DD), which intuitively describe the structure of the data from the perspective of the intra-class compactness and the between-class sparsity, will be first introduced.

Definition 4.1: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system, $D_k \in U/D$ and the attribute subset $B \subseteq C$, the DA of the decision class D_k under B is defined as

$$DA_B(D_k) = \frac{\sum_{x_i \in D_k} \mu_{ik}}{|D_k|} = \frac{\sum_{x_i \in D_k} \mu_B(x_i, D_k)}{|D_k|}, \quad (11)$$

where the μ_{ik} and $\mu_B(x_i, D_k)$ are the membership grades of object x_i with respect to the decision class D_k under B.

The DA is reflected by the membership degree of the instance under the current attribute subset and the decision class D_k. When the sum of membership degrees of an instance under the current attribute subset is large, it indicates that the considered attribute similarity is high and DA is high. The expression (11) indicates that the membership has a negative correlation with the distance, the lager DA_B is, the closer the intraclass objects are.

Theorem 4.1: Let $B \subseteq C$ and $D_k \in U/D$, then $0 \leq DA_B(D_k) \leq 1$.

Proof. $\forall x_i \in D_k$, we obtain $0 \leq \mu_{ik} \leq 1$, such that $0 \leq \sum_{x_i \in D_k} \mu_{ik} \leq |D_k|$. Thus, $0 \leq DA_B(D_k) \leq 1$. □

Definition 4.2: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system, $B \subseteq C$, the DD of the fuzzy decision system under B is defined as

$$DD_B(I) = \frac{\sum_{k=1}^{K} d_B(\overline{C}, C_k)}{|U/D|}, \quad (12)$$

where U represents the universe, D represents decision attribute set, a_i represents the i-th value in attribute set B, $c_k(a_i)$ indicates the i-th component of the center in the k-th cluster center, and $\overline{C} = (\overline{c}(a_1), \overline{c}(a_2), ..., \overline{c}(a_{|B|}))$ is the center of all class centers under B, namely

$$\overline{c}(a_i) = (1/K) \sum_{i=1}^{K} c_k(a_i), \quad (13)$$

where $\overline{c}(a_i)$ represents the i-th component in the mean vector of all class centers, C_k is the center of the k-th cluster, and the $d_B(\overline{C}, C_k)$ denotes the distance from the $\overline{C}$ to C_k under the B.

The DD is reflected by the sum of the distances from the centers of each cluster to the total centers of all instances under the current attribute, indicating that the dispersion is proportional to the distance between instances.

Theorem 4.2: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system, if $B_1 \subseteq B_2 \subseteq C$, then $DD_{B1}(I) \leq DD_{B2}(I)$.

Proof. If $B_1 \subseteq B_2 \subseteq C$, we obtain that $(\sum_{a_i \in B_1} (\overline{c}(a_i) - c_k(a_i))^2 + \sum_{a_i \in B_2 - B_1} (\overline{c}(a_i) - c_k(a_i))^2)^{1/2} \leq d_{B_1}(\overline{C}, C_k)$. Thus, $DD_{B_1}(I) \leq DD_{B_2}(I)$. □

It can be shown that $DD_B(S)$ is the sparsity measure of between-class objects under B. The larger $DD_B(S)$ is, the more scattered the between-class objects are.

DA and DD indicate the intra-class compactness and between-class sparsity, respectively. They are used to compute the significance of attribute is the following content.

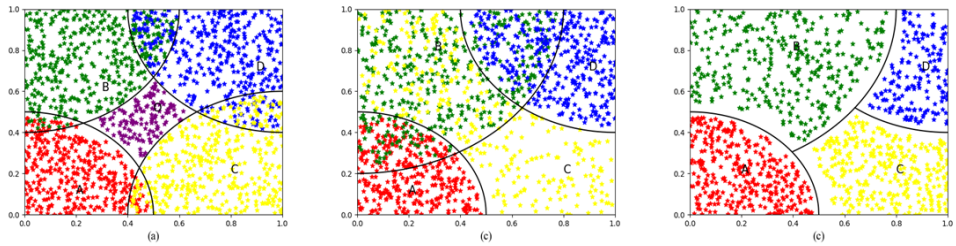

Fig. 2: (a) R_G does not satisfy reflexivity; (b) R_G satisfies reflexivity; (c) R_G satisfies reflexivity, symmetry and transitivity.

Definition 4.3: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system, $B \subseteq C$, the DA of the fuzzy decision system under B is defined as

$$GDA_B(I) = \frac{\sum\limits_{D_k \in U/D} DA_B(D_k)}{|U/D|}, \tag{14}$$

where I denotes the fuzzy decision system.

The $GDA_B(I)$ calculates the dispersion of all instances under attribute B by calculating the average value of different categories of DA.

Theorem 4.3: Let $B \subseteq C$ and $D_k \in U/D$, then $0 \leq GDA_B(D_k) \leq 1$.

Proof. Straight obtained. □

Formula (14) indicates that $GDA_B(D_k)$ is an index of the intra-class compactness of all decision classes under B.

Definition 4.4: Let $GDA_B(I)$ and $DD_B(I)$ are the degrees of aggregation and dispersion of I under B, the separability of the fuzzy decision system under B is defined as

$$DS_B(I) = GDA_B(I) \cdot DD(I). \tag{15}$$

Let $DS_B(I) = 0$ when $B = \emptyset$.

This definition states that $DS_B(S)$ is an index describing the significance of the conditional attribute subset relative to the decision based on the intra-class compactness and between-class sparsity.

Definition 4.5: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system and $B \subseteq C$,$a \in C - B$, $b \in B$.

(1) If $DS_{B \cup a}(I) \leq DS_B(I)$, then a is a redundant attribute for B.

(2) If $DS_{B-b}(I) < DS_B(I)$, then b is an indispensable attribute in B.

(3) If B satisfies $DS_{B \cup a}(I) \leq DS_B(I), DS_{B-b}(I) < DS_B(I)$, then B is a reduction of condition C with respect to decision D in I.

Definition 4.6: Let $I = (U, C \cup D, V, f)$ be a fuzzy decision system, $B \subseteq C$, $a \in C - B$, the significance of an attribute is defined as

$$SIG(a, B, D) = DS_{B \cup a}(I) - DS_B(I). \tag{16}$$

$SIG(a, B, D)$ is an index which describes the significance of attribute a with respect to B under decision D. Then the FHFS based on the significance measurement is provided in Algorithm 3. When the $SIG(a, B, D)$ is larger, the attribute a is more important in the attribute subset B.

Let us explain the Algorithm 3 as follows. First, we initialize an empty set AT to act as a collection used to deposit selected attributes. Then, the difference set $C - AT$ could be traversed, each attribute a_i in which will be computed the separability $DS_{B \cup \{a_i\}}(S)$ and the significance of attribute $SIG(a_i, AT, D)$. Next, the attribute a_k, which corresponds to the maximum value of $SIG(a_k, AT, D)$ in the difference set $C - AT$, will be deposited in the AT set. The process above as a loop will be a continued until the difference set $C - AT$ changes in to an empty set or the cardinality of AT is greater than the value of δ, which is a termination parameter and should be set in advance. How to set δ is a significant issue, which will be discussed in detail in the experimental part.

Algorithm 3 Forward heuristic feature selection (FHFS)

Input: A fuzzy decision system I and its justifiable granularity.
Output: Attribute set AT.

```
Initialize: AT ⇐ ∅.
while C − AT ≠ ∅ ∧ |AT| ≤ δ do
    for each a_i ∈ C − AT do
        Compute the degree of aggregation DA_{B∪{a_i}}(I), the degree of dispersion DD_{B∪{a_i}}(I), and the degree of aggregation of I under B: GDA_{B∪{a_i}}(I);
        Compute the separability DS_{B∪{a_i}}(I) and attribute SIG(a_i, AT, D);
    end for
    Find a_k with maximum value of SIG(a_k, AT, D);
    AT ⇒ a_k;
end while
return AT;
```

V. Experimental Analysis

In this section, we conduct a series of experiments to show the accuracy and effectiveness of the proposed feature selection algorithm. The FHFS is compared with one recent technique (denoted as EFSF) from reference [29], and three traditional techniques, namely principal component analysis (PCA) [30], nonnegative matrix factorization (NMF) [31], factor analysis (FA) [32], on nine public high-dimensional datasets. Four kinds of measurement indicators, namely precision, recall, f1-score, and accuracy, are utilized to show the performance of different comparison algorithms. Then, FHFS and other feature selection algorithms are compared in image feature selection through precision, recall, f1-score and accuracy. Moreover, the novel BRIG is applied to the process of image segmentation, and the influence of the value of α on the BRIG is discussed in detail. All algorithms are executed in Python 3.7 and run in a hardware environment with Intel Core i5-7200 CPU @2.50 and 2.70 GHz with 4-GB RAM.

Significance tests are designed for algorithms to verify whether the FHFS is better than other algorithms. In these tests, reasonable original assumptions and alternative assumptions are formulated as follows: (1) Original hypothesis (h_0). As algorithm is not superior to other algorithms in feature selection. (2) Alternative hypothesis (h_1). As algorithm is superior to other algorithms in feature selection. Moreover, we set the one-tail test with significance level $\alpha = 0.05$. The standard deviation of the corresponding performance of each algorithm is calculated by formula $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)}}$, where n means the freedom degree of data, and standard deviation of each group of data is expressed as $SE_{(\overline{x}_1 - \overline{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Here, the assumption of deviation is made between algorithm performance presents t-distribution, that is, $t = \frac{\overline{x}_1 - \overline{x}_2}{SE}$. In the formula above, $\overline{x}$ indicates the average, σ expresses the standard deviation.

A. Experimental design for high-dimensional datasets

In order to verify the advantages and superiority of the algorithm of FHFS, the comparisons about performance between FHFS and other algorithms (PCA, NMF, FA, EFSF) are made by using nine UCI datasets. Detailed description of the nine datasets are shown in Table II.

TABLE II: DATASET DESCRIPTION

High-dimensional datasets	Instances	Features	Classes
Seed	15,287	68	8
House	14,784	87	16
Sensor	4,107,458	128	6
Skin	145,298	107	8
Synthesis	43,897,238	847	6
Website	130,245	1013	7
Stock	12,878	973	2
Plants	11,848	2183	3
Segmentation	159,876	3012	2

Each group of data is normalized to avoid the influence caused by the difference of value, and the five algorithms, namely PCA, NMF, FA, EFSF and FHFS are used to conduct the process of feature selection. Then, classification algorithms, including KNN, decision tree (DT), and Bayes, are used to finish the task of classification. In KNN, weighted voting method are used to find the nearest five samples of the tested sample to predict its label according to the Euclidean distance. In DT, the feature partition samples selection criteria is set to random. In Bayes, the assumption of conditional probability distribution of each attribute satisfying the gaussian distribution is made before the process of classification.

Among this experiment, we use the ten-fold cross validation method to randomly and evenly divide each group of data samples into ten parts, and take turns to use nine of them to train the model and one to test precision, recall, f1-score and accuracy of the model. The calculated results are expressed in the form of $\mu \pm \sigma$, where μ represents the mean value of different indexes and σ represents the standard deviation of each index. The experimental results and related data of different datasets are shown in Tables III-V.

From the results displayed in Tables III-V, the FHFS algorithm has strong robustness and rationality compared with other algorithms in the process of high-dimension data preprocessing during classification tasks. For example, in the Skin dataset, after using FHFS algorithm for feature selection, the classifiers based on different classification algorithms perform all better than other feature selection algorithms in the indexes of precision, recall, f1-score and accuracy. In the KNN classifier, the classification accuracy of the data processed by FHFS is improved by 6% compared with the raw data and 2% compared with the EFSF algorithm. In the DT classifier, the classification accuracy of the data processed by FHFS is improved by 5% compared with the raw data and 2% compared with the EFSF algorithm. In the Bayes classifier, the classification accuracy of the data processed by FHFS is improved by 16% compared with the raw data and 4% compared with the NMF algorithm.

As for other datasets, although the classification performance after using FHFS algorithm is not completely superior to other algorithms in the indexes of precision, recall, f1-score and accuracy, the classification performance after using FHFS is still at a high level. For example, in the dataset of Seed with DT, the classification accuracy after using FHFS is 2% lower than that after using PCA, but it is still better than that after using other algorithms. These experimental results also indicate that performance of classifiers can be improved by using FHFS algorithm to process high-dimensional datasets. The performances of feature selection algorithms above in the KNN, DT and Bayes are shown in the Fig. 3-Fig. 11.

The value of function J in the process of clustering is decreased significantly with the increase of number of iterations of the algorithm. High convergence rate can be obtained originally, and the value of J can tend to be stable gradually. After the $[45, 50]$ epochs of algorithm operation, the value of J hardly changes, and extraordinary results of clustering will be obtained when the termination factor ε is set at the range of $[10^{-6}, 10^{-4}]$. The change trend of the value J is shown in the Fig. 12. The value J converges to a stable value when the epoch of iteration is less than fifty. It means that FCM algorithm has fast convergence speed, which also shows that the clustering results are reliable.

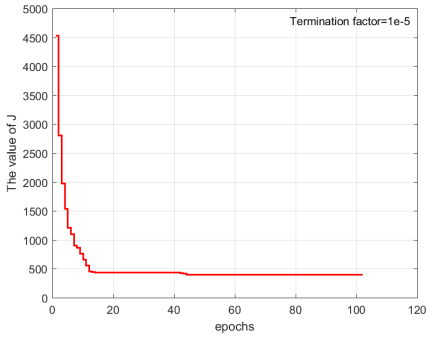

Fig. 12: Values of J in successive iteration.

From the result of significance test ($n = 38$) on high dimensional datasets shown in Table VI, FHFS is not superior to other feature selection algorithms (less than 0.05), which indicates that the original hypothesis (h_0) is not tenable.

TABLE III: EXPERIMENTAL RESULTS FOR DIFFERENT DATASETS (I)

Classifier	Method	Seed				House				Sensor			
		precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
KNN	Raw data	0.84 ± 0.03	0.81 ± 0.01	0.83 ± 0.01	0.81 ± 0.03	0.86 ± 0.02	0.82 ± 0.01	0.85 ± 0.01	0.83 ± 0.03	0.79 ± 0.02	0.78 ± 0.03	0.78 ± 0.04	0.75 ± 0.02
	PCA	0.81 ± 0.03	0.80 ± 0.01	0.78 ± 0.04	0.81 ± 0.01	0.78 ± 0.02	0.80 ± 0.01	0.82 ± 0.02	0.78 ± 0.03	0.79 ± 0.02	0.83 ± 0.02	0.81 ± 0.04	0.80 ± 0.02
	NMF	0.87 ± 0.04	0.85 ± 0.03	0.87 ± 0.02	0.88 ± 0.02	0.85 ± 0.02	0.88 ± 0.01	0.86 ± 0.01	0.87 ± 0.03	0.83 ± 0.01	0.80 ± 0.01	0.85 ± 0.02	0.80 ± 0.02
	FA	0.90 ± 0.01	0.89 ± 0.01	0.89 ± 0.02	0.90 ± 0.02	0.87 ± 0.04	0.87 ± 0.02	0.89 ± 0.01	0.86 ± 0.02	0.79 ± 0.01	0.78 ± 0.03	0.81 ± 0.01	0.78 ± 0.04
	EFSF	0.84 ± 0.02	0.87 ± 0.04	0.88 ± 0.02	0.89 ± 0.04	0.85 ± 0.03	0.86 ± 0.04	$\mathbf{0.92 \pm 0.01}$	0.88 ± 0.03	0.83 ± 0.02	0.80 ± 0.01	0.86 ± 0.01	$\mathbf{0.82 \pm 0.01}$
	FHFS	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.92 \pm 0.04}$	0.90 ± 0.01	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.89 \pm 0.01}$	0.80 ± 0.03
DT	Raw data	0.87 ± 0.01	0.87 ± 0.03	0.86 ± 0.01	0.86 ± 0.03	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.02	0.87 ± 0.03	0.88 ± 0.03	0.87 ± 0.03	0.89 ± 0.01	0.89 ± 0.01
	PCA	$\mathbf{0.93 \pm 0.03}$	0.91 ± 0.04	0.92 ± 0.04	0.92 ± 0.03	0.87 ± 0.03	0.88 ± 0.04	0.89 ± 0.01	0.87 ± 0.02	0.87 ± 0.03	0.86 ± 0.02	0.88 ± 0.02	0.89 ± 0.01
	NMF	0.87 ± 0.01	0.87 ± 0.01	0.89 ± 0.03	0.84 ± 0.01	0.86 ± 0.04	0.87 ± 0.03	0.82 ± 0.01	0.88 ± 0.03	0.89 ± 0.04	0.86 ± 0.01	0.88 ± 0.03	0.90 ± 0.04
	FA	0.89 ± 0.01	0.88 ± 0.03	0.89 ± 0.04	0.88 ± 0.04	0.88 ± 0.02	$\mathbf{0.90 \pm 0.03}$	0.86 ± 0.02	0.87 ± 0.04	0.81 ± 0.01	0.77 ± 0.02	0.80 ± 0.02	0.75 ± 0.02
	EFSF	0.86 ± 0.02	0.90 ± 0.01	0.90 ± 0.03	0.86 ± 0.02	0.90 ± 0.02	0.89 ± 0.03	0.87 ± 0.01	0.86 ± 0.04	0.88 ± 0.01	0.86 ± 0.02	0.86 ± 0.04	0.89 ± 0.02
	FHFS	0.91 ± 0.03	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.91 \pm 0.03}$	0.89 ± 0.04	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.92 \pm 0.03}$
Bayes	Raw data	0.78 ± 0.03	0.79 ± 0.02	0.76 ± 0.01	0.78 ± 0.02	0.78 ± 0.03	0.78 ± 0.03	0.77 ± 0.04	0.80 ± 0.01	0.80 ± 0.02	0.75 ± 0.02	0.79 ± 0.02	0.80 ± 0.03
	PCA	$\mathbf{0.92 \pm 0.01}$	0.89 ± 0.02	0.88 ± 0.02	0.91 ± 0.04	0.88 ± 0.03	0.81 ± 0.03	0.85 ± 0.04	0.85 ± 0.03	0.84 ± 0.04	0.83 ± 0.03	0.84 ± 0.01	0.84 ± 0.01
	NMF	0.76 ± 0.02	0.75 ± 0.01	0.79 ± 0.03	0.78 ± 0.02	0.82 ± 0.03	0.82 ± 0.02	0.83 ± 0.02	0.83 ± 0.02	0.77 ± 0.02	0.84 ± 0.04	0.83 ± 0.03	0.78 ± 0.02
	FA	0.85 ± 0.01	0.88 ± 0.04	0.87 ± 0.01	0.88 ± 0.04	0.86 ± 0.01	0.84 ± 0.03	0.85 ± 0.04	0.85 ± 0.04	0.79 ± 0.03	0.77 ± 0.02	0.78 ± 0.01	0.82 ± 0.01
	EFSF	0.85 ± 0.01	0.83 ± 0.03	0.85 ± 0.04	0.83 ± 0.02	0.81 ± 0.03	0.88 ± 0.02	0.85 ± 0.01	0.85 ± 0.02	0.82 ± 0.03	0.88 ± 0.03	0.84 ± 0.03	0.81 ± 0.03
	FHFS	0.90 ± 0.02	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.90 \pm 0.04}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.91 \pm 0.04}$

TABLE IV: EXPERIMENTAL RESULTS FOR DIFFERENT DATASETS (II)

Classifier	Method	Skin				Synthesis				Website			
		precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
KNN	Raw data	0.82 ± 0.01	0.80 ± 0.02	0.83 ± 0.04	0.83 ± 0.03	0.88 ± 0.02	0.89 ± 0.01	0.89 ± 0.01	0.87 ± 0.04	0.89 ± 0.04	0.88 ± 0.03	0.89 ± 0.03	0.87 ± 0.03
	PCA	0.85 ± 0.02	0.84 ± 0.02	0.80 ± 0.04	0.84 ± 0.04	0.84 ± 0.02	0.83 ± 0.04	0.85 ± 0.03	0.83 ± 0.03	0.85 ± 0.03	0.80 ± 0.03	0.85 ± 0.02	0.82 ± 0.04
	NMF	0.82 ± 0.03	0.83 ± 0.04	0.82 ± 0.02	0.84 ± 0.04	0.88 ± 0.01	0.87 ± 0.01	0.87 ± 0.03	0.85 ± 0.02	0.84 ± 0.03	0.83 ± 0.01	0.86 ± 0.04	0.87 ± 0.02
	FA	0.84 ± 0.01	0.84 ± 0.01	0.82 ± 0.04	0.84 ± 0.02	0.87 ± 0.02	0.88 ± 0.03	0.88 ± 0.02	0.85 ± 0.01	0.87 ± 0.03	0.86 ± 0.04	0.84 ± 0.04	0.82 ± 0.04
	EFSF	0.86 ± 0.02	0.81 ± 0.02	0.85 ± 0.01	0.84 ± 0.04	0.83 ± 0.03	0.80 ± 0.03	0.85 ± 0.01	0.83 ± 0.04	0.86 ± 0.01	0.81 ± 0.02	0.83 ± 0.03	0.83 ± 0.03
	FHFS	$\mathbf{0.88 \pm 0.03}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.90 \pm 0.04}$	$\mathbf{0.89 \pm 0.02}$
DT	Raw data	0.86 ± 0.04	0.85 ± 0.02	0.89 ± 0.03	0.88 ± 0.01	0.87 ± 0.01	0.89 ± 0.03	0.86 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.86 ± 0.03	0.88 ± 0.03	0.87 ± 0.04
	PCA	0.88 ± 0.03	0.88 ± 0.03	0.86 ± 0.01	0.84 ± 0.03	0.88 ± 0.01	0.84 ± 0.04	0.88 ± 0.01	0.85 ± 0.02	0.84 ± 0.03	0.85 ± 0.01	0.84 ± 0.04	0.87 ± 0.04
	NMF	0.87 ± 0.01	0.86 ± 0.02	0.89 ± 0.01	0.87 ± 0.04	0.87 ± 0.02	0.88 ± 0.04	0.84 ± 0.02	0.84 ± 0.03	0.84 ± 0.02	0.87 ± 0.04	0.84 ± 0.02	0.88 ± 0.04
	FA	0.85 ± 0.04	0.89 ± 0.02	0.84 ± 0.03	0.86 ± 0.01	0.84 ± 0.03	0.86 ± 0.02	0.87 ± 0.02	0.84 ± 0.03	0.88 ± 0.03	0.87 ± 0.03	0.86 ± 0.02	0.84 ± 0.03
	EFSF	0.89 ± 0.01	0.89 ± 0.04	0.86 ± 0.04	0.84 ± 0.03	0.88 ± 0.03	0.84 ± 0.03	0.87 ± 0.01	0.84 ± 0.02	0.85 ± 0.04	0.87 ± 0.03	0.84 ± 0.01	0.84 ± 0.03
	FHFS	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
Bayes	Raw data	0.76 ± 0.03	0.77 ± 0.01	0.80 ± 0.04	0.77 ± 0.04	0.79 ± 0.04	0.79 ± 0.01	0.75 ± 0.01	0.81 ± 0.02	0.78 ± 0.01	0.81 ± 0.01	0.76 ± 0.03	0.75 ± 0.04
	PCA	0.83 ± 0.02	0.78 ± 0.02	0.83 ± 0.02	0.82 ± 0.04	0.83 ± 0.01	0.82 ± 0.02	0.81 ± 0.04	0.79 ± 0.01	0.81 ± 0.03	0.82 ± 0.04	0.84 ± 0.04	0.84 ± 0.01
	NMF	0.76 ± 0.03	0.82 ± 0.02	0.81 ± 0.02	0.79 ± 0.02	0.84 ± 0.02	0.88 ± 0.04	0.81 ± 0.03	0.82 ± 0.01	0.85 ± 0.02	0.79 ± 0.01	0.83 ± 0.01	0.86 ± 0.03
	FA	0.87 ± 0.01	0.80 ± 0.01	0.82 ± 0.03	0.84 ± 0.01	0.83 ± 0.03	0.83 ± 0.02	0.86 ± 0.03	0.78 ± 0.01	0.82 ± 0.04	0.79 ± 0.02	0.84 ± 0.01	0.82 ± 0.02
	EFSF	0.88 ± 0.02	0.80 ± 0.03	0.85 ± 0.03	0.84 ± 0.03	0.80 ± 0.03	0.85 ± 0.02	0.83 ± 0.02	0.82 ± 0.02	0.82 ± 0.03	0.85 ± 0.01	0.87 ± 0.01	0.87 ± 0.02
	FHFS	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$

TABLE V: EXPERIMENTAL RESULTS FOR DIFFERENT DATASETS (III)

Classifier	Method	Stock				Plants				Segmentation			
		precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
KNN	Raw data	0.83 ± 0.03	0.82 ± 0.02	0.81 ± 0.04	0.84 ± 0.02	0.80 ± 0.03	0.81 ± 0.01	0.82 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.80 ± 0.02	0.82 ± 0.04	0.83 ± 0.04
	PCA	0.83 ± 0.02	0.82 ± 0.01	0.82 ± 0.03	0.81 ± 0.04	0.84 ± 0.02	0.81 ± 0.04	0.81 ± 0.04	0.81 ± 0.02	0.84 ± 0.04	0.80 ± 0.02	0.80 ± 0.01	0.80 ± 0.03
	NMF	0.82 ± 0.03	0.82 ± 0.03	0.84 ± 0.01	0.83 ± 0.03	0.81 ± 0.04	0.82 ± 0.02	0.82 ± 0.01	0.81 ± 0.04	0.84 ± 0.02	0.80 ± 0.01	0.82 ± 0.02	0.81 ± 0.01
	FA	0.82 ± 0.02	0.84 ± 0.02	0.83 ± 0.01	0.83 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.84 ± 0.01	0.81 ± 0.01	0.83 ± 0.03	0.81 ± 0.04	0.82 ± 0.03	0.82 ± 0.03
	EFSF	0.82 ± 0.03	0.83 ± 0.01	0.79 ± 0.02	0.78 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.84 ± 0.02	0.83 ± 0.01	0.82 ± 0.03	0.81 ± 0.01	0.85 ± 0.00	0.84 ± 0.02
	FHFS	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.94 \pm 0.04}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.94 \pm 0.02}$
DT	Raw data	0.81 ± 0.04	0.80 ± 0.02	0.77 ± 0.02	0.81 ± 0.04	0.82 ± 0.03	0.83 ± 0.03	0.78 ± 0.01	0.81 ± 0.01	0.84 ± 0.03	0.85 ± 0.01	0.84 ± 0.01	0.81 ± 0.02
	PCA	0.87 ± 0.02	0.88 ± 0.01	0.89 ± 0.01	0.89 ± 0.03	0.87 ± 0.00	0.86 ± 0.03	0.87 ± 0.03	0.88 ± 0.03	0.88 ± 0.02	0.89 ± 0.00	0.88 ± 0.02	0.89 ± 0.01
	NMF	0.86 ± 0.03	$\mathbf{0.89 \pm 0.00}$	0.87 ± 0.02	0.88 ± 0.01	0.86 ± 0.02	0.86 ± 0.00	0.85 ± 0.02	0.87 ± 0.03	0.89 ± 0.00	0.89 ± 0.03	$\mathbf{0.92 \pm 0.01}$	0.90 ± 0.01
	FA	0.86 ± 0.04	0.83 ± 0.04	0.85 ± 0.01	0.84 ± 0.03	0.86 ± 0.02	0.79 ± 0.03	0.79 ± 0.02	0.85 ± 0.02	0.86 ± 0.03	$\mathbf{0.89 \pm 0.03}$	0.87 ± 0.01	0.82 ± 0.01
	EFSF	0.88 ± 0.02	0.80 ± 0.01	0.84 ± 0.02	0.82 ± 0.04	0.86 ± 0.04	0.83 ± 0.01	0.84 ± 0.04	0.84 ± 0.03	0.85 ± 0.02	0.81 ± 0.04	0.82 ± 0.02	0.84 ± 0.02
	FHFS	$\mathbf{0.91 \pm 0.02}$	0.88 ± 0.02	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.90 \pm 0.04}$	0.88 ± 0.03	0.91 ± 0.02	$\mathbf{0.91 \pm 0.04}$
Bayes	Raw data	0.82 ± 0.01	0.81 ± 0.03	0.83 ± 0.01	0.82 ± 0.02	0.81 ± 0.02	0.80 ± 0.03	0.81 ± 0.02	0.81 ± 0.03	0.79 ± 0.03	0.82 ± 0.02	0.83 ± 0.04	0.85 ± 0.01
	PCA	0.79 ± 0.01	0.81 ± 0.04	0.80 ± 0.02	0.79 ± 0.02	0.80 ± 0.03	0.79 ± 0.02	0.83 ± 0.02	0.78 ± 0.02	0.80 ± 0.04	0.79 ± 0.01	0.85 ± 0.01	0.78 ± 0.04
	NMF	0.80 ± 0.03	0.84 ± 0.02	0.84 ± 0.03	0.79 ± 0.03	0.83 ± 0.03	0.82 ± 0.02	0.85 ± 0.04	0.80 ± 0.02	0.84 ± 0.03	0.80 ± 0.03	0.82 ± 0.01	0.79 ± 0.01
	FA	0.85 ± 0.01	0.78 ± 0.03	0.81 ± 0.01	0.85 ± 0.03	0.79 ± 0.03	0.83 ± 0.01	0.80 ± 0.02	0.83 ± 0.03	0.83 ± 0.04	0.80 ± 0.03	0.84 ± 0.02	0.84 ± 0.04
	EFSF	0.85 ± 0.03	0.80 ± 0.01	0.79 ± 0.03	0.85 ± 0.02	0.83 ± 0.02	0.85 ± 0.01	0.80 ± 0.01	0.83 ± 0.03	0.79 ± 0.03	0.79 ± 0.01	0.80 ± 0.04	0.79 ± 0.04
	FHFS	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.92 \pm 0.02}$

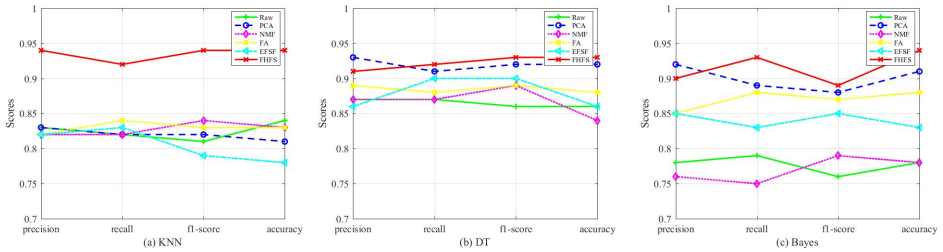

Fig. 3: Performance of different feature selection algorithms on Seed dataset.

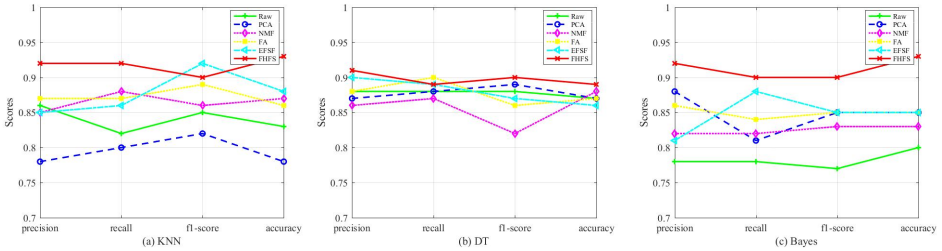

Fig. 4: Performance of different feature selection algorithms on House dataset.

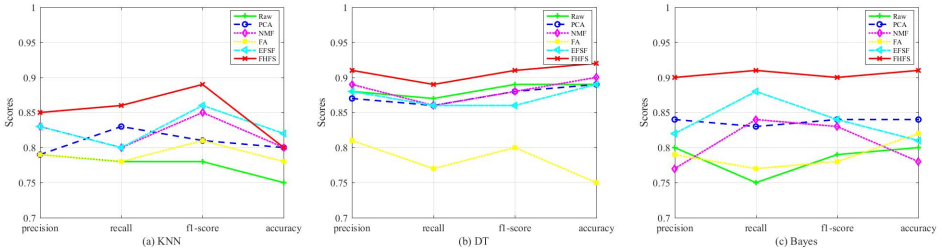

Fig. 5: Performance of different feature selection algorithms on Sensor dataset.

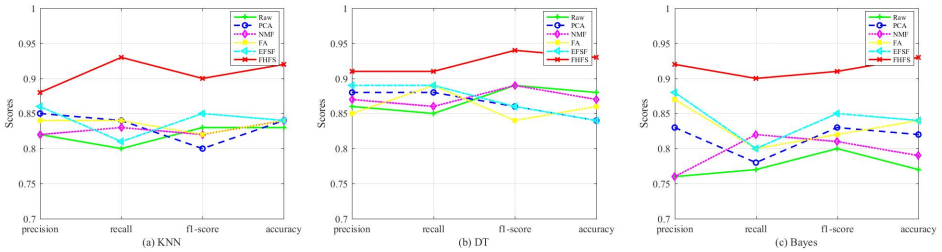

Fig. 6: Performance of different feature selection algorithms on Skin dataset.

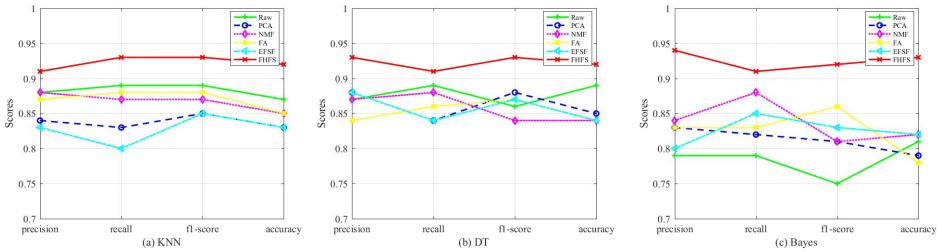

Fig. 7: Performance of different feature selection algorithms on Synthesis dataset.

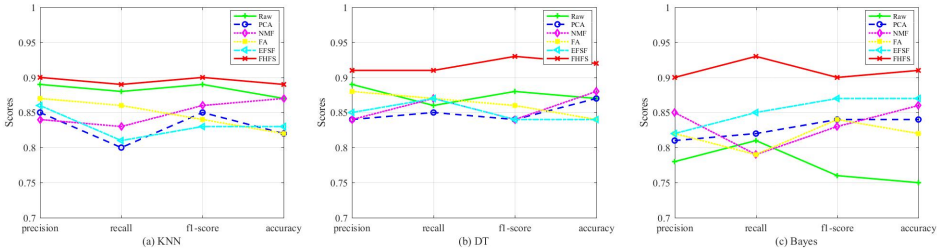

Fig. 8: Performance of different feature selection algorithms on Website dataset.

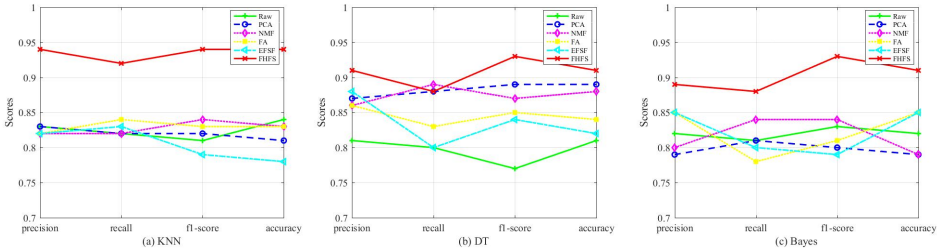

Fig. 9: Performance of different feature selection algorithms on Stock dataset.

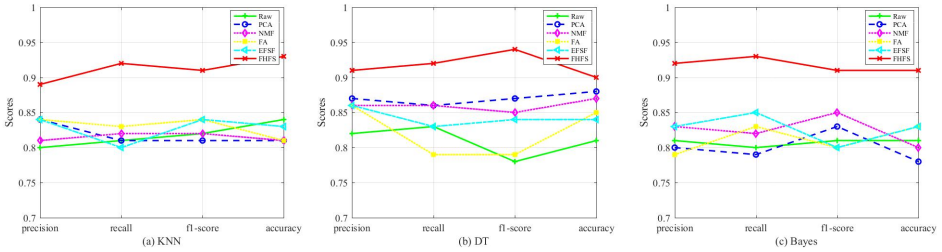

Fig. 10: Performance of different feature selection algorithms on Plants dataset.

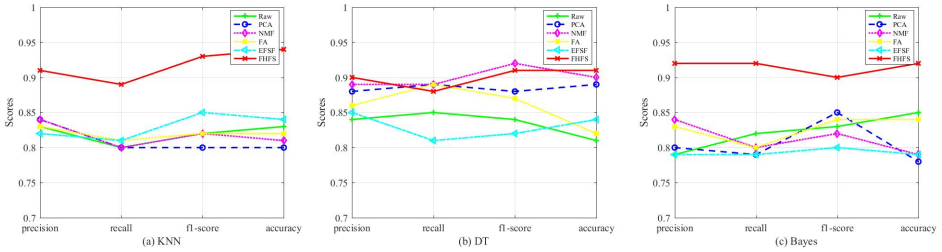

Fig. 11: Performance of different feature selection algorithms on Segmentation dataset.

TABLE VI: T-SCORES FOR HIGH DIMENSIONAL DATASETS

Classifier	Raw data	PCA	NMF	FA	EFSF	FHFS
KNN	4.39	3.93	3.92	3.76	3.74	3.96
DT	3.86	3.44	3.31	3.30	3.36	3.34
Bayes	4.13	3.40	3.42	3.31	3.34	3.36

B. Experimental design about image datasets

We collect animal image datasets from the UCI, each image is composed of 1024×1024 pixels. Detailed description of these image datasets is shown in the Table VII. Then, the algorithm of FCM and refinement of information granularity criterion is used to fulfill the process of image segmentation. Furthermore, comparisons among FHFS, algorithms from [33] (FCBF), [34] (CMIM), [35] (RFS), [36] (MIFS), [37] (m-RMR), [38] (Relief-f), are conducted in the process of image feature selection.

At the same time, four classifiers, namely KNN, Decision Tree (DT), Bayes, Convolutional Neural Network (CNN) are used to realize image recognition. In KNN, weighted voting method is used to find the nearest five samples of the tested sample to predict its label according to the Euclidean distance. In DT, the feature partition samples selection criteria is set to random. In Bayes, the assumption of conditional probability distribution of each attribute satisfying gaussian distribution is made before the process of machine learning. In CNN, the model of LeNet-5, including input layer, three convolution layers, two lower sampling layers and full connection layer is utilized. In this experiment, we also use the ten-fold cross validation method. The calculated results are expressed in the form of $\mu \pm \sigma$. The experimental results are shown in Tables VIII-IX.

From the results displayed in in Tables VIII-IX, after using FHFS, the image classification performance is improved compared with other algorithms. We have the following analysis: (1) In the classifiers KNN and DT, after using FHFS algorithm for feature selection, the classifiers based on different classification algorithms perform all better than other feature selection algorithms in the indexes of precision, recall, f1-score and accuracy. In the KNN classifier, the classification accuracy of the data processed by FHFS is improved by 19% compared with the raw data and 2% compared with the RFS

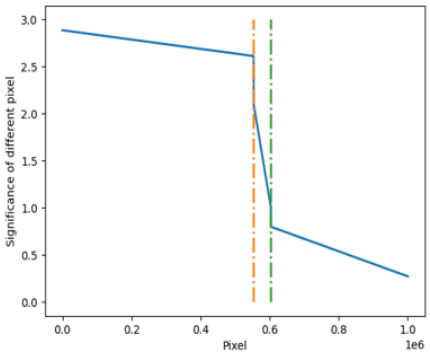

Fig. 14: Significance of each pixel.

algorithm. (2) In the classifiers Bayes and CNN, although the classification performance after using FHFS algorithm is not completely superior to other algorithms in the indexes of precision, recall, f1-score and accuracy, the classification performance after using FHFS is still at a high level. The classification f1-score after using FHFS is 1% lower than that after using RFS, but it is still better than that after using other algorithms in other index.

TABLE VII: DESCRIPTION OF IMAGE DATASETS

Dataset	Number of training set	Number of testing set	Quantity
cat	478	120	598
dog	485	122	607
elephant	4913	1229	6142
horse	5616	1405	7021
panda	4786	1197	5983
bear	8171	2043	10214
tiger	10001	2501	12502
lion	14427	3607	18034
Total	48877	12224	61101

In Fig. 13, red area in the images indicates unselected pixels as the results of feature selection. It is easy to observe that after applying the FHFS method to the feature screening in image segmentation, almost all the important features could be retained. However, other methods will lose some important features more or less. The FHFS performs better than other algorithms in the process of feature extraction. In fact, FHFS algorithm can select more valuable feature to support image recognition. Moreover, this algorithm is good for reducing the response time of image recognition. Moreover, high precision and accuracy of image recognition can be easily obtained by FHFS method.

As shown in Fig. 14, the pixels of image are nearly divided into three parts according to the measurement of significance. Almost 55% of pixels are attached to high significance, 40% of pixels are accompanied with low significance, and 5% of pixels are situated in the mutation region of significance. Based on plenty of experimental results, we can obtained that when we select high pixels to conduct the process of image recognition, excellent performance of feature selection and recognition can be obtained.

TABLE X: T-SCORES FOR IMAGE DATASETS

Classifier	Raw data	MIFS	Relief-f	mRMR	FCBF	CMIM	RFS
KNN	4.11	3.14	3.01	3.28	3.86	3.11	3.21
DT	4.04	3.15	3.02	3.14	3.37	3.26	3.37
Bayes	3.82	3.75	3.44	3.67	3.17	3.12	3.64
CNN	4.89	6.94	3.57	4.04	6.09	5.74	6.15

From the results of significance test ($n = 14$) on the image dataset shown in Table X, the FHFS algorithm is not superior to other feature selection algorithms (less than 0.05), which indicates that the original hypothesis (h_0) is not tenable.

C. Discussion of parameter

A positive weight factor α $(\alpha \geq 0)$ is an important parameter which primarily effects the BRIG, and can be reflected in the process of image segmentation. In this part, the effects of weight factor α assuming values in different range will be discussed, and performance of changed α is shown on the results of image segmentation and image recognition.

From Fig. 15 and Fig. 16, effects of weight factor on image segmentation and information granularity can be reflected, and the regularity of BRIG can also be comprehended by the experimental results. When $\alpha \in [0.1, 0.5]$, the BRIG just satisfies symmetry, but does not satisfy reflexivity, because in this range, some pixels do not belong to any of the information granules. Moreover, some pixels belong to multiple information granules at the same time. When $\alpha \in (0.5, +\infty)$, BRIG satisfies symmetry and reflexivity simultaneously, because in this range, all pixels belong to one or more of these information granules. Therefore, information granularity which are already generated can constitute the coverage of the universe. In this case, BRIG is a kind of similarity relation. Furthermore, when $\alpha \in [1.2, 1.5]$, just 5% pixels belong to multiple information granularity at the same time, information granularity can constitute the partition of the universe approximately. In this case, BRIG is almost an equivalence relation, and the results of image segmentation and recognition is excellent under this condition.

Based on the discussion of weight factor, the experiment of image recognition is conducted about cat dataset with different value of α, and the algorithm FHFS is used in the process of preprocessing. From Fig. 17, when $\alpha \in [1.2, 1.5]$, measurements of image recognition can be maintained in a high range. It indicates that the value of α and different BRIG can lead to performance of image recognition directly.

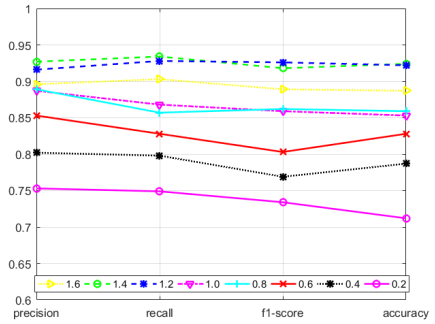

Fig. 17: Image recognition with different weight factors

Fig. 13: Results of image segmentation and feature selection

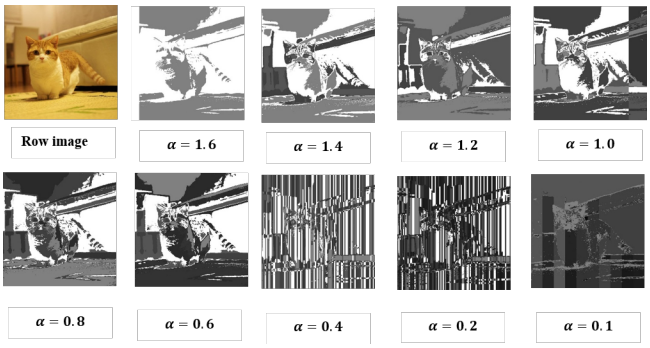

Fig. 15: Different performance of image segmentation with different weight factor

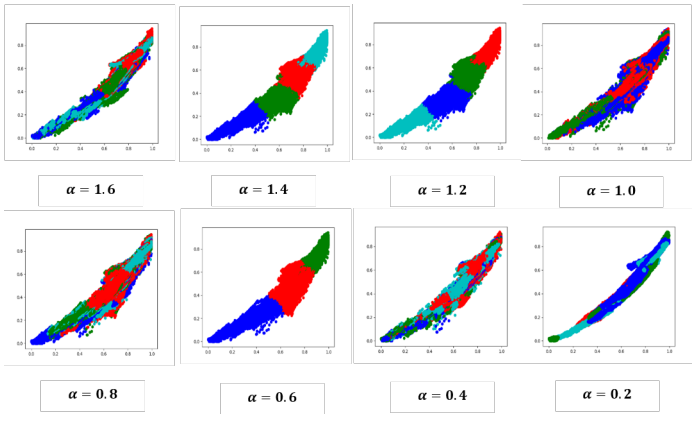

Fig. 16: Results of justifiable information granularity with different weight factor

TABLE VIII: RESULTS FOR IMAGE RECOGNITION WITH DIFFERENT CLASSIFIERS (I)

Algorithm of feature selection	KNN				DT			
	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
Raw data	0.60 ± 0.01	0.63 ± 0.03	0.64 ± 0.02	0.64 ± 0.01	0.60 ± 0.03	0.64 ± 0.02	0.61 ± 0.03	0.60 ± 0.03
MIFS	0.68 ± 0.02	0.63 ± 0.04	0.63 ± 0.03	0.62 ± 0.04	0.64 ± 0.01	0.68 ± 0.01	0.65 ± 0.02	0.62 ± 0.01
Relief-f	0.63 ± 0.04	0.62 ± 0.03	0.66 ± 0.02	0.68 ± 0.01	0.65 ± 0.03	0.65 ± 0.01	0.67 ± 0.03	0.64 ± 0.03
mRMR	0.64 ± 0.02	0.62 ± 0.01	0.68 ± 0.04	0.65 ± 0.04	0.66 ± 0.04	0.63 ± 0.03	0.65 ± 0.04	0.64 ± 0.02
FCBF	0.63 ± 0.04	0.64 ± 0.02	0.64 ± 0.04	0.66 ± 0.04	0.61 ± 0.04	0.68 ± 0.02	0.67 ± 0.02	0.69 ± 0.02
CMIM	0.68 ± 0.01	0.72 ± 0.03	0.71 ± 0.01	0.71 ± 0.03	0.72 ± 0.02	0.73 ± 0.01	0.67 ± 0.03	0.68 ± 0.02
RFS	0.77 ± 0.03	0.76 ± 0.01	0.74 ± 0.04	0.77 ± 0.02	0.74 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	$\mathbf{0.79 \pm 0.01}$
FHFS	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.79 \pm 0.04}$	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.79 \pm 0.01}$	0.78 ± 0.03

TABLE IX: RESULTS FOR IMAGE RECOGNITION WITH DIFFERENT CLASSIFIERS (II)

Algorithm of feature selection	Bayes				CNN			
	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
Raw data	0.68 ± 0.03	0.70 ± 0.01	0.75 ± 0.03	0.74 ± 0.01	0.78 ± 0.02	0.77 ± 0.03	0.78 ± 0.03	0.72 ± 0.04
MIFS	0.73 ± 0.03	0.70 ± 0.04	0.73 ± 0.01	0.71 ± 0.01	0.84 ± 0.04	0.84 ± 0.01	0.83 ± 0.02	0.85 ± 0.02
Relief-f	0.77 ± 0.02	0.78 ± 0.02	0.79 ± 0.01	0.77 ± 0.02	0.88 ± 0.02	0.88 ± 0.03	0.86 ± 0.02	0.88 ± 0.02
mRMR	0.71 ± 0.02	0.72 ± 0.03	0.71 ± 0.03	0.76 ± 0.01	0.83 ± 0.01	0.88 ± 0.02	0.84 ± 0.02	0.87 ± 0.03
FCBF	0.76 ± 0.03	0.74 ± 0.01	0.76 ± 0.04	0.72 ± 0.03	0.77 ± 0.01	0.76 ± 0.01	0.75 ± 0.04	0.81 ± 0.01
CMIM	0.72 ± 0.03	0.70 ± 0.03	0.74 ± 0.03	0.71 ± 0.02	0.83 ± 0.02	0.83 ± 0.03	0.83 ± 0.04	0.82 ± 0.02
RFS	0.74 ± 0.02	0.77 ± 0.01	0.81 ± 0.02	0.80 ± 0.02	0.85 ± 0.02	0.86 ± 0.02	$\mathbf{0.91 \pm 0.02}$	0.91 ± 0.02
FHFS	$\mathbf{0.81 \pm 0.01}$	$\mathbf{0.84 \pm 0.04}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.87 \pm 0.03}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	0.90 ± 0.03	$\mathbf{0.93 \pm 0.03}$

D. Time complexity analysis and comparison

For step 7 in Algorithm 1, the time complexity for computing the distance between samples is $O(d)$. Step 10 compute the updated prototypes $\{\nu_1, \nu_2, ..., \nu_c\}$, and the time complexity is $O(dc)$. The time complexity of steps 6-11 is $O(ndc^2)$, where data elements and clusters are traversed. In summary, the time complexity of Algorithm 1 is $O(ndc^2t)$, where n indicates the number of samples, d indicates the dimension of the data, c indicates the number of clusters, t represents the number of convergence iterations of the algorithm. For Algorithm 2, the time complexity of steps 6-10, 4-17 are respectively, $O(n)$, $O(tn)$. In summary, the time complexity of Algorithm 2 is $O(cpn)$, where c indicates the number of clusters, p indicates the number of the set P, n represents the number of samples. As to Algorithm 3, for the $(i+1)th$ traversal, the time complexity from step 3 to step 6 is $O(|C| - i)$. In summary, the time complexity of Algorithm 3 is $O(|C|^2 + |C|)$.

Then the total time complexity of our method including Algorithms 1-3 is $O(ndc^2t + cpn + |C|^2 + |C|)$. The time complexity comparison between the proposed method and other feature selection algorithms is shown in Table XI.

TABLE XI: ALGORITHMS COMPLEXITY

Algorithm	Time Complexity
PCA	$O(nd^2 + d^3)$
NMF	$O(tdnk + n^2d + d^2)$
FA	$O(dn^2)$
EFSF	$O(nd^2)$
MIFS	$O(n^2)$
Relief-f	$O(tnd^2)$
mRMR	$O(ndk^2)$
FCBF	$O(nd^3)$
CMIM	$O(nd^2)$
RFS	$O(nd^2log(d))$
Our method	$O(ndc^2t + cpn + \|C\|^2 + \|C\|)$

E. Robustness analysis

Based on the above experiments, we analyze the robustness of the proposed FHFS algorithm as follows. We randomly add noise to the data and delete the data to make the original data set change. Meanwhile, we directly use the classification algorithm and the classification algorithm after using FHFS to calculate the classification accuracy of the obtained data. Repeat the above operation for ten times and calculate the standard deviation. From Fig. 18 and Fig. 19, we intuitively observe that the data after processing of the FHFS exhibits better robustness than the original data, and the FHFS does not change the distribution of the original data.

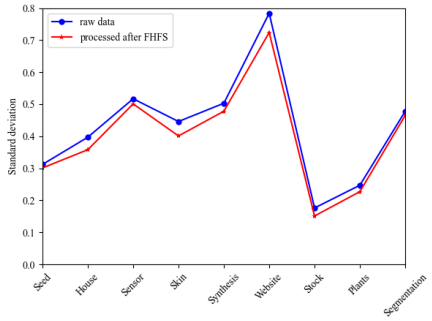

Fig. 18: Robustness of FHFS on high-dimensional dataset.

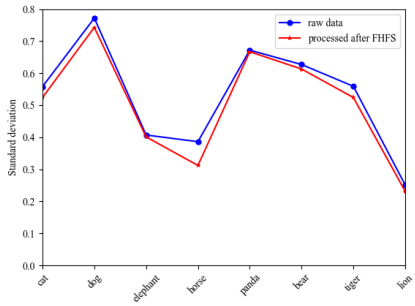

Fig. 19: Robustness of FHFS on image dataset.

VI. Conclusions

Based on the improved FCM algorithm and justifiable information granularity, a novel binary relation called BRIG is proposed to improve the limitations of the traditional FCM method. The new BRIG delivers a significant way to conduct the process of data-preprocessing during the task of machine learning, image recognition, and data mining. Meanwhile, criteria based on the degrees of aggregation and dispersion for measuring the importance of attributes are developed. Different effects and results can be obtained by resetting the weight factor, which has been displayed obviously in the process of image segmentation. Furthermore, rules of improved FCM-based FHFS with justifiable granularity are provided, and the related algorithm for FHFS is derived. The FHFS exhibits extraordinary performance during data dimensionality reduction compared with other algorithms in the same category. The selection based on the measurement of attribute importance is good for selecting valuable and important attributes, which is performed significantly in processing high dimensional information system and image feature extraction. The generalization ability of the feature selection algorithm proposed in this paper is mainly reflected by testing different datasets and studying the variance in the test results. This paper mainly focuses on improving the FCM clustering through the principle of justifiable granularity, in the future work, we will take into account the different clustering algorithms, focus on the variant of the proposed algorithm, and study the feature selection methods for heterogeneous data with different justifiable granularity classifiers.

References

[1] Z. Yuan, H. Chen, P. Zhang, J. Wan and T. Li, "A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information," *IEEE Trans. Fuzzy Syst.*, 2021, DOI: 10.1109/TFUZZ.2021.3114734.

[2] T.M. Nguyen and Q.M.J. Wu, "Online feature selection based on fuzzy clustering and its applications," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 6, pp. 1294-1306, 2016.

[3] W. Li, H. Zhou, W. Xu, X.Z. Wang, W. Pedrycz, "Interval dominance-based feature selection for interval-valued ordered data," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, DOI: 10.1109/TNNLS.2022.3184120.

[4] N. V. Kumar and D. S. Guru, "A novel feature ranking criterion for supervised interval valued feature selection for classification," *14th Int. Conf. Doc. Analys. Recogn.*, pp. 71-76, 2017.

[5] F. Jimnez, C. Martlnez, E. Marzano, J.T. Palma, G. Snchez and G. Sciavicco, "Multiobjective evolutionary feature selection for fuzzy classification," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 5, pp. 1085-1099, 2019.

[6] Y. Lin, Q. Hu, J. Liu, J. Li and X. Wu, "Streaming feature selection for multilabel learning based on fuzzy mutual information," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1491-1507, 2017.

[7] H. Zhao, P. Wang, Q. Hu and P. Zhu, "Fuzzy rough set based feature selection for large-scale hierarchical classification," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 10, pp. 1891-1903, 2019.

[8] W. Pedrycz, "Granular computing analysis and design of intelligent systems," *CRC Press Taylor & Francis Group*, 2013.

[9] L.A. Zadeh, "Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 19, pp. 111-127, 1997.

[10] Y. Qian, J. Liang, W.Z. Wu and C. Dang, "Information granularity in fuzzy binary GrC model," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 2, pp. 253-264, 2011.

[11] X. Zhu, W. Pedrycz and Z. Li, "A development of hierarchically structured granular models realized through allocation of information granularity," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3845-3858, 2021.

[12] T. T. Nguyen, X. C. Pham, A. W. Liew and W. Pedrycz, "Aggregation of classifiers: a justifiable information granularity approach," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2168-2177, 2019.

[13] F. Liu, Y. Wu and W. Pedrycz, "A modified consensus model in group decision making with an allocation of information granularity," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3182-3187, 2018.

[14] X. Zhu, W. Pedrycz and Z. Li, "Granular data description: designing ellipsoidal information granules," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4475-4484, 2017.

[15] P.A. Kowalski and E. Jeczmionek, "Parallel complete gradient clustering algorithm and its properties," *Inf. Sci.*, vol. 600, pp. 155-169, 2022.

[16] A. Gupta and S. Das, "On efficient model selection for sparse hard and fuzzy center-based clustering algorithms," *Inf. Sci.*, vol. 590, pp. 29-44, 2022.

[17] C. Wu and X. Zhang, "A novel kernelized total Bregman divergence-based fuzzy clustering with local information for image segmentation," *Int. J. Approx. Reason.*, vol. 136, pp. 281-305, 2021.

[18] V. Antoine, J.A. Guerrero and J. Xie, "Fast semi-supervised evidential clustering," *Int. J. Approx. Reason.*, vol. 133, pp. 116-132, 2021.

[19] F. Nie, W. Zhu and X. Li, "Unsupervised large graph embedding based on balanced and hierarchical K-means," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 2008-2019, 2022.

[20] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123-1136, 2019.

[21] G. Huo, S. X. Yang, Q. Li and Y. Zhou, "A robust and fast method for sidescan sonar image segmentation using nonlocal despeckling and active contour model," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 855-872, 2017.

[22] K. Zhang, Q. Liu, H. Song and X. Li, "A variational approach to simultaneous image segmentation and bias correction," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1426-1437, 2015.

[23] K. Zhang, L. Zhang, K.M. Lam and D. Zhang, "A level set approach to image segmentation with intensity inhomogeneity," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 546-557, 2016.

[24] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng and A.K. Nandi, "Superpixel-based fast fuzzy c-means clustering for color image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753-1766, 2019.

[25] J.K. Parker and L.O. Hall, "Accelerating fuzzy c-means using an estimated subsample size," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 5, pp. 1229-1244, 2014.

[26] X. Hu, Y. Shen, W. Pedrycz, X. Wang, X. Wang, A. Gacek and B. Liu, "Identification of fuzzy rule-based models with collaborative fuzzy clustering," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6406-6419, 2022.

[27] W. Lu, C. Ma, W. Pedrycz and J. Yang, "Design of granular model: a method driven by hyper-box iteration granulation," *IEEE Trans. Cybern.*, 2021, DOI: 10.1109/TCYB.2021.3124235.

[28] H. Ju, W. Ding, X. Yang, H. Fujita, S. Xu. "Robust supervised rough granular description model with the principle of justifiable granularity," *Applied Soft Comput.*, vol. 110, no. 107612, pp. 1-19, 2021.

[29] N.Z. Joodaki, M. Bagher Dowlatshahi and M. Joodaki, "A novel ensemble feature selection method through Type I fuzzy," *9th Iranian Joint Congr. Fuzzy Intell. Syst.*, pp. 1-6, 2022.

[30] C. Xu, H. Jiang and J. Yu, "Robust two-dimensional principle component analysis," *27th Chinese Contr. Conf.*, pp. 452-455, 2008.

[31] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[32] H. Akaike, "Factor analysis and AIC," *Psychometrika*, vol. 52, no. 3, pp. 317-332, 1987.

[33] C. Ying, G. Wang and H. Li, "Design of feature selection algorithm based on improved FCBF," *6th Int. Conf. Intell. Comput. Signal Proc.*, pp. 323-327, 2021.

[34] T. Sylvain, F. Dutil, T. Berthier, L.D. Jorio, M. Luck, D. Hjelm and Y. Bengio, "CMIM: cross-modal information maximization for medical imaging" *Int. Conf. Acoust. Speech Sign. Process.*, pp. 1190-1194, 2021.

[35] S. Dong, Y. Quan, W. Feng, Q. Li, G. Dauphin and M. Xing, "Ensemble CNN based on pixel-pair and random feature selection for hyperspectral image classification with small-size training set," *Int. Geoscienc. Rem. Sens. Symp.*, pp. 2353-2356, 2021.

[36] H.E. Maia, A. Hammouch and M. Bakrim, "Color texture feature selection by MIFS for image classification," *5th Int. Symp. I/V Commun. Mobile Network*, pp. 1-4, 2010.

[37] Y. Sun, L. Ma, N. Qin, M. Zhang and Q. Lv, "Analog filter circuits feature selection using MRMR and SVM," *14th Int. Conf. Contr., Autom. Syst.*, pp. 1543-1547, 2014.

[38] Y. Wang and F. Makedon, "Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data," *IEEE Comput. Syst. Bioinf. Conf.*, pp. 497-498, 2004.