Instance and Feature Selection Using Fuzzy Rough Sets: A Bi-Selection Approach for Data Reduction

Xiao Zhang, Changlin Mei, Jinhai Li, Yanyan Yang, and Ting Qian

***Abstract*—Data reduction, aiming to reduce the original data by selecting the most representative information, is an important technique of preprocessing data. At present, large-scale or huge data are very common and the development of data reduction techniques for such data has attracted much attention. As a powerful tool for handling uncertainty in real-valued data, the fuzzy rough set theory has been widely applied to data reduction including extensive feature selection methods and some instance selection approaches. Nevertheless, not much work has been devoted to the simultaneous selection of feature and instance based on fuzzy rough sets. In this paper, we investigate the fuzzy rough set-based bi-selection issue for data reduction. Specifically, the unified concepts of the importance degrees of fuzzy granules are presented to select the representative instances first and then the critical features. An instance selection algorithm with a noise elimination technique is provided to firstly remove the noise and then select the representative instances according to the importance degrees of fuzzy granules. Then, the importance-degree-preserved attribute reduction is proposed, and a corresponding feature selection algorithm with a wrapper technique is given to search for a best feature subset. Lastly, the bi-selection method based on fuzzy rough sets (BSFRS) is presented for data reduction by integrating the instance selection and the feature selection methods. Moreover, some numerical experiments are conducted to assess the performance of BSFRS, and the results show that BSFRS performs well in terms of the effectiveness.**

***Index Terms*—fuzzy rough sets, data reduction, instance selection, feature selection.**

I. Introduction

DUE to the rapid development of society, larger and larger observations and simulations are collected in many fields, such as engineering, economics, biology and so on. Dealing with huge data brings great challenges to data mining, machine learning and pattern recognition.

Data reduction [1] is an important data preprocessing task. The objective of data reduction is to reduce the original data by selecting the most representative information. The obvious advantages of data reduction focus on alleviating time complexity and avoiding excessive storage. On the other hand, the simplified models could be built by the reduced data and the analysis results may be improved. The well known data reduction techniques include instance selection [2], [3], feature selection [4], instance generation [5], attribute discretization [6] and so on.

Instance selection, as one of the important techniques for data reduction, has been used to choose a subset of data to achieve the original purpose of a data mining application as if the whole data is used [2]. Via instance selection, one can usually obtain the representative instances by removing redundant instances, irrelevant instances or errors from original data, which helps to mine critical information and easily acquire high quality results with less computation time especially for the large-scale data. The issue of instance selection has been studied in many application domains [7]–[13]. Feature selection, as another important technique for data reduction, has been used for both dimensionality reduction and learning performance improvement by removing redundant or irrelevant features (attributes). Many relevant algorithms with different search strategies and evaluation functions have been proposed for feature selection [14]–[17]. In addition, some researches on simultaneous feature and instance selection have been developed, which mainly bases on genetic algorithms [18]–[25]. For example, Tsai *et al.* [18] performed feature and instance selection based on genetic algorithms using different priorities. Ros *et al.* [21] proposed a hybrid genetic approach to select a subdatabase of the original one. However, it should be pointed out that the simultaneous feature and instance selection techniques using genetic algorithms generally need to be pre-specified multiple parameters and take much time to be implemented.

Traditional rough set theory is originally proposed by Pawlak [26] as a mathematical tool for data analysis and knowledge discovery. The theory is powerful in discovering knowledge in such databases that are described by nominal attributes. As an extension of the traditional rough set theory, fuzzy rough sets have been proposed to deal with real-valued or even mixed data [27], [28], and have been successfully applied in the fields of both data mining and machine learning [29]–[33]. One of the main applications of fuzzy rough sets is data reduction including feature selection (often called attribute reduction) [34]–[44], instance selection [45]–[47], and so on.

The existing fuzzy-rough-set-based attribute reduction methods may be mainly partitioned into the discernibility matrix methods [39], [40], the fuzzy-dependency-function-based heuristic algorithms [31], [41], [42], and the fuzzy-information-entropy-based heuristic algorithms [43], [44]. Rel-

This work was supported by the National Natural Science Foundation of China under Grant 12171388, Grant 12271420, Grant 11971211, and Grant 12171386, and the Natural Science Foundation of Shaanxi Province of China under Grant 2021JQ-465. *(Corresponding author: Xiao Zhang)*

X. Zhang is with the Department of Applied Mathematics, Xi'an University of Technology, Xi'an 710054, China (e-mail: zhangxiao@xaut.edu.cn).

C.L. Mei is with the Department of Finance and Statistics, Xi'an Polytechnic University, Xi'an 710048, China (e-mail: clmei@xpu.edu.cn).

J.H. Li is with the Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China (e-mail: jhlixjtu@163.com).

Y.Y. Yang is with the School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: yangyy@bjtu.edu.cn).

T. Qian is with the College of Science, Xi'an Shiyou University, Xi'an 710065, China (e-mail: qiant2000@126.com).

ative to the much research on fuzzy-rough-set-based feature selection, there exists less work on the instance selection. A preliminary work of instance selection based on fuzzy rough sets can be found in [45], which selects the instances with the membership to the fuzzy positive region being not less than a pre-specified threshold. Verbiest *et al.* [46] put forward the fuzzy rough prototype selection method in which a fuzzy rough measure was used to characterize the quality of the instances, and a wrapper approach was provided to determine the selected instances. Tsang *et al.* [47] designed a weighted sampling technique to select the representative instances for K-nearest neighbor rule (KNN rule). Recently, Zhang *et al.* [48] presented a fuzzy-rough-set-based method of searching for a representative instance set according to the discriminating ability of the fuzzy granular rules in a fuzzy decision system. Nevertheless, removing the noise from the original data is not considered in this method. That's to say, the obtained representative instance set may include the noise. Moreover, based on fuzzy rough sets, Zhang *et al.* [49] proposed an approach to select the representative instances from an incoming instance set in dynamic environment, and then used the incoming representative instance set rather than the whole incoming instance set to accomplish the incremental feature selection so as to reduce the computation time and complexity.

It should be pointed out that there exists some fuzzy-rough-set-based work on the simultaneous feature and instance selection [19], [50], [52], [53] which focuses primarily on the intelligent optimization algorithms. Here, intelligent optimization algorithm is a kind of optimization algorithm that simulates natural phenomena and behaviors with population-based iterations [54]. Specifically, Derrac *et al.* [19] presented such a steady-state genetic algorithm that is added in a fuzzy-rough-set-based feature selection process to select the instances. Via a shuffled frog leaping algorithm, Anaraki *et al.* [50] proposed a simultaneous feature and instance selection approach based on fuzzy rough sets. The concept of a bireduct was firstly put forward in [51], which is an extension of the notion of a reduct of rough sets. A bireduct includes both some instances and attributes, and seems to be a class of classification rules. As the extension of the work in [51], simultaneous feature and instance selection using fuzzy rough sets was investigated in [52], and a corresponding algorithm with a frequency-based approach taking as a heuristic was designed to select the features or instances alternatively. Furthermore, the fuzzy-rough-set-based bireducts were further researched in [53], and a harmony-search-based algorithm was provided to discover the bireducts.

The motivation of this paper is twofold. On the one hand, it's known that the fuzzy granules generated by fuzzy rough sets imply the common information of some similar instances with respect to a given attribute set. Thus, the fuzzy granule can be seemed as a cluster of similarity information. If the importance degrees of the fuzzy granules could be numerically evaluated, then the representative instances may be intuitively obtained according to the magnitude of the importance degrees, and the redundant attributes may be also reduced via the importance degrees. Then, the process of the instance and feature selection will be not randomized iterative as intelligent optimization algorithms. Nevertheless, at present, the instance and feature selection using fuzzy rough sets has not been researched from the viewpoint of the importance degrees of the fuzzy granules. On the other hand, it is worth noting that the presence of errors is likely to make the mining algorithms applied on the whole data inefficient. However, the noise elimination technique has not been considered in either the existing fuzzy-rough-set-based instance selection methods or the simultaneous feature and instance selection methods. Thus, a noise elimination technique is suggested to add in the corresponding instance selection procedure.

In this paper, we investigate the bi-selection of instance and feature using fuzzy rough sets for data reduction. Firstly, the concept of the importance degree of a fuzzy granule is proposed. Via the importance degrees of the fuzzy granules, an instance selection method with a noise elimination technique is provided. Secondly, an importance-degree-preserved attribute reduction is presented. Then, a corresponding feature selection algorithm is given, and a wrapper technique of searching for a best feature subset is provided. Thirdly, a bi-selection method based on fuzzy rough sets (BSFRS) is put forward for data reduction by integrating the instance selection method and the importance-degree-preserved feature selection method. Lastly, the performance of BSFRS is empirically evaluated by some numerical experiments.

The remainder of the paper is organized as follows. Some basic knowledge of fuzzy rough sets is introduced in Section II. In Section III, an instance selection method with a noise elimination technique is put forward according to the importance degrees of the fuzzy granules. In Section IV, an importance-degree-preserved feature selection method is presented, and a wrapper technique of searching for a best feature subset is provided. Then, BSFRS is formulated in Section V. In Section VI, some numerical experiments are conducted to assess the performance of BSFRS. Finally, this paper is ended with a summary and the future work in Section VII.

II. Preliminaries

In order to facilitate the subsequent discussions, we first introduce some basic knowledge about fuzzy rough sets in this section.

A. Fuzzy Rough Sets

Let U be a nonempty universe of discourse and $F(U \times U)$ be the fuzzy power set on $U \times U$. R is called a fuzzy relation on $U \times U$ if $R \in F(U \times U)$, where $R(x, y)$ measures the strength of relationship between $x \in U$ and $y \in U$. A fuzzy relation R is reflexive if $R(x_i, x_i) = 1$ for any $x_i \in U$; R is symmetric if $R(x_i, x_j) = R(x_j, x_i)$ for any $x_i, x_j \in U$; and R is T-transitive if $R(x_i, x_j) \geq T(R(x_i, x_k), R(x_k, x_j))$ for a triangular norm T and any $x_i, x_j, x_k \in U$. If R is reflexive, symmetric and T-transitive, R is called a T-similarity relation . Specially, if $T = \min$, R is called a fuzzy equivalence relation.

A pair of lower and upper approximation operators of a fuzzy set X based on a T-similarity relation R is defined in

[27], for each $x_i \in U$, as

$$\underline{R}X(x_i) = \inf_{x_j \in U} \max\{1 - R(x_i, x_j), X(x_j)\} \tag{1}$$

and

$$\overline{R}X(x_i) = \sup_{x_j \in U} \min\{R(x_i, x_j), X(x_j)\} \tag{2}$$

to measure the degree of x_i certainly belonging to X and the degree of x_i possibly belonging to X, respectively. Then, the fuzzy rough set of X is defined by $(\underline{R}X, \overline{R}X)$. Some other general fuzzy approximation operators are also presented and one can refer to [28], [55]–[57]. Since the fuzzy approximation operators (1) and (2) are pioneering pair which have been extensively used in fuzzy rough sets, the work of this paper also takes the operators (1) and (2). Moreover, it should be pointed out that the fuzzy relations in the operators (1) and (2) in our subsequent work only need to satisfy reflexivity and symmetry.

B. Fuzzy information systems and fuzzy decision systems

A fuzzy information system is a pair (U, A) in which $U = \{x_1,\ x_2,\ \ldots,\ x_n\}$ is the universe of discourse and $A = \{a_1,\ a_2,\ \ldots,\ a_m\}$ is the attribute set. For each attribute $a_t \in A$, a mapping $a_t : U \to V_{a_t}$ holds where V_{a_t} is the domain of a_t, and a fuzzy relation $R_{\{a_t\}}$ can be defined. Here, the fuzzy relation $R_{\{a_t\}}$ is a fuzzy set that is defined on the fuzzy power set $F(U \times U)$, and $R_{\{a_t\}}(x_i, x_j)$ is used to reflect the similarity degree between the objects x_i and x_j with respect to the attribute a_t. Furthermore, the fuzzy relation of a subset $B \subseteq A$ is defined by $R_B = \bigcap_{a_t \in B} R_{\{a_t\}}$.

Let (U, A) be a fuzzy information system. By adding an attribute set $D = \{d\}$ with $A \cap D = \emptyset$ into $(U,\ A)$, we obtain a fuzzy decision system $(U,\ A \cup D)$ where A is called the conditional attribute set and D is called the decision attribute set. It should be pointed out that the decision attribute d is nominal with a mapping $d : U \to V_d$ being held. Here, V_d is the domain of the decision attribute d. Then, an equivalence relation R_D can be defined for d, and the universe U is partitioned by R_D into a family of disjoints subsets $U/D = \{[x_i]_D : x_i \in U\}$, where $[x_i]_D = \{x_j \in U : d(x_i) = d(x_j)\}$ is called the decision class to which the object x_i belongs. Furthermore, the membership function of the crisp set $[x_i]_D$ is

$$[x_i]_D(x_j) = \begin{cases} 1, x_j \in [x_i]_D; \\ 0, \text{otherwise.} \end{cases}$$

Therefore, for a fuzzy decision system $(U,\ A \cup D)$ with $U = \{x_1,\ x_2,\ \ldots,\ x_n\}$ and $B \subseteq A$, we then have, for each $x_i \in U$,

$$\underline{R_B}[x_j]_D(x_i) = \begin{cases} \inf\limits_{x_k \notin [x_j]_D} \{1 - R_B(x_i, x_k)\}, & x_i \in [x_j]_D; \\ 0, & \text{otherwise.} \end{cases}$$

and

$$\overline{R_B}[x_j]_D(x_i) = \begin{cases} 1, & x_i \in [x_j]_D; \\ \sup\limits_{x_k \in [x_j]_D} \{R_B(x_i, x_k)\}, & \text{otherwise.} \end{cases}$$

C. Fuzzy granules

Let $(U, A \cup D)$ be a fuzzy decision system with $U = \{x_1, x_2, \ldots, x_n\}$ and $B \subseteq A$. Each object or instance $x_i \in U$ can correspond to the fuzzy granule $[x_i]_B^{\lambda_i}$ with respect to B (one can refer to [39], [44]) as follows

$$[x_i]_B^{\lambda_i}(x_j) = \begin{cases} \lambda_i,\ 1 - R_B(x_i, x_j) < \lambda_i; \\ 0,\ \text{otherwise} \end{cases} \tag{3}$$

where $\lambda_i = \underline{R_A}[x_i]_D(x_i) > 0$. Here, $[x_i]_B^{\lambda_i}$ is called the fuzzy granule induced by x_i with respect to B. Then, the set of the fuzzy granules $[x_i]_B^{\lambda_i}$ induced by all the objects in U with respect to B is denoted by $GrS(U, B)$, i.e.,

$$\begin{aligned} GrS(U, B) = \quad & \{[x_i]_B^{\lambda_i} : x_i \in U, i = 1, 2, \ldots, n, \\ & B \subseteq A, \lambda_i = \underline{R_A}[x_i]_D(x_i)\}. \end{aligned} \tag{4}$$

It should be noticed that, for the two attribute subsets B and C, both the notations λ_i in $[x_i]_B^{\lambda_i} \in GrS(U, B)$ and λ_i in $[x_i]_C^{\lambda_i} \in GrS(U, C)$ represent the same value, i.e., $\lambda_i = \underline{R_A}[x_i]_D(x_i)$.

III. A Novel Instance Selection Method With a Noise Elimination Technique

As aforementioned in Introduction, deleting noise has not been considered in either the existing fuzzy-rough-set-based instance selection methods or the simultaneous feature and instance selection methods. Thus, a novel instance selection method with a noise elimination technique is put forward in this section.

In order to measure the importance degree of the fuzzy granule $[x_i]_B^{\lambda_i}$, the following definition is firstly proposed.

Definition 1. *Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$, and $[x_i]_B^{\lambda_i} \in GrS(U, B)$. For a given object $x_j \in U$, if $[x_i]_B^{\lambda_i}(x_j) > 0$, we say that the object x_j is covered by the fuzzy granule $[x_i]_B^{\lambda_i}$. The set of all the objects covered by $[x_i]_B^{\lambda_i}$ is denoted by*

$$Cov\left([x_i]_B^{\lambda_i}\right) = \left\{x_j : [x_i]_B^{\lambda_i}(x_j) > 0, x_j \in U\right\}. \tag{5}$$

According to Definition 1, we know that the objects in $Cov\left([x_i]_B^{\lambda_i}\right)$ can be discriminated by $[x_i]_B^{\lambda_i}$ from U. The larger $\left|Cov\left([x_i]_B^{\lambda_i}\right)\right|$ is, the more the objects covered by $[x_i]_B^{\lambda_i}$ are, which means that $[x_i]_B^{\lambda_i}$ may possess more powerful discriminating ability. Here, $|\cdot|$ is the cardinality of a set.

Proposition 1. *Let $(U, A \cup D)$ be a fuzzy decision system with $U = \{x_1, x_2, \ldots, x_n\}$. Given an arbitrary object $x_i \in U$, both $[x_i]_B^{\lambda_i} \subseteq [x_i]_C^{\lambda_i}$ and $\left|Cov\left([x_i]_B^{\lambda_i}\right)\right| \leqslant \left|Cov\left([x_i]_C^{\lambda_i}\right)\right|$ hold for $C \subseteq B \subseteq A$.*

Proof: For any object x_j which satisfies $[x_i]_B^{\lambda_i}(x_j) > 0$, we obtain $1 - R_B(x_i, x_j) < \lambda_i$. Since $R_C \supseteq R_B$ holds for $C \subseteq B$, we have $1 - R_C(x_i, x_j) \leqslant 1 - R_B(x_i, x_j)$ which yields $1 - R_C(x_i, x_j) < \lambda_i$ and then $[x_i]_C^{\lambda_i}(x_j) = \lambda_i > 0$. Thus, $[x_i]_B^{\lambda_i}(x_j) = [x_i]_C^{\lambda_i}(x_j)$.

On the other hand, for any object $x_{j'}$ which satisfies $[x_i]_B^{\lambda_i}(x_{j'}) = 0$, we have $1 - R_B(x_i, x_{j'}) \geq \lambda_i$. Since $1 - R_C(x_i, x_{j'}) \leqslant 1 - R_B(x_i, x_{j'})$, then either $1 - R_C(x_i, x_{j'}) <$

λ_i or $1 - R_C(x_i, x_{j'}) \geq \lambda_i$ holds, i.e., $[x_i]_C^{\lambda_i}(x_{j'}) = \lambda_i > 0$ or $[x_i]_C^{\lambda_i}(x_{j'}) = 0$. Thus, $[x_i]_B^{\lambda_i}(x_{j'}) \leqslant [x_i]_C^{\lambda_i}(x_{j'})$.

In summary, $[x_i]_B^{\lambda_i}(x_j) \leqslant [x_i]_C^{\lambda_i}(x_j)$ holds for any $x_j \in U$, which yields $[x_i]_B^{\lambda_i} \subseteq [x_i]_C^{\lambda_i}$. Then, we obviously have $\left|Cov\left([x_i]_B^{\lambda_i}\right)\right| \leqslant \left|Cov\left([x_i]_C^{\lambda_i}\right)\right|$. ■

It is easily known from Proposition 1 that the fuzzy granule $[x_i]_B^{\lambda_i}$ satisfies monotonicity. Specially, $[x_i]_{B\cup\{a\}}^{\lambda_i} \subseteq [x_i]_B^{\lambda_i}$ holds for any attribute $a \in A \setminus B$, which yields $\left|Cov\left([x_i]_{B\cup\{a\}}^{\lambda_i}\right)\right| \leqslant \left|Cov\left([x_i]_B^{\lambda_i}\right)\right|$.

Definition 2. *Let $(U, A \cup D)$ be a fuzzy decision system with $U = \{x_1, x_2, \ldots, x_n\}$. The importance degree of the fuzzy granule $[x_i]_A^{\lambda_i} \in GrS(U, A)$ is defined as*

$$F\left([x_i]_A^{\lambda_i}\right) = \lambda_i \cdot \left|Cov\left([x_i]_A^{\lambda_i}\right)\right|. \tag{6}$$

According to Definition 2, we know that the importance degree of the fuzzy granule $[x_i]_A^{\lambda_i}$ is determined by two factors: the fuzzy lower approximation value λ_i and the cardinality of the coverage object set $\left|Cov\left([x_i]_A^{\lambda_i}\right)\right|$. The larger the value $\left|Cov\left([x_i]_A^{\lambda_i}\right)\right|$ is, the more powerful discriminating ability of the fuzzy granule $[x_i]_A^{\lambda_i}$ possesses. Furthermore, λ_i can be used to reflect the determinacy of the discriminating ability of $[x_i]_A^{\lambda_i}$. Since the fuzzy granule $[x_i]_A^{\lambda_i}$ is induced by the instance x_i, the importance degree $F\left([x_i]_A^{\lambda_i}\right)$ can be also called the importance degree of the instance x_i.

Let $(U, A \cup D)$ be a fuzzy decision system and $U/D = \{D_1, D_2, \ldots, D_l\}$ be the decision partition. We put forward a novel instance selection method which includes two main procedures as what follows.

(1) Firstly, we consider to identify the noise in each decision class D_t $(t = 1, 2, \ldots, l)$ and remove the noise from D_t. Compute $Cov([x_i]_A^{\lambda_i})$ for each $x_i \in D_t$. Since the coverage ability of the noise is usually lower, we take the instance x_{i_0} with $\left|Cov([x_{i_0}]_A^{\lambda_{i_0}})\right| = 1$ as the potential possible noise in $(U, A \cup D)$. Then, we need to further determine whether x_{i_0} is noise or not. Specifically, we need to search for the k-nearest neighbors of x_{i_0}, namely the k instances that are closest to x_{i_0}. Here, the proximity of x_{i_0} to the other instances in U is defined by

$$Pro_A(x_{i_0}, x_i) = 1 - R_A(x_{i_0}, x_i), (x_i \in U, i \neq i_0), \tag{7}$$

where R_A is the fuzzy relation with respect to the conditional attribute set A. The set of the k-nearest neighbors of x_{i_0} is denoted by $knn(x_{i_0})$. If the number of the instances in $knn(x_{i_0})$ with different labels from the label of x_{i_0} is greater than $k/2$, the instance x_{i_0} is regarded as the noise. It should be pointed out that the label of an instance x_i is factually the value of the decision attribute d, i.e., $d(x_i)$. Additionally, the choice of the parameter k is suggested in the section of the numerical experiments.

(2) Secondly, we select the representative instances according to equation (6). For notational simplicity, we still use D_t $(t = 1, 2, \ldots, l)$ to denote the decision class in which the noise has been deleted. Considering that the importance degree of each instance can also be reflected by equation (6), for each instance $x_j \in D_t$, we compute $F\left([x_j]_A^{\lambda_j}\right)$ and select the instance x_{j_0} satisfying $\max\limits_{j\in\{1,2,\ldots,|D_t|\}} F\left([x_j]_A^{\lambda_j}\right)$. Then, the instances covered by $[x_{j_0}]_A^{\lambda_{j_0}}$ except x_{j_0} are removed from D_t. For each $x_{j'} \in Cov\left([x_{j_0}]_A^{\lambda_{j_0}}\right)$, let $Cov([x_{j'}]_A^{\lambda_{j'}}) = \emptyset$ and remove $x_{j'}$ from $Cov\left([x_j]_A^{\lambda_j}\right)$ $(x_j \notin Cov([x_{j_0}]_A^{\lambda_{j_0}}, j = 1, 2, \ldots, |D_t|)$. The procedure is repeated until $Cov\left([x_j]_A^{\lambda_j}\right) = \emptyset$ for each $x_j \in D_t (j = 1, 2, \ldots, |D_t|)$ which is equivalent to $\max\limits_{j\in\{1,2,\ldots,|D_t|\}} F\left([x_j]_A^{\lambda_j}\right) = 0$.

In summary, Algorithm 1 is provided to firstly delete the noise and then select the representative instances from an original data set.

Algorithm 1 Instance selection algorithm of deleting the noise and selecting the representative instances

Input: A fuzzy decision system $(U, A \cup D)$ with $U = \{x_1, x_2, \ldots, x_n\}$, the fuzzy lower approximation values $\lambda_i = \underline{R_A}[x_i]_D(x_i)$ $(i = 1, 2, \ldots, n)$, the decision partition $U/D = \{D_1, D_2, \ldots, D_l\}$, and a parameter k.

Output: A representative instance set U^* of $(U, A \cup D)$.

```
Initialize U* = ∅;
for t = 1 to l do
    D_temp = D_t
    for i = 1 to |D_t| do
        compute Cov([x_i]_A^{λ_i}) by equation (5);
        if there exists x_{i_0} such that |Cov([x_{i_0}]_A^{λ_{i_0}})| = 1 and
        |{x_i | x_i ∈ knn(x_{i_0}) ∧ d(x_i) ≠ d(x_{i_0})}| > k/2 then
            D_temp = D_temp − {x_{i_0}};
            let Cov([x_{i_0}]_A^{λ_{i_0}}) = ∅;
            remove x_{i_0} from Cov([x_i]_A^{λ_i}) (i = 1, 2, ..., |D_t|);
        end if
    end for
    for j = 1 to |D_temp| do
        compute F([x_j]_A^{λ_j}) by equation (6);
    end for
    imp_degree = max_{j∈{1,2,...,|D_temp|}} F([x_j]_A^{λ_j});
    while imp_degree > 0 do
        denote j_0 = arg max_j F([x_j]_A^{λ_j});
        D_temp = D_temp − Cov([x_{j_0}]_A^{λ_{j_0}}) \ {x_{j_0}};
        for each x_{j'} ∈ Cov([x_{j_0}]_A^{λ_{j_0}}) do
            Cov([x_{j'}]_A^{λ_{j'}}) = ∅;
            delete x_{j'} from Cov([x_j]_A^{λ_j})(x_j ∉ Cov([x_{j_0}]_A^{λ_{j_0}});
        end for
        imp_degree = max_j F([x_j]_A^{λ_j});
    end while
end for
U* = U* ∪ D_temp
return U*.
```

The time complexity of Algorithm 1 is polynomial. Given the decision class D_t, the complexity of computing $Cov([x_i]_A^{\lambda_i})$ $(x_i \in D_t)$ is $O(|D_t|^2|A|)$. The complexity

of running the noise elimination technique (Steps 6–10) is $O(|D_t||U||A|)$. Steps 12–24 is at most $O(|D_t|^2|A|)$. Furthermore, Steps 2–25 need to be run l times. Totally, the time complexity of Algorithm 1 is at most $O(l|D_t||U||A|)$.

There mainly exist two differences between Algorithm 1 and the instance selection method in [48]. On the one hand, the noise is firstly removed (see Steps 6–10) and then the representative instances are selected by Algorithm 1. However, in [48], the representative instances are directly selected without eliminating the noise; On the other hand, the evaluation indexes of selecting the representative instances by Algorithm 1 and the method in [48] are different. The former is the importance degree of the fuzzy granule (equation (6)), and the latter is the discriminating ability of the fuzzy granular rule.

IV. Importance-Degree-Preserved Feature Selection Method

Given an original fuzzy decision system $(U, A \cup D)$, a representative instance set U^* can be obtained by Algorithm 1, which can yield a fuzzy decision subsystem $(U^*, A \cup D)$. Generally speaking, redundant features or attributes may be included in data. For $(U^*, A \cup D)$, if selecting the critical feature set A^* to preserve the importance degrees of the fuzzy granules, we can acquire a more compact fuzzy decision system $(U^*, A^* \cup D)$ whose scale may be smaller in both the lengthwise and widthwise directions, and then achieve the objective of data reduction. Thus, in this section, we propose the importance-degree-preserved attribute reduction for $(U^*, A \cup D)$ and design the corresponding feature selection algorithm.

Hereinafter, all the related work orients towards the fuzzy decision subsystem $(U^*, A \cup D)$. Then, the universe U^* is partitioned by the decision equivalence relation R_D into a family of disjoints subsets $U^*/D = \{[x_i]^*_D, x_i \in U^*\}$, where $[x_i]^*_D$ is the decision class of $(U^*, A \cup D)$ to which x_i belongs. It should be pointed out that, for $(U^*, A \cup D)$, the fuzzy lower approximation value of x_i belonging to $[x_i]^*_D$ is denoted by

$$\lambda_i^* = \underline{R_A}[x_i]^*_D(x_i) = \inf_{x_j \in U^*} \max\{1 - R_A(x_i, x_j), [x_i]^*_D(x_j)\}$$

which may be different from

$$\lambda_i = \underline{R_A}[x_i]_D(x_i) = \inf_{x_j \in U} \max\{1 - R_A(x_i, x_j), [x_i]_D(x_j)\}$$

for $(U, A \cup D)$. Besides, for $(U^*, A \cup D)$, the set of all the fuzzy granules induced by the instances in U^* with respect to B is denoted by

$$GrS(U^*, B) = \left\{ [x_i]_B^{\lambda_i^*} : x_i \in U^*, B \subseteq A, \lambda_i^* = \underline{R_A}[x_i]^*_D(x_i) \right\}.$$

Definition 3. *Let $(U^*, A \cup D)$ be a fuzzy decision system with $B \subseteq A$. The importance degree of the fuzzy granule $[x_i]_B^{\lambda_i^*} \in GrS(U^*, B)$ is*

$$F\left([x_i]_B^{\lambda_i^*}\right) = \frac{\left|Cov\left([x_i]_A^{\lambda_i^*}\right)\right|^2}{\left|Cov\left([x_i]_B^{\lambda_i^*}\right)\right|} \cdot \underline{R_B}[x_i]^*_D(x_i) \qquad (8)$$

where

$$\underline{R_B}[x_i]^*_D(x_i) = \inf_{x_j \in U^*} \max\{1 - R_B(x_i, x_j), [x_i]^*_D(x_j)\}.$$

Specially, if $U^* = U$, then it is easily known from Definition 3 that the importance degree of the fuzzy granule $[x_i]_B^{\lambda_i} \in GrS(U, B)$ is

$$F\left([x_i]_B^{\lambda_i}\right) = \frac{\left|Cov\left([x_i]_A^{\lambda_i}\right)\right|^2}{\left|Cov\left([x_i]_B^{\lambda_i}\right)\right|} \cdot \underline{R_B}[x_i]_D(x_i),$$

and then the importance degree of the fuzzy granule $[x_i]_A^{\lambda_i} \in GrS(U, A)$ is

$$\begin{aligned} F\left([x_i]_A^{\lambda_i}\right) &= \frac{\left|Cov\left([x_i]_A^{\lambda_i}\right)\right|^2}{\left|Cov\left([x_i]_A^{\lambda_i}\right)\right|} \cdot \underline{R_A}[x_i]_D(x_i) \\ &= \left|Cov\left([x_i]_A^{\lambda_i}\right)\right| \cdot \lambda_i \end{aligned}$$

which is consistent with equation (6).

Property 1. *Let $(U^*, A \cup D)$ be a fuzzy decision system. $F\left([x_i]_C^{\lambda_i^*}\right) \leqslant F\left([x_i]_B^{\lambda_i^*}\right)$ holds for any $x_i \in U^*$ and $C \subseteq B \subseteq A$.*

Proof: For $(U^*, A \cup D)$ and an arbitrary object $x_i \in U^*$, it is known from Proposition 1 that $\left|Cov\left([x_i]_B^{\lambda_i^*}\right)\right| \leqslant \left|Cov\left([x_i]_C^{\lambda_i^*}\right)\right|$ holds for $C \subseteq B$. Besides, $\underline{R_C}[x_i]^*_D(x_i) = \inf_{x_j \in U^*} \max\{1 - R_C(x_i, x_j), [x_i]^*_D(x_j)\} \leqslant \underline{R_B}[x_i]^*_D(x_i) = \inf_{x_j \in U^*} \max\{1 - R_B(x_i, x_j), [x_i]^*_D(x_j)\}$ holds for any $x_i \in U^*$. Therefore,

$$\begin{aligned} F\left([x_i]_C^{\lambda_i^*}\right) &= \frac{\left|Cov\left([x_i]_A^{\lambda_i^*}\right)\right|^2}{\left|Cov\left([x_i]_C^{\lambda_i^*}\right)\right|} \cdot \underline{R_C}[x_i]^*_D(x_i) \\ &\leqslant \frac{\left|Cov\left([x_i]_A^{\lambda_i^*}\right)\right|^2}{\left|Cov\left([x_i]_B^{\lambda_i^*}\right)\right|} \cdot \underline{R_B}[x_i]^*_D(x_i) = F\left([x_i]_B^{\lambda_i^*}\right) \end{aligned}$$

■

In order to keep the importance degree of each fuzzy granule $[x_i]_A^{\lambda_i^*}$ $(x_i \in U^*)$, both the concepts of an importance-degree-preserved consistent set and the reduct are presented as follows.

Definition 4. *Let $(U^*, A \cup D)$ be a fuzzy decision system. $B \subseteq A$ is called an importance-degree-preserved consistent set of $(U^*, A \cup D)$ if $F\left([x_i]_B^{\lambda_i^*}\right) = F\left([x_i]_A^{\lambda_i^*}\right)$ holds for any $x_i \in U^*$; B is called an importance-degree-preserved reduct of $(U^*, A \cup D)$ if B is an importance-degree-preserved consistent set of $(U^*, A \cup D)$, and there exists $x_{i_0} \in U^*$ such that $F\left([x_{i_0}]_{B-\{a\}}^{\lambda_{i_0}^*}\right) < F\left([x_{i_0}]_B^{\lambda_{i_0}^*}\right)$ holds for any $a \in B$.*

According to Definition 4, an importance-degree-preserved reduct B is factually a minimal subset of A that preserves $F\left([x_i]_B^{\lambda_i^*}\right) = F\left([x_i]_A^{\lambda_i^*}\right)$ for each $x_i \in U^*$.

Definition 5. *Let $(U^*, A \cup D)$ be a fuzzy decision system with $B \subseteq A$. The sum of the importance degrees of all the instances*

in U^* with respect to B is denoted by

$$\Psi_{U^*}(B) = \sum_{x_i \in U^*} F\left([x_i]_B^{\lambda_i^*}\right). \tag{9}$$

It is known from Definition 5 that $\Psi_{U^*}(B)$ characterizes the global information of the importance degrees of all the fuzzy granules $[x_i]_B^{\lambda_i^*}$ $(x_i \in U^*)$.

Property 2. *Let $(U^*, A \cup D)$ be a fuzzy decision system. $\Psi_{U^*}(C) \leqslant \Psi_{U^*}(B)$ holds for $C \subseteq B \subseteq A$.*

Proof: According to Definition 5, it can be implied by Property 1. ∎

Theorem 1. *Let $(U^*, A \cup D)$ be a fuzzy decision system. $B \subseteq A$ is an importance-degree-preserved reduct of $(U^*, A \cup D)$ if and only if $\Psi_{U^*}(B) = \Psi_{U^*}(A)$ and $\Psi_{U^*}(B - \{a\}) < \Psi_{U^*}(B)$ for any $a \in B$.*

Proof: $B \subseteq A$ is an importance-degree-preserved consistent set of $(U^*, A \cup D) \iff F\left([x_i]_B^{\lambda_i^*}\right) = F\left([x_i]_A^{\lambda_i^*}\right)$ for each $x_i \in U^* \iff \sum_{x_i \in U^*} F\left([x_i]_B^{\lambda_i^*}\right) = \sum_{x_i \in U^*} F\left([x_i]_A^{\lambda_i^*}\right)$ $\iff \Psi_{U^*}(B) = \Psi_{U^*}(A)$ according to equation (9).

Furthermore, if there exists $x_{i_0} \in U^*$ such that $F\left([x_{i_0}]_{B-\{a\}}^{\lambda_{i_0}^*}\right) < F\left([x_{i_0}]_B^{\lambda_{i_0}^*}\right)$ holds for any $a \in B \iff \sum_{x_i \in U^*} F\left([x_i]_{B-\{a\}}^{\lambda_i^*}\right) < \sum_{x_i \in U^*} F\left([x_i]_B^{\lambda_i^*}\right)$ for any $a \in B \iff \Psi_{U^*}(B - \{a\}) < \Psi_{U^*}(B)$ for any $a \in B$.

In conclusion, according to Definition 5, B is an importance-degree-preserved reduct if and only if $\Psi_{U^*}(B) = \Psi_{U^*}(A)$ and $\Psi_{U^*}(B - \{a\}) < \Psi_{U^*}(B)$ for any $a \in B$. ∎

Let $(U^*, A \cup D)$ be a fuzzy decision system with $A = \{a_1, a_2, \ldots, a_m\}$, and U^* be a representative instance set. Assume that the attributes $a_{i_1}, a_{i_2}, \ldots$ are added into the empty set one by one according to the magnitude of their respective obtained Ψ_{U^*}. The process continues until there exists some $t \in \{1, 2, \ldots, m\}$ such that $\Psi_{U^*}(\{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}) = \Psi_{U^*}(A)$. It is obtained from Property 2 that $\Psi_{U^*}(\{a_{i_1}\}) \leqslant \Psi_{U^*}(\{a_{i_1}, a_{i_2}\}) \leqslant \ldots \leqslant \Psi_{U^*}(\{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}) = \Psi_{U^*}(A)$. Thus, a forward additional heuristic algorithm of searching for an importance-degree-preserved reduct is provided as follows.

The time complexity of Algorithm 2 is polynomial. Carrying out Steps 1–3 needs $O(|U^*|^2|A|)$, and the complexity of computing $\Psi_{U^*}(A)$ is $O(|U^*|)$. The complexity of computing $\Psi_{U^*}(B \cup \{a_t\})$ is at most $O(|U^*|^2|A|)$. Carrying out Steps 7–9 needs at most $|A|$ times. Totally, the time complexity of Algorithm 2 is at most $O(|U^*|^2|A|^2)$.

It should be noted that the returned result of Algorithm 2 may be a consistent set which may include some redundant features. As indicated in [44], the fewer attributes may possess more powerful generalization ability in processing the data set. In order to obtain a more compact data set, a wrapper technique is provided to find a best feature subset as what follows.

(1) Firstly, let U^* and TeD be the representative instance set and the testing data set, respectively. Assume that the attributes $a_{i_1}, a_{i_2}, \ldots, a_{i_t}$ are stepwise selected from $(U^*, A \cup D)$ by Algorithm 2.

Algorithm 2 Feature selection algorithm of searching for an importance-degree-preserved reduct

Input: A fuzzy decision system $(U^*, A \cup D)$ with $A = \{a_1, a_2, \ldots, a_m\}$ and the decision partition $U^*/D = \{[x_i]_D^*, x_i \in U^*\}$.
Output: An importance-degree-preserved reduct B.

1: **for** each $x_i \in U^*$ **do**
2: compute the fuzzy lower approximation value $\lambda_i^* = \underline{R_A}[x_i]_D^*(x_i)$ and $Cov\left([x_i]_A^{\lambda_i^*}\right)$;
3: **end for**
4: initialize $B = \emptyset$ and $\Psi = 0$;
5: compute $\Psi_{U^*}(A)$ according to equation (9);
6: **while** $\Psi < \Psi_{U^*}(A)$ **do**
7: **for** each $a_t \in A - B$ **do**
8: compute $\Psi_{U^*}(B \cup \{a_t\})$;
9: **end for**
10: choose an attribute a_{t_0} $(a_{t_0} \in A - B)$ satisfying $\Psi_{U^*}(B \cup \{a_{t_0}\}) = \max_{a_t \in A-B} \Psi_{U^*}(B \cup \{a_t\})$ and update $\Psi = \Psi_{U^*}(B \cup \{a_{t_0}\})$;
11: let $B = B \cup \{a_{t_0}\}$;
12: **end while**
13: **return** B.

(2) Secondly, denote $S_1 = \{a_{i_1}\}$, $S_2 = \{a_{i_1}, a_{i_2}\}$, …, and $S_t = \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$ which are considered as the *candidate sequence feature subsets*. For a candidate feature subset $S_k (k \in \{1, 2, \ldots, t\})$, we build some classifier for such a training data set that corresponds to $(U^*, S_k \cup D)$. The classifier is used to classify the testing data TeD with all the attributes in $A - S_k$ being deleted. Then, the classification accuracy achieved by S_k for TeD is obtained and denoted by $acc(S_k)$.

(3) Thirdly, assume that S_{k_0} $(k_0 \in \{1, 2, \ldots, t\})$ achieves the highest accuracy, i.e., $acc(S_{k_0}) = \max_{k \in \{1,2,\ldots,t\}} acc(S_k)$. Then, S_{k_0} is taken as the candidate best feature subset.

(4) Lastly, a backward elimination method is applied on S_{k_0} to select a best feature subset. Concretely, remove the jth $(j = 1, 2, \ldots, |S_{k_0}|)$ feature from S_{k_0} and denote the obtained feature subset as $S_{k_0}^{(j)}$. Then, compute $acc(S_{k_0}^{(1)})$, $acc(S_{k_0}^{(2)})$,…, and $acc(S_{k_0}^{(|S_{k_0}|)})$, respectively. If there exists j_0 $(j_0 \in \{1, 2, \ldots, |S_{k_0}|\})$ such that $S_{k_0}^{(j_0)}$ acquires the highest accuracy, i.e., $acc(S_{k_0}^{(j_0)}) = \max_{j \in \{1,2,\ldots,|S_{k_0}|\}} acc(S_{k_0}^{(j)})$, and the accuracy is not less than $acc(S_{k_0})$, $S_{k_0}^{(j_0)}$ is selected as the candidate best feature subset. The procedure is repeated until no gain in either classification accuracy improvement or feature dimension reduction. Then, the backward elimination technique terminates.

On the basis of the above procedures, Algorithm 3 is formulated to select a best feature subset. It should be noted that, for Algorithm 3, the classifier used in Step 12 is identical to one used in Step 3.

Algorithm 3 Wrapper technique of searching for a best feature subset

Input: The representative instance set U^*, a testing data set TeD, and the feature subset $B = \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$.
Output: A best feature subset B^*.

denote $S_1 = \{a_{i_1}\}$, $S_2 = \{a_{i_1}, a_{i_2}\}$,..., $S_t = \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$;
for $k = 1$ to t **do**
  compute $acc(S_k)$ for TeD by some employed classifier;
end for
if $acc(S_{k_0}) = \max\limits_{k\in\{1,2,\ldots,t\}} acc(S_k)$ **then**
  Let $S = S_{k_0}$ and $acc = acc(S_{k_0})$;
end if
Let $check = 1$;
while $check == 1$ **do**
  for $j = 1$ to $|S|$ **do**
    remove the jth feature from S and denote the obtained feature subset as $S^{(j)}$;
    compute $acc(S^{(j)})$ for TeD by the employed classifier;
  end for
  if $acc(S^{(j_0)}) = \max\limits_{j\in\{1,2,\ldots,|S|\}} acc(S^{(j)}) \geqslant acc$ **then**
    Update $S = S^{(j_0)}$ and $acc = acc(S^{(j_0)})$;
  else
    $check = 0$;
  end if
end while
return $B^* = S$.

V. Bi-Selection Method Based on Fuzzy Rough Sets

The ultimate objective of our work in this paper is to achieve the reduction of both data capacity and dimensionality for the real-valued data. Then, we can obtain a more compact data set with the most representative information. It can be known from the above work that the importance degrees of fuzzy granules can be taken as the evaluation indexes to select the representative instances and the critical features, respectively. Therefore, according to the unified concepts of the importance degrees of the fuzzy granules, we present a bi-selection method based on fuzzy rough sets (BSFRS) by integrating Algorithms 1–3 for data reduction. Specifically, Algorithm 1 is used to acquire the representative instance set U^* of $(U, A \cup D)$. For $(U^*, A \cup D)$, we obtain an importance-degree-preserved reduct B by Algorithm 2. Furthermore, Algorithm 3 is used to search for a best feature set B^* on the testing data set TeD. At last, we obtain the compact data set $ComDS$ with the representative instance set U^* and the conditional attribute subset B^*. Obtaining the compact data set $ComDS$ achieves the reduction of both data capacity and dimensionality. In summary, the whole procedures of BSFRS for data reduction is shown by Algorithm 4.

Algorithm 4 Bi-selection based on fuzzy rough sets (BSFRS)

Input: A fuzzy decision system $(U, A \cup D)$ and a testing data set TeD.
Output: A compact data set $ComDS$.

acquire a representative instance set U^* of $(U, A \cup D)$ by Algorithm 1;
obtain an importance-degree-preserved reduct B for $(U^*, A \cup D)$ by Algorithm 2;
search for a best feature subset B^* on the testing data set TeD by Algorithm 3;
return a compact data set $ComDS$ with the representative instance set U^* and the conditional attribute subset B^*.

VI. Numerical Experiments

In this section, some numerical experiments are conducted to assess the performance of the proposed BSFRS. The experiments mainly focus on the evaluation of BSFRS and the comparison with some instance and feature selection algorithms. In order to achieve these tasks, we downloaded 16 data sets from UCI Repository of machine learning databases [58]. The data sets are briefly described in Table I.

TABLE I
Description of the Data Sets

Data set	Abbreviation of data set	No. of objects	No. of features	No. of classes
1. Wine	Wine	178	13	3
2. Wisconsin Prognostic Breast Cancer	WPBC	194	33	2
3. Wisconsin Diagnostic Breast Cancer	WDBC	569	30	2
4. Ionosphere	Iono	351	34	2
5. Connectionist Bench (Sonar, Mines vs. Rocks)	Sonar	208	60	2
6. Musk (Version 1)	Musk1	476	166	2
7. Hill-Valley	HV	606	100	2
8. Cardiotocography	CTG	2126	20	3
9. Steel Plates Faults	Steel	1941	27	7
10. Waveform Database Generator (Version 1)	WDG1	5000	21	3
11. Waveform Database Generator (Version 2)	WDG2	5000	40	3
12. Wall-Following Robot Navigation Data	Robot	5456	24	4
13. Statlog (Landsat Satellite)	Stat	6435	36	6
14. Anuran Calls (MFCCs)	Anuran	7195	21	4
15. EEG Eye Stste	EEG	14980	14	2
16. Letter Recognition	Letter	20000	16	26

A. Evaluation metrics

Similar to the evaluation metrics recommended for instance selection algorithms in [13], three metrics are used to evaluate the performances of the developed BSFRS and the compared algorithms. The metrics are as follows.

(1) Reduction ratio (Red.) measures the reduction degree of the data size including both the capability and the dimension-

ality, i.e.,

$$\text{Red.} = 1 - \frac{|U^*|}{|U|} \times \frac{|A^*|}{|A|} \tag{10}$$

where U^*, U, A^* and A are the selected instance subset, the original instance set, the selected feature subset and the original feature set, respectively. Here, $|\cdot|$ is the cardinality of a set.

(2) Classification accuracy (Acc.) measures the ability of the classifier built by the compact data, and is the fraction of the number of the correctly classified instances in the testing data set TeD divided by the total number of the instances in TeD, i.e.,

$$\text{Acc.} = \frac{\text{number of the correctly classified instances in } TeD}{\text{total number of the instances in } TeD} \tag{11}$$

(3) Effectiveness (Eff.) represents the ability of the tradeoff of the reduction ratio and the classification accuracy, which is computed by the product of the reduction ratio and the classification accuracy, i.e.,

$$\text{Eff.} = \text{Red.} \times \text{Acc.} \tag{12}$$

B. Design of Experiments

1) Pretreatment of Data Sets: Firstly, the each given data set needs to be normalized. For each data set, the object set, the conditional attribute set and the decision attribute set are denoted by U', A and D, respectively. For each real-valued attribute $a \in A$, the attribute value of each object is normalized as $a'(x_i) = [a(x_i) - \min_j a(x_j)]/[\max_j a(x_j) - \min_j a(x_j)]$, so that $a'(x_i) \in [0,1]$ for each $x_i \in U'$. Here, a is still used to denote the corresponding normalized conditional attribute for notational simplicity. Then, a fuzzy relation for each normalized conditional attribute $a \in A$ is defined as

$$R_{\{a\}}(x_i, x_j) = 1 - |a(x_i) - a(x_j)|. \tag{13}$$

2) Compared Algorithms: In order to assess the effectivity of the proposed noise elimination technique (see Steps 6–10 of Algorithm 1), we consider to remove from BSFRS the noise elimination technique and denote the corresponding bi-selection method by BSFRS1. BSFRS is factually a linear manner to select the instances first and then the features. In order to perform the comparison experiments, we need to employ some instance and feature selection algorithms which should be conducted like our linear manner. Although there exist some fuzzy-rough-set-based instance selection methods [45]–[48], some methods seem to be not suitable for comparison. For example, the basic fuzzy rough instance selection algorithm (FRIS) [45] selects the instances with the fuzzy positive region membership being not lower than a certain threshold. FRIS strongly depends on the fuzzy indiscernibility relation and the threshold. Therefore, we only employ the representative instance selection method (RIS) [48] and some well-known instance selection algorithms CDIS [7], ICF [8], CNN [9] and ENN [10] to select instances. Then, the fuzzy-rough-set-based feature selection algorithms, i.e., the implication-relationship-preserved filter algorithm IRFA [48] and the modified quick reduction algorithm MQRA [59] are utilized to select features, respectively. Since the proposed BSFRS includes the important step of obtaining a best feature subset by using the wrapper technique Algorithm 3, we also add the same wrapper technique to the feature selection algorithms IRFA and MQRA, and denote the algorithms as IRFWA and MQRWA, respectively. At last, we combine the instance and feature selection algorithms to obtain the compared algorithms RIS-IRFWA, CDIS-MQRWA, ICF-MQRWA, CNN-MQRWA and ENN-MQRWA. It should be pointed out that both RIS and IRFA are the algorithms in [48]. In fact, IRFA is a local reduction algorithm oriented towards RIS. Thus, we don't consider CDIS-IRFWA, ICF-IRFWA, CNN-IRFWA and ENN-IRFWA. Moreover, the threshold of the evaluation measure of MQRA was set to be 0.9999.

3) Design of Experiments: The experiments were designed as follows. Given one of the normalized data sets, the ten-fold cross validation approach was used. Specifically, the instances were randomly divided into ten approximately equal parts. One of the ten parts was chosen as a testing data set TeD and the remainder was taken as the training data set TrD. For TrD, the object set was denoted by U, and the conditional attribute set and the decision attribute set were still denoted by A and D, respectively. According to equation (13), the training data set TrD was transformed into a fuzzy decision system $(U, A \cup D)$. It should be noted that the parameter k (the number of the nearest neighbors) existing in the noise elimination technique (see Step 6 in Algorithm 1) may directly effect the number of the selected representative instances. Then, the obtained reduction ratios, the classification accuracies and the effectiveness may be different with the variation of the parameter k. Thus, the parameter k in Algorithm 1 was specified to be 1, 2,..., 10, respectively. For the fuzzy decision system $(U, A \cup D)$ (or the training data set TrD) and the testing data set TeD, each of BSFRS, BSFRS1, and the compared algorithms can all obtain a more compact data set with an instance subset and a best feature subset. Then, the corresponding reduction ratios, the classification accuracies and the effectiveness levels can be acquired. Here, the classification accuracies were the highest classification accuracies of TeD obtained by the wrapper technique Algorithm 3. Moreover, the k-Nearest Neighbor Classifier with k=3, i.e., 3NN, was taken as the classifier in Algorithm 3, and all the other parameters of 3NN were default. This process was repeated for each of the ten parts.

The experiments were performed by Matlab 2020a (64-bit) on a workstation with Intel(R) Core(TM) i7-8750H CPU @2.20GHz 2.21Ghz and 64G memory in Windows 10 system. All the experiments were repeated 10 times. Moreover, the paired t-test was performed to ensure that the experimental results were significantly different, where the significance level was set to be 0.05.

C. Experimental Results

For each of the data sets, the average number of the selected instances, the average number of the selected features, the average reduction ratio, the average classification accuracy, the average effectiveness and the average running time of each

TABLE II
AVERAGE INSTANCE NUMBERS SELECTED BY BSFRS AND THE COMPARED ALGORITHMS

Data set	k	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	6	18.85	22.36	22.66	74.13	89.31	11.98	155.27	5/0/1
WPBC	10	54.42	73.02	73.84	59.27	104.69	72.18	132.72	6/0/0
WDBC	7	46.68	58.05	55.15	188.62	225.09	46.69	497.34	5/1/0
Iono	2	81.12	94.50	98.93	117.97	191.55	49.24	281.15	5/0/1
Sonar	5	47.23	60.89	63.43	83.44	100.03	14.75	157.47	5/0/1
Musk1	2	109.92	141.49	145.01	163.90	220.03	23.10	353.96	5/0/1
HV	2	174.42	382.59	383.25	277.50	417.55	275.37	284.42	6/0/0
CTG	5	183.35	281.17	289.61	920.56	837.50	282.31	1750.63	6/0/0
Steel	1	309.11	721.38	723.91	806.21	1096.31	365.02	1225.46	6/0/0
WDG1	1	966.43	1512.98	1563.71	1282.12	2154.48	1329.60	3609.55	6/0/0
WDG2	1	1316.90	2088.12	2144.01	1019.82	2602.05	1484.44	3502.70	5/0/1
Robot	4	1005.20	1476.66	1480.62	2409.89	2490.16	650.32	4524.69	5/0/1
Stat	1	650.82	1083.58	1117.83	2126.66	3044.37	796.36	5268.45	6/0/0
Anuran	10	174.72	202.8	228.57	3617.45	3416.08	112.38	6410.68	5/0/1
EEG	1	2571.59	3812.67	3902.94	6904.33	6770.78	2290.47	11366.43	5/0/1
Letter	1	3307.84	4729.75	4830.63	7845.70	9579.87	2598.95	17084.12	5/0/1

TABLE III
AVERAGE FEATURE NUMBERS SELECTED BY BSFRS AND THE COMPARED ALGORITHMS

Data set	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	5.02	4.92	4.54	2.99	3.48	3.13	3.42	0/1/5
WPBC	4.49	4.43	5.23	3.99	6.11	6.59	2.81	2/3/1
WDBC	6.11	5.52	6.65	3.93	4.06	6.27	3.91	0/3/3
Iono	4.28	8.12	7.29	4.78	3.33	7.67	2.82	3/1/2
Sonar	10.25	11.13	11.99	4.69	9.73	3.68	7.09	1/2/3
Musk1	20.45	25.96	21.01	11.37	17.17	1.64	12.08	1/1/4
HV	7.03	7.96	8.00	8.49	10.87	5.78	6.45	1/5/0
CTG	7.89	7.67	5.95	6.39	6.34	9.56	6.55	1/1/4
Steel	11.67	10.96	10.53	9.22	10.30	7.60	8.23	0/0/6
WDG1	17.46	16.34	16.34	14.61	17.76	17.73	16.21	0/2/4
WDG2	34.34	34.60	34.61	23.86	33.57	34.36	22.60	0/4/2
Robot	3.35	3.57	3.09	3.18	3.08	2.87	3.34	1/2/3
Stat	26.81	28.30	26.05	20.54	20.23	25.24	19.26	1/1/4
Anuran	12.67	14.15	12.58	10.23	12.12	11.17	10.37	1/2/3
EEG	6.88	7.49	6.98	6.82	7.33	6.88	7.00	2/4/0
Letter	10.40	10.73	10.65	10.60	11.15	10.32	11.82	4/2/0

algorithm are listed in Tables II–VII, respectively. Here, the second column of Table II is the corresponding parameter k that achieves the highest average effectiveness level by BSFRS for each data set. Then, the reported results of BSFRS in Tables III–VII also correspond to the given value k in Table II. Additionally, the statistical comparison results of BSFRS and the other algorithms using the paired t-test are listed in the last column of Tables II–VII, respectively. It should be pointed out that "w" is the number of win achieved by our BSFRS, in which win means that the performance of BSFRS is significantly better than the performances of the other algorithms; "t" is the number of tie achieved by our BSFRS, where tie means that the results obtained by BSFRS have no statistically difference with the other algorithms; similarly, "l" is the number of lose achieved by BSFRS. Moreover, we underline the values which are significantly better than those of BSFRS in Tables II–VII, respectively. For example, in Table IV, the underline of the value 0.9817 means that the reduction ratio of BSFRS for the data set Wine is significantly lower than that of CNN-MQRWA.

D. Discussions of Experimental Results

1) Comparisons between BSFRS and BSFRS1: Since BSFRS1 is factually the version of the proposed BSFRS that does not perform the noise elimination in the instance selection process, it is expected that the noise elimination technique is effective without degrading the performance of BSFRS. The experimental results are as follows.

- Selected instance number: It can be obtained from 3rd and 4th columns in Table II that, for each of the data sets, the instances selected by BSFRS are significantly fewer than those selected by BSFRS1;

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3216990

TABLE IV
AVERAGE REDUCTION RATIOS ACHIEVED BY BSFRS AND THE COMPARED ALGORITHMS

Data set	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	0.9543	0.9473	0.9506	0.8933	0.8505	0.9817	0.7451	4/1/1
WPBC	0.9578	0.9435	0.9330	0.9587	0.8887	0.9171	0.9352	5/1/0
WDBC	0.9816	0.9790	0.9760	0.9516	0.9405	0.9809	0.8735	4/2/0
Iono	0.9657	0.9242	0.9287	0.9442	0.9368	0.9625	0.9215	5/1/0
Sonar	0.9566	0.9395	0.9324	0.9652	0.9134	0.9951	0.9007	4/0/2
Musk1	0.9683	0.9483	0.9571	0.9738	0.9469	0.9995	0.9398	4/0/2
HV	0.9775	0.9442	0.9438	0.9568	0.9169	0.9709	0.9662	5/1/0
CTG	0.9622	0.9436	0.9549	0.8466	0.8614	0.9295	0.7004	6/0/0
Steel	0.9235	0.8324	0.8383	0.8424	0.7607	0.9412	0.7861	5/0/1
WDG1	0.8215	0.7384	0.7296	0.8017	0.5951	0.7506	0.3809	6/0/0
WDG2	0.7487	0.5987	0.5876	0.8648	0.5149	0.7166	0.5605	5/0/1
Robot	0.9714	0.9553	0.9612	0.9350	0.9349	0.9842	0.8718	5/0/1
Stat	0.9163	0.8529	0.8604	0.7905	0.7047	0.9036	0.5134	6/0/0
Anuran	0.9837	0.9789	0.9789	0.7277	0.6957	0.9908	0.5113	5/0/1
EEG	0.9062	0.8487	0.8556	0.7504	0.7369	0.9165	0.5784	5/0/1
Letter	0.8805	0.8239	0.8214	0.7112	0.6291	0.9068	0.2985	5/0/1

TABLE V
AVERAGE CLASSIFICATION ACCURACIES ACHIEVED BY BSFRS AND THE COMPARED ALGORITHMS

Data set	Original	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	0.9585	0.9460	0.9332	0.8998	0.9704	0.9675	0.7274	0.9799	3/0/3
WPBC	0.7395	0.8681	0.7967	0.7928	0.8447	0.8259	0.8697	0.8362	5/1/0
WDBC	0.9692	0.9629	0.9213	0.9334	0.9789	0.9631	0.9722	0.9828	2/1/3
Iono	0.8464	0.9089	0.8976	0.9316	0.9219	0.9259	0.9445	0.9405	0/1/5
Sonar	0.8342	0.8951	0.8858	0.9102	0.8475	0.9334	0.7204	0.9261	2/2/2
Musk1	0.8358	0.8561	0.9177	0.9035	0.8201	0.8935	0.5913	0.8879	2/0/4
HV	0.5323	0.5890	0.6245	0.6088	0.5872	0.6401	0.6096	0.5772	1/1/4
CTG	0.9109	0.9026	0.8440	0.8548	0.9055	0.8676	0.8953	0.9227	4/1/1
Steel	0.7094	0.6918	0.6307	0.6136	0.7114	0.6954	0.5014	0.7372	3/1/2
WDG1	0.8058	0.8119	0.7642	0.7746	0.8365	0.8123	0.7945	0.8440	3/1/2
WDG2	0.7763	0.8108	0.7861	0.7864	0.8180	0.8114	0.7816	0.8244	3/1/2
Robot	0.8731	0.8931	0.9007	0.8669	0.8762	0.8851	0.8204	0.9199	4/0/2
Stat	0.9095	0.8974	0.8459	0.8543	0.8740	0.8963	0.8916	0.9148	4/1/1
Anuran	0.9912	0.9580	0.9411	0.9525	0.9549	0.9872	0.8580	0.9931	3/1/2
EEG	0.8351	0.7586	0.7363	0.7406	0.7307	0.7849	0.6887	0.8253	4/0/2
Letter	0.9555	0.9174	0.9336	0.9378	0.8882	0.9460	0.9232	0.9548	1/0/5

- Selected feature number: According to Table III, we know that the features selected by BSFRS are significantly fewer than or nearly equivalent to those selected by BSFRS1 for almost all of the data sets except Steel and WDG1;
- Reduction ratio: As a result, it can be seen from Table IV that the average reduction ratios of BSFRS are significantly higher than those of BSFRS1 for almost all of the data sets except WDBC. Thus, the noise elimination technique in Algorithm 1 does make BSFRS select fewer representative instances and have the obvious advantage in terms of reduction ratio;
- Classification accuracy: It can be obtained from 3rd and 4th columns in Table V that the average classification accuracies achieved by BSFRS are significantly higher than those achieved by BSFRS1 for the most of the data sets, and significantly lower than those achieved by BSFRS1 for the data sets Musk1, HV, Robot and Letter;
- Effectiveness: According to 2nd and 3rd columns in Table VI, the average levels of effectiveness obtained by BSFRS are significantly higher than those obtained by BSFRS1 for almost all of the data sets, and significantly lower for the only two data sets Musk1 and HV. It should be pointed out that both Musk1 and HV are high-dimensional data sets. Generally speaking, the data points are more and more sparse with the increase of dimensionality. That's to say, the instances at close distance or the similar instances are fewer in high-dimensional data. Given a fuzzy granule (see equation (3)) generated by the high-dimensional data, it is likely to cover the only one instance which induces the fuzzy granule itself. Then, the instance inducing the fuzzy granule may be, to a great extent, recognized as the noise according to the noise elimination technique in Algorithm 1. Thus, it

TABLE VI
AVERAGE EFFECTIVENESS ACHIEVED BY BSFRS AND THE COMPARED ALGORITHMS

Data set	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	0.9022	0.8835	0.8537	0.8666	0.8216	0.7134	0.7291	6/0/0
WPBC	0.8306	0.7496	0.7375	0.8089	0.7308	0.7957	0.7816	6/0/0
WDBC	0.9451	0.9016	0.9106	0.9314	0.9055	0.9536	0.8583	5/0/1
Iono	0.8774	0.8284	0.8646	0.8700	0.8670	0.9089	0.8662	4/1/1
Sonar	0.8550	0.8307	0.8471	0.8169	0.8510	0.7167	0.8325	4/2/0
Musk1	0.8281	0.8695	0.8637	0.7979	0.8447	0.5910	0.8332	2/1/3
HV	0.5753	0.5883	0.5739	0.5611	0.5848	0.5915	0.5574	2/2/2
CTG	0.8684	0.7957	0.8160	0.7661	0.7469	0.8319	0.6456	6/0/0
Steel	0.6387	0.5235	0.5134	0.5991	0.5281	0.4717	0.5792	6/0/0
WDG1	0.6670	0.5645	0.5652	0.6704	0.4832	0.5963	0.3204	5/1/0
WDG2	0.6071	0.4710	0.4620	0.7067	0.4174	0.5602	0.4573	5/0/1
Robot	0.8675	0.8603	0.8332	0.8192	0.8275	0.8074	0.8020	6/0/0
Stat	0.8223	0.7215	0.7348	0.6903	0.6314	0.8056	0.4688	6/0/0
Anuran	0.9424	0.9212	0.9323	0.6949	0.6868	0.8501	0.5077	6/0/0
EEG	0.6875	0.6249	0.6337	0.5483	0.5784	0.6312	0.4774	6/0/0
Letter	0.8078	0.7692	0.7703	0.6316	0.5952	0.8372	0.2852	5/0/1

TABLE VII
AVERAGE RUNNING TIME (S) OF BSFRS AND THE COMPARED ALGORITHMS

Data set	BSFRS	BSFRS1	RIS-IRFWA	CDIS-MQRWA	ICF-MQRWA	CNN-MQRWA	ENN-MQRWA	Paired t-test(w/t/l)
Wine	0.35	0.47	0.37	0.27	0.30	0.15	0.26	1/1/4
WPBC	0.61	1.40	0.57	0.57	0.63	0.62	0.64	1/5/0
WDBC	1.05	1.26	1.16	0.83	1.00	0.61	2.33	2/2/2
Iono	0.67	2.05	0.99	0.97	0.74	0.71	1.10	5/1/0
Sonar	2.15	5.18	2.77	1.17	2.74	0.39	2.68	4/0/2
Musk1	12.80	90.36	17.85	7.35	16.05	0.42	24.25	4/0/2
HV	5.73	134.46	14.66	5.93	11.54	6.38	6.58	5/1/0
CTG	8.83	10.75	18.10	2.86	10.04	0.84	8.07	3/0/3
Steel	2.91	28.43	18.62	3.30	6.54	1.18	6.55	5/0/1
WDG1	31.54	61.82	240.86	5.26	35.90	6.21	42.69	4/0/2
WDG2	48.72	300.41	300.88	10.50	75.68	24.71	133.01	4/0/2
Robot	44.58	91.38	311.90	19.14	56.53	2.31	94.41	4/0/2
Stat	29.83	113.32	497.77	28.14	74.25	7.30	195.44	4/0/2
Anuran	163.76	165.76	632.29	34.41	191.48	0.82	94.29	3/0/3
EEG	1584.59	1679.59	5993.40	47.89	1563.49	5.93	135.49	2/0/4
Letter	84.20	288.38	13183.79	95.75	171.02	11.59	404.09	5/0/1

is suggested that the noise elimination technique should not be used for the high-dimensional data sets, and then BSFRS1 is more appropriate for the high-dimensional data sets;

- Running time: We know from Table VII that the average running time of BSFRS is less than that of BSFRS1 for each data set. Since both BSFRS and BSFRS1 include the various steps, i.e., the instance selection, feature selection and wrapper technique steps which correspond to Algorithms 1–3, respectively. we depict the running time histogram in Fig. 1 where "1–3" in the horizontal axis means Algorithms 1–3, respectively. It is seen from Fig. 1 that the average running time of the feature selection procedure of BSFRS1 is always more than that of BSFRS, and occupies the most of all the running time for nearly all of the data sets except CTG, Anuran and EGG. In addition, for the data sets CTG, Anuran and EGG, although the instance selection procedure takes the main part of all the running time, the feature selection procedure of BSFRS1 still needs more time than BSFRS. The possible cause may be the fact that BSFRS1 can obtain more representative instances without removing the noise, whereas the feature selection procedure is oriented towards the selected representative instances for preserving the importance degrees of these instances. Therefore, BSFRS1 costs more computation time than BSFRS.

In conclusion, except the high-dimensional data, the proposed noise elimination technique in Algorithm 1 is effective and competitive, which may make BSFRS achieve significantly higher reduction ratio, classification accuracy and effectiveness with less running time.

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3216990

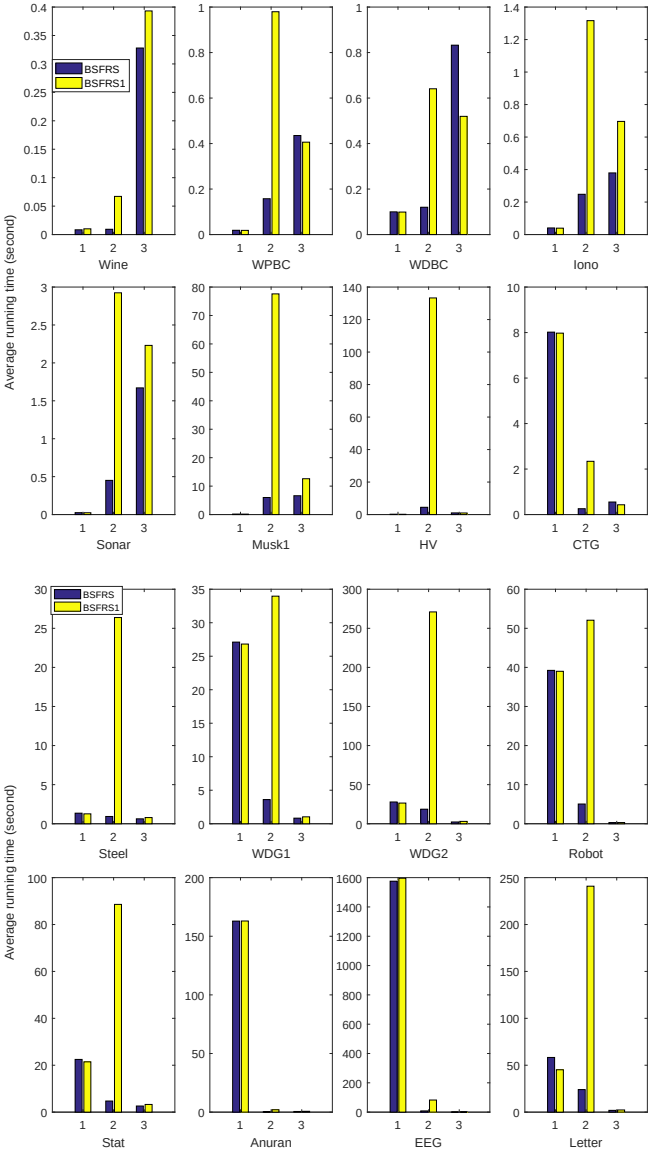

Fig. 1. Average running time (s) of Algorithms 1–3 in BSFRS and BSFRS1

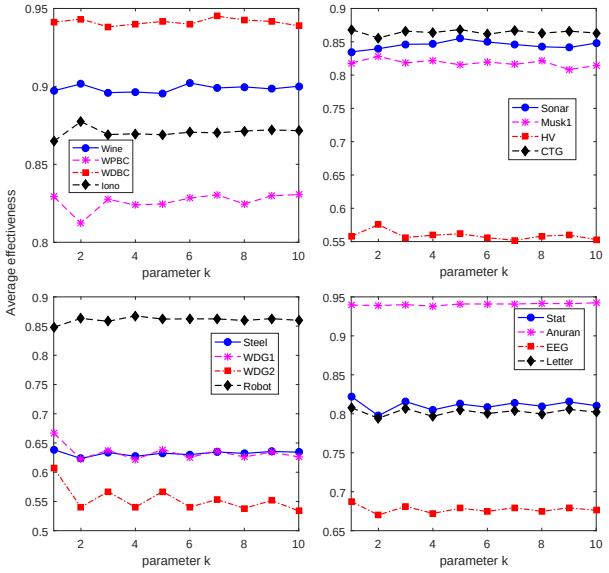

Fig. 2. Average effectiveness of BSFRS with the parameter k being 1 to 10

2) Comparisons With the Other Algorithms: The comparison results are as follows.

- Selected instance number: It can be seen from Table II that, for almost all of the data sets, the instances selected by BSFRS are significantly fewer than those selected by RIS-IRFWA, CDIS-MQRWA, ICF-MQRWA and ENN-MQRWA. In addition, for the half of the data sets (see the underlines in 8th column), the instances selected by BSFRS are significantly more than those acquired by CNN-MQRWA;
- Selected feature number: It is known from Table III that, for nearly half of the data sets, the average numbers of the features selected by BSFRS are significantly more than those obtained by ICF-MQRWA and CNN-MQRWA. BSFRS obtains significantly more features than CDIS-MQRWA and ENN-MQRWA for the most of the data sets. Moreover, BSFRS obtains significantly more features than RIS-IRFWA for the five data sets Wine, CTG, Steel, WDG1 and Robot. Since the wrapper technique Algorithm 3 is added to all the other compared algorithms, a best feature subset can be obtained by each of the compared algorithms. Thus, the number of the features obtained by BSFRS is comparable with or larger than that acquired by the compared algorithms depending on the specific data sets;
- Reduction ratio: The conclusions of reduction ratios obtained from Table IV are nearly similar to those obtained from Table II;
- Classification accuracy: We know from Table V that the average classification accuracies achieved by BSFRS are significantly lower than those of RIS-IRFWA, CDIS-MQRWA, ICF-MQRWA and CNN-MQRWA for a few of the data sets. Although CNN-MQRWA can obtain the significantly higher average reduction ratios for the half of the data sets, it achieves significantly higher average classification accuracies for only the four data sets WDBC, Iono, HV and Letter. Besides, ENN-MQRWA acquires significantly lower average reduction ratios for almost all of the data sets, but it achieves significantly higher average classification accuracies for 14 data sets. It should be pointed out that a competitive data reduction algorithm should consider the tradeoff between the reduction ratio and the classification accuracy;
- Effectiveness: It can be obtained from Table VI that, for almost all of the data sets, the average levels of effectiveness of BSFRS are significantly higher than those of RIS-IRFWA, CDIS-MQRWA, ICF-MQRWA and ENN-MQRWA. Additionally, BSFRS obtains significantly higher effectiveness than CNN-MQRWA for the most of the data sets. Thus, the proposed BSFRS is of competition and achieves the satisfactory effectiveness;

- Running time: It can be known from Table VII that the average running time of BSFRS is significantly more than that of CDIS-MQRWA and CNN-MQRWA for nearly all of the data sets, and is significantly more than that of ICF-MQRWA and ENN-MQRWA for 2 and 4 data sets, respectively. Moreover, the running time of BSFRS is significantly less than that of RIS-IRFWA for almost all of the data sets.

In summary, although BSFRS needs more running time than some of the compared algorithms, it achieves the satisfactory effectiveness for the most of the data sets.

3) The choice of Parameter k: We have already listed the corresponding parameter k which achieves the highest average effectiveness in 2nd column of Table II. In order to clearly illustrate the impact of the parameter k on the effectiveness, we depict in Fig. 2 the average effectiveness of each data set obtained by BSFRS with $k = 1, 2, \ldots, 10$. It's seen from Fig. 2 that, for the most of the data sets, the average effectiveness level is generally steady with the variation of k. In addition, it's known from 2nd column of Table II that BSFRS with $k = 1$ obtains the highest average effectiveness for the six data sets Steel, WDG1, WDG2, Stat, EEG and Letter, and BSFRS with $k = 2$ achieves the highest average effectiveness for the three data sets Iono, Musk1 and HV. For the six data sets Wine, WPBC, WDBC, Sonar, CTG and Robot, we find that the highest average effectiveness obtained by BSFRS with the values k listed in 2nd column of Table II has no significant difference with that obtained by BSFRS with $k = 1, 2$ or 3. For example, for CTG, the highest average effectiveness acquired by BSFRS with $k = 5$ has no significant difference with that obtained by BSFRS with $k = 1$ or 3. Moreover, for the remaining one data set Anuran, the average effectiveness levels obtained by BSFRS with $k = 1, 2, 3$ are 0.9394, 0.9390 and 0.9400, respectively. Although the highest average effectiveness 0.9424 acquired by BSFRS with $k = 10$ is significantly higher than that obtained by BSFRS with $k = 1, 2$ and 3, there doesn't exist too much bias between 0.9424 and the average effectiveness corresponded to $k = 1, 2$ or 3. Therefore, it is suggested that a good parameter k could be chosen among 1, 2 and 3.

VII. Conclusion

In this paper, we study the bi-selection of instance and feature using fuzzy rough sets for data reduction, and present a bi-selection method BSFRS. The highlights of this work are as follows.

- The unified concepts of the importance degrees of fuzzy granules are presented to select the representative instances first and then the critical features;
- An instance selection algorithm (Algorithm 1) with the noise elimination technique is provided to firstly delete the noise and then select the representative instances according to the importance degrees of the fuzzy granules. The noise elimination technique in Algorithm 1 has been experimentally validated to make BSFRS achieve significantly better performance (except the high-dimensional data) with significantly less computation time;
- The importance-degree-preserved attribute reduction is proposed, and Algorithm 2 is formulated to compute an importance-degree-preserved reduct. Furthermore, a wrapper technique (Algorithm 3) is provided to search for a best feature subset;
- Algorithms 1–3 are integrated to form a bi-selection method BSFRS which can reduce both the capacity and the dimensionality of the real-valued data.

The performance of BSFRS is assessed by conducting the extensive numerical experiments. The experimental results show that BSFRS can achieve significantly higher effectiveness for the most of the data sets. It should be pointed out that BSFRS without the noise elimination technique is suggested to deal with the high-dimensional data for data reduction, which may obtain higher classification accuracy and effectiveness.

In the future, we will investigate bi-selection methods for complex data and incomplete data. Moreover, the incremental bi-selection mechanisms under the dynamic environment will be paid more attention.

References

[1] D. Pyle, *Data preparation for data mining*. San Diego, CA, USA: Morgan Kaufmann, 1999.

[2] H. Liu and H. Motoda, "On issues of instance selection," *Data Mining Knowl. Discov.*, vol. 6, no. 2, pp. 115–130, 2002.

[3] S. García, J. Derrac, J.R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 417–435, 2012.

[4] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. New York, USA: Kluwer Academic Publishers, 1998.

[5] I. Triguero, J. Derrac, S. García, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 42, no. 1, pp. 86–100, 2011.

[6] H. Liu, F. Hussain, C.L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining Knowl. Discov.*, vol. 6, no. 4, pp. 393–423, 2002.

[7] J.L. Carbonera and M. Abel, "A novel density-based approach for instance selection," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell.*, 2016, pp. 549–556.

[8] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Min. Knowl. Disc.*, vol. 6, no. 2, pp. 153–172, 2002.

[9] P.E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 515–516, 1968.

[10] D.L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. 2, no. 3, pp. 408–421, 1972.

[11] P.S. Bradley, U. Fayyad, C. Reina, "Scaling clustering algorithms to large databases," in *Proc. 4th Int. Conf. Knowl. Disc. Data Min.*, 1998, pp. 9–15 .

[12] S. Guha, R. Rastogi, K. Shim, "CURE: An efficient clustering algorithm for large databases," in *Proc. 1998 ACM-SIGMOD Int. Conf. on Management of Data*, 1998, pp. 73–84.

[13] M. Malhat, M.E. Menshawy, H. Mousa H, and A.E. Sisi, "A new approach for instance selection: Algorithms, evaluation, and comparisons," *Expert Syst. Appl.*, vol. 149, pp. 113297, 2020.

[14] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, pp. 131–156, 1997.

[15] R. Kohavi and G.H.John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.

[16] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, 2002.

[17] H.C. Peng, F.H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[18] C.F. Tsai, W. Eberle, and C.Y. Chu, "Genetic algorithms in feature and instance selection," *Knowl.-Based Syst.*, vol. 39, pp. 240–247, 2013.

[19] J. Derrac, C. Cornelis, S. García, and F. Herrera, "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection," *Inf. Sci.*, vol. 186, pp. 73–92, 2012.

[20] J. Derrac, S. García, and F. Herrera, "IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule," *Pattern Recognit.*, vol. 43, pp. 2082–2105, 2010.

[21] F. Ros, S. Guillaume, M. Pintore, and J.R. Chrétien, "Hybrid genetic algorithm for dual selection," *Pattern Anal. Applic.*, vol. 11, no. 2, pp. 179–198, 2008.

[22] L.I. Kuncheva and L.C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," *Pattern recognit. lett.*, vol. 20, pp. 1149–1156, 1999.

[23] I.M.R. Albuquerque, B.H. Nguyen, B. Xue, and M.J. Zhang, "A novel genetic algorithm approach to simultaneous feature selection and instance selection," in *Proc. IEEE Symposium Series Comput. Intell.*, 2020, pp. 616–623.

[24] D. Fragoudis, D. Meretakis, S. Likothanassis, "Integrating feature and instance selection for text classification," in *Proc. 8th Int. Conf. Knowl. Disc. Data Min.*, 2002, pp. 501–506.

[25] C.C. Lin, J.R. Kang, Y.L. Liang, and C.C. Kuo, "Simultaneous feature and instance selection in big noisy data using memetic variable neighborhood search," *Appl. Soft Comput.*, vol. 112, pp. 107855, 2021.

[26] Z. Pawlak, "Rough sets," *Int. J. Comput. & Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.

[27] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. General Syst.*, vol. 17, no. 2-3, pp. 191–209, 1990.

[28] A.M. Radzikowska and E.E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 126, no. 2, pp. 137–155, 2002.

[29] J.H. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 211–221, 2013.

[30] Q.H. Hu, D.R. Yu, W. Pedrycz, and D.G. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, 2011.

[31] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets Syst.*, vol. 141, pp. 469–485, 2004.

[32] S. Vluymans, L. D'eer, Y. Saeys, and C. Cornelis, "Applications of fuzzy rough set theory in machine learning: A survey," *Fund. Inform.*, vol. 142, no. 1-4, pp. 53–86, 2015.

[33] F.F. Xu, D.Q. Miao, and L. Wei, "Fuzzy-rough attribute reduction via mutual information with an application to cancer classification," *Comput. Math. Appl.*, vol. 57, no. 6, pp. 1010–1017, 2009.

[34] J.K. Chen, J.S. Mi, and Y.J. Lin, "A graph approach for fuzzy-rough feature selection," *Fuzzy Sets Syst.*, 2020, 391: 96-116.

[35] B.B. Sang, H.M. Chen, L. Yang, T.R. Li, and W.H. Xu, "Incremental feature selection using a conditional entropy based on fuzzy dominance neighborhood rough sets," *IEEE Trans. Fuzzy Syst.*, to be published. DOI:10.1109/TFUZZ.2021.3064686

[36] Q.H. Hu, L.J. Zhang, Y.C. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, 2018.

[37] Y.J. Lin, Y.W. Li, C.X. Wang, and J.K. Chen, "Attribute reduction for multi-label learning with fuzzy rough set," *Knowl-Based Syst.*, vol. 152, pp. 51–61, 2018.

[38] Y.Y. Yang, D.G. Chen, H. Wang, and X.Z. Wang, "Incremental perspective for feature selection based on fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1257–1273, 2018.

[39] E.C.C. Tsang, D.G. Chen, D.S. Yeung, X.Z. Wang, and J.W.T. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, 2008.

[40] J.H. Dai, H. Hu, W.Z. Wu, Y.H. Qian, and D.B. Huang, "Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 2174–2187, 2018.

[41] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 4, pp. 824–837, 2009.

[42] C.Z. Wang, Y.L. Qi, M.W. Shao *et al.*, "A fitting model for feature selection with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 741–753, 2017.

[43] Q.H. Hu, D.R. Yu, Z.X. Xie, and J.F. Liu, "Fuzzy probabilistic approximations spaces and their information measures," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 2, pp. 191–201, 2006.

[44] X. Zhang, C.L. Mei, D.G. Chen, and J.H. Li, "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy," *Pattern Recognit.*, vol. 56, pp. 1–15, 2016.

[45] R. Jensen and C. Cornelis, "Fuzzy-rough instance selection," In *Proc. WCCI 2010 IEEE World Congress Comput. Intell.*, 2010, pp.1–7.

[46] N. Verbiest, C. Cornelis, and F. Herrera, "FRPS: A fuzzy rough prototype selection method," *Pattern Recognit.*, vol. 46, pp. 2770–2782, 2013.

[47] E.C.C. Tsang, Q.H. Hu, and D.G. Chen, "Feature and instance reduction for PNN classifiers based on fuzzy rough sets," *Int. J. Mach. Learn. Cybern.*, vol. 7, pp. 1–11, 2016.

[48] X. Zhang, C.L. Mei, D.G. Chen, and Y.Y. Yang, "A fuzzy rough set-based feature selection method using representative instances," *Knowl-Based Syst.*, vol. 151, pp. 216–229, 2018.

[49] X. Zhang, C.L. Mei, D.G. Chen, Y.Y. Yang, and J.H. Li, "Active incremental feature selection using a fuzzy-rough-set-based information entropy," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 901–915, 2019.

[50] J.R. Anaraki, S. Samet, J.H. Lee, and C.W. Ahn, "SUFFUSE: Simultaneous fuzzy-rough feature-sample selection," *J. Adv. Inf. Technol.*, vol. 6, no. 3, pp. 103–110, 2015.

[51] D. Ślęzak and A, Janusz, "Ensembles of bireducts: Towards robust classification and simple representation," In *Proc. Int. Conf. Future Generation Inf. Technol.*, 2011, pp. 64–77.

[52] N.M. Parthaláin and R. Jensen, "Simultaneous feature and instance selection using fuzzy-rough bireducts," In *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2013, pp. 1–8.

[53] N.M. Parthalain, R. Jensen, and R. Diao, "Fuzzy-rough set bireducts for data reduction," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 8, pp. 1840–1850, 2020.

[54] F. Tao, L. Zhang, and Y. Laili, *Configurable intelligent optimization algorithm*. Heidelberg, Germany: Springer, 2015.

[55] J. S. Mi and W. X. Zhang, An axiomatic characterization of a fuzzy generalization of rough sets, *Inf. Sci.*, vol. 160, pp. 235–249, 2004.

[56] N. N. Morsi and M. M. Yakout, Axiomatics for fuzzy rough sets, *Fuzzy Sets Syst.*, vol. 100, pp. 327–342, 1998.

[57] W.Z. Wu, J.S. Mi, and W.X. Zhang, "Generalized fuzzy rough sets," *Inf. Sci.* vol. 151, pp. 263–282, 2003.

[58] 2007. [Online]. Available: http://www.ics.uci.edu/ mlearn/ML Repository.html

[59] C. Cornelis, R. Jensen, G. Hurtado, and D. Slezak, "Attribute selection with fuzzy decision reducts," *Inf. Sci.*, vol. 180, pp. 209–224, 2010.

Xiao Zhang received the B.Sc. degree in science from Henan Normal University, Xinxiang, China, in 2008, the M.Sc. degree in science from North China Electric Power University, Beijing, China, in 2011, and the Ph.D. degree in science from Xi'an Jiaotong University, Xi'an, China, in 2014.

She is currently with the School of Science, Xi'an University of Technology, Xi'an. Her current research interests include rough sets, granular computing, data mining, and machine learning.

Changlin Mei received the B.Sc., M.Sc., and Ph.D. degrees in science from Xi'an Jiaotong University, Xi'an, China, in 1983, 1989, and 2000, respectively.

He was a Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University, before July, 2019, and is currently a Professor with the School of Science, Xi'an Polytechnic University, Xi'an. He has published more than 70 papers in international journals. His current research interests include data mining, spatial data analysis, and nonparametric regression analysis.

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3216990

Jinhai Li received the M.Sc. degree in science from Guangxi University, Nanning, China, in 2009, and the Ph.D. degree in science from Xi'an Jiaotong University, Xi'an, China, in 2012.

He is currently a Professor with the Kunming University of Science and Technology, Kunming, China. His current research interests include big data, cognitive computing, granular computing, and formal concept analysis.

Yanyan Yang received the M.Sc. degree in science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree in engineering from North China Electric Power University, Beijing, China, in 2017.

From 2015 to 2016, she visited the University of Ulster as a Joint Ph.D. Student sponsored by China Scholarship Council. From 2017 to 2019, she was a Postdoctoral Fellow with Tsinghua University, Beijing. She is currently with the School of Software Engineering, Beijing Jiaotong University, Beijing, China. Her current research interests include rough sets, fuzzy sets, and machine learning.

Ting Qian received the B.Sc. degree in science from Shangqiu Normal University, Shangqiu, China, in 2009, the M.Sc. degree and the Ph.D. degrees in science from Northwest University, Xi'an, China, in 2012 and 2016, respectively.

She is currently with the School of Science, Xi'an Shiyou University, Xi'an. Her current research interests include rough sets, granular computing, formal concept analysis and three-way concept analysis.