entropy

MDPI

Article

Ensemble Fuzzy Feature Selection Based on Relevancy, Redundancy, and Dependency Criteria

Omar A. M. Salem [1,2], Feng Liu [1,*], Yi-Ping Phoebe Chen [3] and Xi Chen [1,*]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China; omarsalem@whu.edu.cn or omarsalem@ci.suez.edu.eg

[2] Department of Information System, Faculty of Computers and Informatics, Suez Canal University, Ismailia 41522, Egypt

[3] Department of Computer Science and Information Technology, La Trobe University, Melbourne 3086, Australia; phoebe.chen@latrobe.edu.au

* Correspondence: fliuwhu@whu.edu.cn (F.L.); robertcx@whu.edu.cn (X.C.)

Received: 7 May 2020; Accepted: 28 June 2020; Published: 9 July 2020

Abstract: The main challenge of classification systems is the processing of undesirable data. Filter-based feature selection is an effective solution to improve the performance of classification systems by selecting the significant features and discarding the undesirable ones. The success of this solution depends on the extracted information from data characteristics. For this reason, many research theories have been introduced to extract different feature relations. Unfortunately, traditional feature selection methods estimate the feature significance based on either individually or dependency discriminative ability. This paper introduces a new ensemble feature selection, called fuzzy feature selection based on relevancy, redundancy, and dependency (FFS-RRD). The proposed method considers both individually and dependency discriminative ability to extract all possible feature relations. To evaluate the proposed method, experimental comparisons are conducted with eight state-of-the-art and conventional feature selection methods. Based on 13 benchmark datasets, the experimental results over four well-known classifiers show the outperformance of our proposed method in terms of classification performance and stability.

Keywords: feature selection; fuzzy sets; mutual information; rough set

1. Introduction

Nowadays, classification systems have a lot of contributions in different domains such as bioinformatics, medical analysis, text categorization, pattern recognition, and intrusion detection [1]. The main challenge of these systems is to deal with high dimensionality data, which may include redundant or irrelevant features [2]. These features have a negative effect on classification systems which can lead to (1) reducing the classification accuracy, (2) reducing the classification speed, (3) increasing the classification complexity. To overcome these limitations, features selection introduces an effective solution to reduce the dimensionality of data by selecting the significant features and discarding the undesirable ones [3].

Feature selection methods are divided into three categories: filter [4], embedded [5], and wrapper [6]. These methods can be also classified into two groups according to the role of classifiers in the feature selection process: classification-independent (filter method), and classification dependent group (embedded, and wrapper method) [3]. The former depends only on the data characteristics without considering classifiers in the selection process, while the latter depends on classifiers to assess the significance of features in the selection process. Although the classification-dependent group can return the best feature selection subset, it requires more computational cost as a result of the

Entropy **2020**, *22*, 757; doi:10.3390/e22070757 www.mdpi.com/journal/entropy

classification process. Moreover, the selected features related only to the used classifier in the feature selection process. For this reason, classification-independent is more practical for high dimensionality data [7]. In this study, filter feature selection is our interest rather than embedded and wrapper due to its benefits such as simplicity, practicality, scalability, efficiency, and generality [8].

The success of filter methods depends on the amount of extracted information from data characteristics [9]. Motivated by this hypothesis, many theories have been introduced to find the best filter feature selection method such as information theory [10], and rough set theory [11]. Information theory measures can rank the features not only according to their relevancy to class but also with respect to the redundancy of features [12]. Moreover, These measures outperform other measures as correlation due to its ability to deal with linear and non-linear relations [3]. Rough set theory can select a subset of features according to their dependency to class [13]. The main advantages of rough set measures are simplicity, and no user-defined parameter is required. However, the traditional measures of these theories share common limitation, they can not deal directly with continuous features. To overcome this limitation, many research studies have been extended by integrating the previous theories with fuzzy set theory [14–16]. Feature selection based fuzzy sets is not only suitable for any kind of data but also extracts more information from classes compared with the traditional feature selection methods [14]. In addition to its ability to deal with noise data [17].

Traditional methods based on previous theories estimate the feature significance based on either individually or dependency discriminative ability. Consequently, there is no general feature selection method, which returns the best feature subset with all datasets [18]. The traditional solution is to understand the data characteristics before the feature selection process. This solution is not efficient because of the high computational cost of expert analysis. To overcome this limitation, a new research direction, called an ensemble feature selection, is introduced, which combines more than one feature selection to cover all situations [2].

In this study, we propose a new ensemble feature selection method (fuzzy feature selection based on relevancy, redundancy, and dependency (FFS-RRD)) to utilize the previous theories. Firstly, we proposed a new method, called fuzzy weighted relevancy-based FS (FWRFS) to estimate the individually discriminative ability. Then, we combined it with fuzzy lower approximation-based FS (L-FRFS) to estimate the dependency discriminative ability [16]. The former method extracts two relations: relevancy and redundancy, while the latter extracts the dependency relation. The aim is to investigate these relations and produce a unique and effective feature selection method to improve classification methods.

The paper is organized as follows: Section 2 presents the main criteria of feature selection: relevancy, redundancy, and dependency. Then, the related work is presented in Section 3. Section 4 introduces the proposed method: fuzzy feature selection based on relevancy, redundancy, and dependency (FFS-RRD). After that, the experiment setup is showed in Section 5. Section 6 analyzes the experimental results. Finally, the conclusion is reported in Section 7.

2. Relevancy, Redundancy, and Dependency Measures

Filter-based FS methods try to find the best feature subset based on data characteristics without depending on classification models [4]. For this reason, they depend on the characteristics of data to find the most significant features. Consequently, filter-based feature selection methods study different data relations such as the relation between features and class, and the relation among features. There are three well-known feature relations: relevancy, redundancy, and dependency.

Firstly, relevancy relation measures the amount of shared information between features and the class [15]. However, some features may have the same relevancy relation and do not add new information to discriminate the classes. These features are considered redundant and no need to be selected. Redundancy relation measures the amount of shared information among features [15]. Another important feature relation is dependency [16]. Dependency relation measures the membership degree of feature subset to class. In the following, we present the definitions of these relations based on the fuzzy set theory [15,16].

Given a dataset $D = (U, F \cup C)$, where $U = \{u_1, u_2, \ldots, u_m\}$ is a finite set of m instances, $F = \{f_1, f_2, \ldots, f_n\}$ is a finite set of n features, and $C = \{c_1, c_2, \ldots, c_l\}$ is a finite set of l classes. Let $f : U \to V_f$, where V_f is the feature value on U. Every feature $f \in F$ can be represented by fuzzy equivalence relation E_f on U and defined by the following fuzzy relation matrix $M(E_f)$.

$$M(E_f) = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & \cdots & \cdots & e_{2m} \\ & \cdots\cdots\cdots\cdots & & \\ e_{m1} & e_{m2} & \cdots & e_{mm} \end{pmatrix} \tag{1}$$

where $e_{ij} = E(x_i, x_j)$ is the fuzzy equivalence relation that defines the similarity degree between x_i and x_j, where $x_i, x_j \in U$.

Fuzzy equivalence class $[x_i]_{E_f}$ of x_i is defined by the following fuzzy set on U:

$$[x_i]_{E_f} = \frac{e_{i1}}{x_1} + \frac{e_{i2}}{x_2} + \cdots + \frac{e_{im}}{x_m} \tag{2}$$

Fuzzy entropy of feature f based on E_f is defined as

$$H(f) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{m}{|[x_i]_{E_f}|} \tag{3}$$

where, the cardinal value of $[x_i]_{E_f}$ is defined as $|[x_i]_{E_f}|= \sum_{j=1}^{m} e_{ij}$.

Indiscernibility relation $IND(\bar{F})$ defines a set of objects that have the same equivalence class, where $\bar{F} \subseteq F$. The fuzzy partition of U on $IND(\bar{F})$ is defined by $U/IND(\bar{F}) = \{X_1, X_2, \ldots, X_{\bar{F}}\}$, where $X_k = \{x_j \in U | [x_i]_E = [x_j]_E\}$, and $x_i \in X_x, k = \{1, 2, \ldots, \bar{F}\}$.

The fuzzy lower approximation of a single fuzzy equivalence class X is defined as

$$\mu_{\underline{E_{\bar{F}}X}}(x_i) = \inf_{x_j \in U} I(\mu_{E_{\bar{F}}}(x_i, x_j), \mu_X(x_j)) \tag{4}$$

where $\mu_{E_{\bar{F}}}(x_i, x_j) = \bigcap_{f \in \bar{F}} \mu_{E_f}(x_i, x_j)$, and $I = min(1, 1 - x_i + x_j)$ is the fuzzy Łukasiewicz implicator.

The fuzzy positive region determines all the objects on U that discriminate the classes of $U/IND(C)$ based on a set of features $\bar{F}$. The fuzzy positive region is defined as

$$\mu_{POS_{E_{\bar{F}}}(C)}(x_i) = \sup_{X \in U/C} \mu_{\underline{E_{\bar{F}}}X}(x_i) \tag{5}$$

2.1. Relevancy

Let E_f and E_C are two fuzzy relations of feature f and class C on U, respectively. Then, the fuzzy mutual information between f and C is defined as

$$I(f; C) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{m|[x_i]_{E_f} \cap [x_i]_{E_C}|}{|[x_i]_{E_f}|.|[x_i]_{E_C}|} \tag{6}$$

2.2. Redundancy

Let E_{f_1} and E_{f_2} be two fuzzy relations of features f_1 and f_2 on U, respectively. Then, the fuzzy mutual information between f_1 and f_2 is defined as

$$I(f_1; f_2) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{m|[x_i]_{E_{f_1}} \cap [x_i]_{E_{f_2}}|}{|[x_i]_{E_{f_1}}|.|[x_i]_{E_{f_2}}|} \tag{7}$$

2.3. Dependency

Let $\bar{F}$ is a set of features, the dependency degree of $\bar{F}$ is defined as

$$\Upsilon_{\bar{F}}(C) = \frac{\sum_{x_i \in U} \mu_{POS_{E_{\bar{F}}}(C)}(x_i)}{|[x_i]_{E_f}|} \tag{8}$$

2.4. Example

To illustrate the computations of previous relations, a small example is presented in Table 1. Firstly, we estimate the relation matrix of each feature based on the following similarity equation [15]:

$$E_f(x_i, x_j) = \exp\left(-\|x_i - x_j\|\right) \tag{9}$$

As C contains discrete values, we estimate the relation matrix according to the crisp way [19].

Table 1. An example of a small dataset, contains two features (f_1, f_2), and class C.

f_1	f_2	C
0.2	0.1	1
0.8	0.5	0
0.4	0.3	1
0.6	0.4	0
0.2	0.1	1

The relation matrix of f_1 is:

$$M(E_{f1}) = \begin{pmatrix} 1.00 & 0.55 & 0.82 & 0.67 & 1.00 \\ 0.55 & 1.00 & 0.67 & 0.82 & 0.55 \\ 0.82 & 0.67 & 1.00 & 0.82 & 0.82 \\ 0.67 & 0.82 & 0.82 & 1.00 & 0.67 \\ 1.00 & 0.55 & 0.82 & 0.67 & 1.00 \end{pmatrix}$$

The relation matrix of f_2 is:

$$M(E_{f2}) = \begin{pmatrix} 1.00 & 0.67 & 0.82 & 0.74 & 1.00 \\ 0.67 & 1.00 & 0.82 & 0.90 & 0.67 \\ 0.82 & 0.82 & 1.00 & 0.90 & 0.82 \\ 0.74 & 0.90 & 0.90 & 1.00 & 0.74 \\ 1.00 & 0.67 & 0.82 & 0.74 & 1.00 \end{pmatrix}$$

The relation matrix of C is:

$$M(E_C) = \begin{pmatrix} 1.0 & 0.0 & 1.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 1.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 1.0 & 0.0 & 1.0 \end{pmatrix}$$

The fuzzy entropy of f_1 is:

$$H(f_1) = \frac{1}{5}(\log\frac{5}{4.04} + \log\frac{5}{3.59} + \log\frac{5}{4.13} + \log\frac{5}{3.98} + \log\frac{5}{4.04}) = 0.34$$

The fuzzy entropy of C is:

$$H(C) = \frac{1}{5}(\log\frac{5}{3} + \log\frac{5}{2} + \log\frac{5}{3} + \log\frac{5}{2} + \log\frac{5}{3}) = 0.97$$

The relevancy between f_1 and C is:

$$I(f_1, C) = \frac{1}{5}(\log\frac{5*2.82}{4.04*3} + \log\frac{5*1.82}{3.59*2} + \log\frac{5*2.64}{4.13*3} + \log\frac{5*1.82}{3.98*2} + \log\frac{5*2.82}{4.04*3}) = 0.21$$

The redundancy between f_1 and f_1 is:

$$I(f_1, f_2) = \frac{1}{5}(\log\frac{5*4.04}{4.04*4.23} + \log\frac{5*3.59}{3.59*4.06} + \log\frac{5*4.13}{4.13*4.36} + \log\frac{5*3.98}{3.98*4.29} + \log\frac{5*4.04}{4.04*4.23}) = 0.24$$

For the first object x_1 of f_1, the fuzzy lower approximation of a single fuzzy equivalence class $X = 1$ is:

$$\mu_{\underline{E_{f_1}X=1}}(x_1) = \inf_{x_j \in U} I(\mu_{E_{\bar{F}}}(x_1, x_j), \mu_X = 1(x_j)) = \inf\{I(1.00, 1.0), I(0.55, 0.0), I(0.82, 1.0), I(0.67, 0.0), I(1.00, 1.0)\} = 0$$

For the first object x_1 of f_1, The fuzzy lower approximation of a single fuzzy equivalence class $X = 0$ is:

$$\mu_{\underline{E_{f_1}X=0}}(x_1) = \inf_{x_j \in U} I(\mu_{E_{\bar{F}}}(x_1, x_j), \mu_X = 0(x_j)) = \inf\{I(1.00, 0.0), I(0.55, 1.0), I(0.82, 0.0), I(0.67, 1.0), I(1.00, 0.0)\} = 0.33$$

Similarly, for the remaining objects of f_1:

$$\mu_{\underline{E_{f_1}X=1}}(x_2) = 0.33, \mu_{\underline{E_{f_1}X=0}}(x_2) = 0$$
$$\mu_{\underline{E_{f_1}X=1}}(x_3) = 0.0, \mu_{\underline{E_{f_1}X=1}}(x_3) = 0.18$$
$$\mu_{\underline{E_{f_1}X=1}}(x_4) = 0.18, \mu_{\underline{E_{f_1}X=1}}(x_4) = 0.0$$
$$\mu_{\underline{E_{f_1}X=1}}(x_5) = 0.0, \mu_{\underline{E_{f_1}X=1}}(x_5) = 0.33$$

The fuzzy positive region for the first object $x = 1$ is:

$$\mu_{POS_{E_{f_1}}(C)}(x_1) = \sup_{X \in 1,0} \mu_{\underline{E_{f_1}}X}(x_1) = \sup_{X \in 1,0} (0, 1) = 0.33$$

For the remaining objects, the fuzzy positive region are:

$$\mu_{POS_{E_{f_1}}(C)}(x_2) = 0.33$$
$$\mu_{POS_{E_{f_1}}(C)}(x_3) = 0.18$$
$$\mu_{POS_{E_{f_1}}(C)}(x_4) = 0.18$$
$$\mu_{POS_{E_{f_1}}(C)}(x_5) = 0.33$$

The dependency degree of f_1 is:

$$\Upsilon_{f_1}(C) = \frac{\sum_{x_i \in U} \mu_{POS_{E_{f_1}}(C)}(x_i)}{|U|} = \frac{0.33 + 0.33 + 0.18 + 0.18 + 0.33}{5} = 0.27$$

3. Related Works

Filter approach evaluates the feature significant based on the characteristics of data only with full independence of classification models [1]. Although the filter approach has many benefits over embedded and wrapper approaches, it may fail to find the best feature subset [20]. For this reason, a great research effort has been introduced to study the feature characteristics with the aim to find the significant features that improve classification models.

Among a variety of evaluation measures, mutual information (MI) has a popularity solution in feature selection based information theory due to its ability to define different relation of features such as relevancy, and redundancy. The main advantages of MI are [3]: (1) ability to deal with deal linear and non-linear relations among features; (2) ability to deal with both categorical and numerical features. In the past decades, MI has been used in many feature selection methods. Mutual information maximization (MIM) [21] defines the significance of features based on the relevancy relation. It suffers from the redundant features. After that, mutual information based feature selection (MIFS) [22] has been introduced and improved in MIFS-U [23] to define the significance of features based on both relevancy and redundancy relation. However, both methods require a predefined parameter to balance between the relevancy and redundancy relations. In [24], minimum redundancy maximum relevance (mRMR) proposes automatic value to estimate the predefined parameter of MIFS, and MIFS-U. In the literature, several feature selection methods have been proposed to find the best estimation of the relevancy and redundancy relations such as joint mutual information (JMI) [25], conditional mutual information maximization (CMIM) [26], joint mutual information maximization (JMIM) [27], and max-relevance and max-independence (MRI) [28]. However, previous studies of feature selection based mutual information do not consider the balance of selected/candidate feature relevancy relation. To avoid this limitation, Zhang et al. [29] has introduced a new method to keep the balance between the feature relevancy relations, called feature selection based on weighted relevancy (WRFS).

Another important solution in the filter approach is rough set which used to measure the dependency relation of features. Feature selection based rough set tries to find the minimal feature subset that maximizes the informative structure of all features (termed a reduct) [30]. The main advantages of the rough set are (1) analyzing only the hidden facts in data, (2) extracting the hidden knowledge of data without additional user-defined information, and (3) returning a minimal knowledge structure of data [19]. Many studies on feature selection based on rough set have been done. Rough Set Attribute Reduction (RSAR) defines the significance of a subset of features based on the dependency relation [31]. However, there is no guarantee to return the minimum feature subset. Han et al. [32] proposes an alternative dependency relation to reduce the computational cost of the feature selection process. Zhong et al. [33] defines the significance of the feature subset based on the discernibility matrix. However, it is impractical for high dimensionality data. In Entropy Based Reduction (EBR) [34], the significance of the feature subset is defined based on entropy which returns the maximum amount of information. In the literature of rough sets, further feature selection methods have been introduced such as Variable precision rough sets (VPRS) [35], and parameterized average support heuristic (PASH) [36].

However, both MI and rough set share common limitations when dealing with features of continuous values [19,37]. There are two traditional solutions have been proposed to overcome this limitation: parzen window [38], and discretization process [39]. The former has some limitations: firstly it requires a predefined parameter to compute the window function [40]. Secondly, it does not work efficiently with high dimensional data of spare samples [15]. The latter may lead to loss of feature information [41]. To overcome these limitations, FS based information theory and FS based rough set have been extended by fuzzy set theory to deal with continuous features directly [14,19]. However, most of FS methods based information theory focus on relevancy and redundancy relation, while FS methods based on rough set focus on dependency relation. The former depends on individually discriminative ability, while the latter depends on dependency discriminative ability. As a result, the traditional methods do not take the benefits of all types of discriminative ability.

4. Fuzzy Feature Selection Based on Relevancy, Redundancy, and Dependency (FFS-RRD)

In this section, we present our proposed method, called FFS-RRD, as a filter feature selection method. The effectiveness of filter methods depends on the amount of extracted information from the data characteristics. To promote our proposed method, we used both individually and dependency discriminative ability based on three criteria: relevancy, redundancy, and dependency. FFS-RRD aims to maximize both relevancy and dependency relations and minimize the redundancy ones. To design our proposed method, firstly, we modified WRFS to overcome their limitations: (1) it can not deal with continuous features without the discretization process which may lead to loss of feature information. (2) WRFS does not consider dependency relation in the feature selection process. To overcome these limitations, we estimated WRFS based on the fuzzy concept instead of the probability concept. The extended method, called FWRFS, can deal with any numerical data without the discretization process. Then, we combined FWRFS with fuzzy-rough lower approximations (L-FRFS) [16] to extract the dependency relation. Consequently, we proposed a unique FS method, called FFSRRD, which maximizes both relevancy and dependency, and minimizes the redundancy relation. The three relations can extract more information from the dataset to promote the discriminate ability of feature selection. Figure 1 shows the process of the proposed method FFS-RRD. Both FWRFS and L-FRFS are applied on the same dataset. FWRFS selects the most relevant features and removes the redundancy ones, while L-FRFS selects the most dependency feature subset. The results of each method are combined to return the final feature selection subset. In our study, we used one of the popular combination methods called MIN [2]. MIN method assigns the minimum position of each feature among different results of feature selection methods to be ranked position in the final result.

The algorithm of the proposed method is presented in Algorithm 1. FFS-RRD depends on a combination of two methods: For the first method, FWRFS is used to return the ranked feature set that maximizes the relevancy and minimizes the redundancy. In the first step (Lines 1–3), The main parameters are initialized: ranked feature set (R_1), candidate feature (*candidate*), and the current selected feature set (*selected*). Then, the feature of maximum relevancy with class is selected to be the first ranked feature in R_1, and removed from the feature set F (Lines 4–8). After that, the feature of maximum relevancy with class and minimum redundancy with selected features is added to R_1, and removed from F. This process is repeated until all features of F are ranked in R_1 (Lines 9–14). For the second method, L-FRFS is used to return the subset of features that maximizes the dependency relation. In the first step (Lines 15–17), the main parameters are initialized: selected feature subset (R_2), temporary feature (T), maximum dependency degree (Υ_{select}), and the last maximum dependency degree (Υ_{last}). Then, the feature of maximum dependency is added to R_2. This process is repeated until the maximum possible dependency degree of features be produced (Lines 18–25). Finally, the result of both methods is combined by $MIN(R_1, R_2)$ to select the final feature subset (Line 26–27).

Algorithm 1: FFS-RRD: fuzzy feature selection based relevancy, redundancy, and dependency.

```
   Input: A dataset $D = (F \cup C)$, where $F = \{f_1, f_2, \dots, f_n\}$ is a set of $n$ features, and $C$ is a class label
   Output: $R$ Ranked set of features
   // Method 1: fuzzy weighted relevancy-based FS (FWRFS)
1  $R_1 \leftarrow \emptyset$
2  $selected \leftarrow \emptyset$
3  $candidate \leftarrow \emptyset$
4  for $i \leftarrow 1$ to $n$ do
5      $candidate(f_i) \leftarrow I(f_i;C)$
6  $selected(f_i) \leftarrow max_{f_i \in F} candidate(f_i)$;
7  $R_1 \leftarrow R_1 \cup selected(f_i)$;
8  $F \leftarrow F - selected(f_i)$;
9  while $length(R_1) \neq 0$ do
10     for $j \leftarrow 1$ to $length(F)$ do
11         $candidate(f_j) \leftarrow \sum_{f_j \in R_1} \frac{I(f_i,f_j;C)}{H(C)} * I(f_i;C|f_j) + \frac{I(f_j;C)}{H(C)} * I(f_j;C|f_i) - I(f_i;f_j)$
12     $selected(f_j) \leftarrow max_{f_j \in F} candidate(f_j)$
13     $R_1 \leftarrow R_1 \cup selected(f_j)$
14     $F \leftarrow F - selected(f_i)$
   // Method 2: fuzzy lower approximation-based FS (L-FRFS)
15 $R_2 \leftarrow \emptyset$
16 $\Upsilon_{select} \leftarrow 0$
17 $\Upsilon_{last} \leftarrow 0$
18 while $\Upsilon_{select} \neq \Upsilon_{last}$ do
19     $T \leftarrow R_2$
20     $\Upsilon_{last} \leftarrow \Upsilon_{select}$
21     foreach $f_i \in (F - R_2)$ do
22         if $\Upsilon_T(C) \leq \Upsilon_{R_2 \cup f_i}(C)$ then
23             $T \leftarrow R_2 \cup f_i$
24             $\Upsilon_{select} \leftarrow \Upsilon_T(C)$
25     $R_2 \leftarrow T$
   // combination: FWRFS and L-FRFS
26 $R \leftarrow MIN(R_1, R_2)$
27 return $R$
```

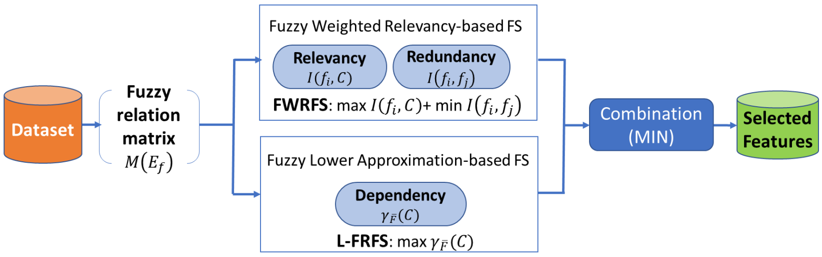
Dataset
Fuzzy relation matrix $M(E_f)$
Fuzzy Weighted Relevancy-based FS
Relevancy $I(f_i, C)$
Redundancy $I(f_i, f_j)$
FWRFS: max $I(f_i, C)$+ min $I(f_i, f_j)$
Fuzzy Lower Approximation-based FS
Dependency $\gamma_{\bar{F}}(C)$
L-FRFS: max $\gamma_{\bar{F}}(C)$
Combination (MIN)
Selected Features

Figure 1. The process of our proposed method fuzzy feature selection based relevancy, redundancy, and dependency (FFS-RRD): firstly, the fuzzy relation matrix is generated for each feature in the dataset. Then, fuzzy mutual information maximizes the relevancy and minimizes the redundancy, while fuzzy rough set maximizes the dependency. Finally, the results are combined to find the selected features.

5. Experiment Setup

The main goal of the feature selection process is to improve the classification performance with the minimum feature selection subset. To validate our proposed method, we used four classifiers to compare the proposed method with eight feature selection methods based on benchmark datasets. Figure 2 shows the framework of our experiment. In the following, we present more details about the experiment setup.

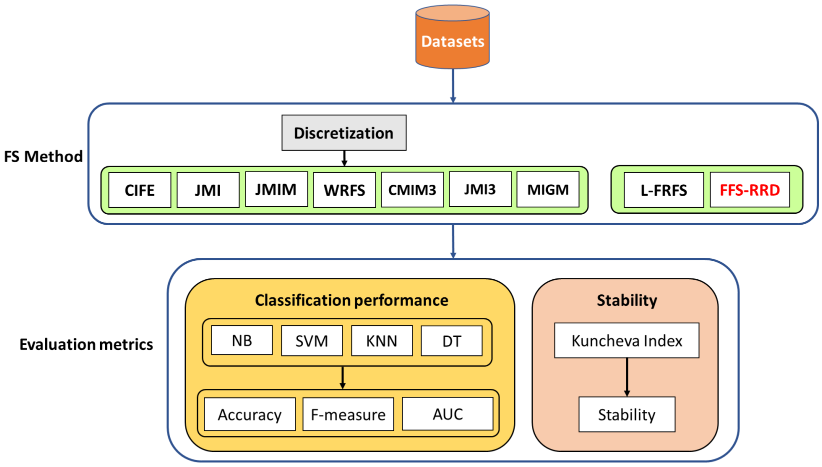

Figure 2. An experimental framework of the proposed method fuzzy feature selection based relevancy, redundancy, and dependency (FFS-RRD): firstly, a discretization process is applied before probability-based methods. Then, the compared methods are evaluated in terms of classification performance, and stability.

5.1. Dataset

Our experiment was conducted based on 13 benchmark datasets from machine learning repository (UCI) [42]. The datasets support different classification problems of binary and multi-class data. Table 2 presents a brief description of the experimental datasets.

5.2. Compared Feature Selection Methods

Table 3 shows the compared FS methods and their discriminative ability. The compared methods can be divided into two groups: probability-based, and fuzzy-based. Firstly, probability-based group uses the probability concept to estimate information measures. The probability-based group consists of CIFE [43], JMI [25], JMIM [27], WRFS [29], CMIM3 [44], JMI3 [44], and MIGM [45]. This group depends on the discretization process before implementation of feature selection methods. In our experiment, the discretization process transforms the continuous features into discrete features with ten equal intervals [46].

Unlike a probability-based group which requires discretization preprocess, fuzzy-based group uses the fuzzy concept to estimate information measures. The fuzzy-based group includes L-FRFS [16], and the proposed method FFS-RRD. This group depends on similarity relation which transforms each feature into a fuzzy equivalence relation. In our experiment, we used the following similarity relation [15].

$$E_f(x_i, x_j) = \exp\left(-\|x_i - x_j\|\right) \tag{10}$$

Table 2. Description of datasets used in the experiments.

Dataset	Brief	# Instances	# Features	#Classes
Breast Cancer Wisconsin (Prognostic)	BCW Prognostic	198	33	2
Breast Cancer Wisconsin (Diagnostic)	BCW Diagnostic	569	31	2
Climate Model Simulation Crashes	CMSC	540	18	2
Credit Approval	Credit Approval	690	15	2
Dermatology	Dermatology	336	34	6
Diabetic Retinopathy Debrecen	DRD	1151	19	2
Fertility	Fertility	100	9	2
Statlog (Heart)	Heart	270	13	2
Ionosphere	Ionosphere	351	34	2
Iris	Iris	150	4	3
Libras Movement	Libras Movement	360	90	15
QSAR biodegradation	QSAR	1055	41	2
Zoo	Zoo	101	16	7

Table 3. The extracted feature relations of compared feature selection methods.

Ref.	FS Group	FS Method	Discriminative Ability	
			Individually	Dependency
[43]	Probability-based	CIFE	✓	
[25]		JMI	✓	
[27]		JMIM	✓	
[29]		WRFS	✓	
[44]		CMIM3	✓	
[44]		JMI3	✓	
[45]		MIGM		✓
[16]	Fuzzy-based	L-FRFS		✓
Proposed		FFS-RRD	✓	✓

5.3. Evaluation Metrics

The main factors characterize the quality of feature selection methods are its classification performance and stability [47]. The evaluation of our experiment is divided into two parts: classification performance and stability evaluation. Classification performance requires classification models to evaluate the effect of feature selection methods on improving the classification performance, while stability measures the robustness of feature selection methods.

5.3.1. Classification Performance

To evaluate the classification performance, we used three metrics: classification accuracy, F-measure ($\beta = 1$), AUC. The experiment depends on four classifiers: Naive bayes (NB), support vector machine (SVM), K-nearest neighbors (KNN, K = 3), and decision tree (DT). To find reliable results, we used 10-fold cross-validation where the dataset is divided into ten equal parts, nine for the training phase and one for the test phase [48]. This process is repeated ten times. Then, we calculate the average results to compute the score of accuracy, F-measure, and AUC.

In this experiment, we used a threshold to cut the ranked features and return a subset of selected features. The threshold is the median position of the ranked features (or the nearest integer position if the number of ranked features is even). For L-FRFS, we used the same threshold if the size of the returned subset is more than the median of all features.

5.3.2. Stability Evaluation

The confidence of feature selection method is not only about the improvement of classification performance but also related to the robustness of the method [49]. The robustness of feature selection method against any small change of data, as a noise, is called feature selection stability [50]. In the stability experiment, we injected the data by 10% of noise which is generated based on standard deviation and the gaussian distribution of each feature [51]. Then, we run the feature selection method to return the sequence of features. This process is repeated for ten times with a new returned sequence each time. After that, we measure the stability for each feature selection method based on Kuncheva stability measure which is defined as [52]:

$$Kun_{stab} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} Kun_{index}(R_i, R_j) \tag{11}$$

where p is the number of feature selection sequences, and $Kun_{index}(R_i, R_j)$ is the Kuncheva stability index between two feature selection sequences R_i, and R_j which is defined as:

$$Kun_{index}(R_i, R_j) = \frac{wn - r^2}{r(n-r)}, \tag{12}$$

where $w = |R_i \cap R_j|$, $r = |R_i| = |R_j|$, and n is the total number of features.

6. Results Analysis

6.1. Classification Performance

6.1.1. Accuracy

Based on NB classifier, it is obvious that FFS-RRD achieved the maximum average accuracy with score 83.4%, as shown in Table 4. The proposed method was more accurate than compared methods by the range from 0.4% to 1.8%. The order of methods ranked after FFS-RRD was JMIM, followed by JMI, both CMIM3 and JMI3, MIGM, WRFS, L-FRFS, and CIFE.

According to SVM classifier, FFS-RRD achieved the maximum average accuracy of all datasets by 86.4%, while L-FRFS achieved the minimum average accuracy by 84.1%, as shown in Table 5. The proposed method outperformed other methods in the range from 0.5% to 2.3%. The second-best feature selection method was JMI, followed by CMIM3, both JMIM and JMI3, WRFS, MIGM, and CIFE.

In the case of KNN classifier, FFS-RRD also was the best feature selection method in the term of average accuracy by 85.4%, while L-FRFS was the worst method by 82.5%, as shown in Table 6. After that, MIGM achieved the second-best method, followed by both JMI and JMIM, JMI3, WRFS, CMIM3, CIFE. The proposed method achieved better accuracy in the range from 0.5% to 2.9%.

Similarly, FFS-RRD kept the best average accuracy of DT classifier by 84.5%, as shown in Table 7. The proposed method outperformed other methods in the range from 0.4% to 1.4%. In contrast, both CIFE and L-FRFS achieved the worst results by 83.1%. The second-best feature selection method was JMI, followed by JMIM, both WRFS and JMI3, MIGM, and CMIM3.

Table 4. Average classification accuracy on Naive bayes (NB) classifier: our proposed method achieved the best result.

Dataset	CIFE	JMI	JMIM	WRFS	CMIM3	JMI3	MIGM	L-FRFS	FFS-RRD
BCW Prognostic	73.9	69.5	73.8	65.4	68.6	68.8	67.7	71.3	69.8
BCW Diagnostic	92.0	93.4	93.4	93.6	93.2	92.4	92.9	85.9	93.6
CMSC	91.9	93.8	91.9	94.1	93.6	92.8	93.8	93.7	93.8
Credit Approval	85.6	83.5	83.7	86.9	82.4	83.7	84.8	76.8	85.4
Dermatology	96.1	93.9	95.3	93.1	98.0	94.6	95.3	96.2	96.1
DRD	57.6	60.3	57.6	57.5	57.7	60.6	60.3	57.6	57.5
Fertility	87.9	87.9	88.0	88.0	88.0	88.0	87.9	88.0	88.0
Heart	80.1	83.0	81.0	84.1	83.0	83.9	80.1	75.1	81.5
Ionosphere	77.0	82.6	86.2	80.8	84.8	84.8	77.4	89.8	89.0
Iris	94.6	94.6	94.6	94.6	93.5	93.5	93.5	95.0	95.0
Libras Movement	51.4	61.7	60.8	50.5	59.0	59.0	59.4	61.0	60.0
QSAR	78.2	78.7	77.1	78.5	78.7	78.1	77.6	77.7	80.0
Zoo	94.2	95.1	96.0	96.0	96.1	96.0	96.9	93.5	94.9
Average	81.6	82.9	83.0	81.8	82.8	82.8	82.1	81.7	83.4

Table 5. Average classification accuracy on support vector machine (SVM) classifier: our proposed method achieved the best result.

Dataset	CIFE	JMI	JMIM	WRFS	CMIM3	JMI3	MIGM	L-FRFS	FFS-RRD
BCW Prognostic	76.3	76.8	77.9	76.3	76.4	77.7	76.6	76.3	77.5
BCW Diagnostic	96.3	97.3	97.3	97.5	96.9	95.0	95.6	88.4	96.3
CMSC	91.5	92.0	91.5	93.2	91.9	91.6	91.9	92.1	91.9
Credit Approval	85.5	85.5	85.5	85.5	85.5	85.5	85.5	73.7	85.5
Dermatology	95.8	95.5	95.8	94.2	98.2	96.8	95.7	96.8	96.5
DRD	68.0	67.2	68.0	67.7	67.5	67.0	67.2	68.0	67.7
Fertility	88.0	88.0	88.0	88.0	88.0	88.0	88.0	88.0	88.0
Heart	79.7	84.3	84.3	84.2	81.3	83.7	79.7	76.3	82.5
Ionosphere	76.6	79.0	78.5	81.2	81.7	82.9	78.0	86.7	87.6
Iris	95.7	95.7	95.7	95.9	93.9	93.9	93.9	94.5	95.0
Libras Movement	72.2	76.1	74.7	67.6	75.5	74.9	74.5	76.5	75.3
QSAR	84.2	84.3	84.5	82.3	84.2	83.7	84.4	83.8	84.1
Zoo	92.8	95.3	88.7	93.3	93.9	89.5	94.1	92.0	95.3
Average	84.8	85.9	85.4	85.1	85.8	85.4	85.0	84.1	86.4

Table 6. Average classification accuracy on K-nearest neighbors (KNN) classifier: our proposed method achieved the best result.

Dataset	CIFE	JMI	JMIM	WRFS	CMIM3	JMI3	MIGM	L-FRFS	FFS-RRD
BCW Prognostic	75.2	75.0	77.1	76.6	73.5	74.1	76.9	74.0	77.0
BCW Diagnostic	93.3	93.5	97.1	95.4	95.1	94.6	95.2	87.5	94.8
CMSC	90.1	92.3	90.1	92.0	90.9	92.4	92.6	93.4	93.0
Credit Approval	84.7	83.9	85.1	85.5	84.4	85.3	85.1	76.4	84.6
Dermatology	95.0	95.9	95.6	93.7	97.7	96.9	96.1	96.1	96.3
DRD	64.3	63.7	64.3	64.1	62.3	63.7	63.6	64.7	64.1
Fertility	88.7	88.7	87.2	86.7	84.0	90.5	88.7	85.0	89.4
Heart	77.7	79.1	79.1	81.7	79.3	77.7	77.6	71.7	78.2
Ionosphere	81.8	80.9	80.3	83.0	82.3	83.3	83.5	82.7	84.2
Iris	93.9	93.9	93.9	93.9	91.8	91.8	91.8	92.7	92.7
Libras Movement	70.7	76.8	77.9	70.4	77.1	74.7	78.3	75.8	78.3
QSAR	84.0	84.7	83.3	81.5	83.8	84.4	83.6	82.5	83.0
Zoo	90.1	94.1	91.6	93.1	92.1	92.1	91.2	89.7	94.3
Average	83.8	84.8	84.8	84.4	84.2	84.7	84.9	82.5	85.4

Table 7. Average classification accuracy on decision tree (DT) classifier: our proposed method achieved the best result.

Dataset	CIFE	JMI	JMIM	WRFS	CMIM3	JMI3	MIGM	L-FRFS	FFS-RRD
BCW Prognostic	73.8	73.7	73.1	72.5	72.8	73.0	73.0	76.6	73.2
BCW Diagnostic	93.7	94.5	94.1	94.2	94.0	93.5	93.5	88.6	94.6
CMSC	89.6	91.2	89.6	90.3	91.4	91.2	91.3	91.3	91.6
Credit Approval	85.5	85.5	85.7	86.3	84.8	85.7	85.3	75.3	86.3
Dermatology	93.6	92.3	92.2	91.3	93.5	94.1	92.5	94.3	94.9
DRD	68.0	65.8	68.0	67.7	65.0	66.8	65.9	67.7	67.6
Fertility	87.1	87.1	87.1	87.6	86.6	86.6	87.1	87.5	87.5
Heart	73.9	80.3	81.3	80.6	76.3	77.6	74.0	74.3	75.5
Ionosphere	88.4	88.4	88.4	89.7	88.8	88.4	88.0	88.6	89.0
Iris	90.8	90.8	90.8	90.8	91.9	91.9	91.9	91.7	91.7
Libras Movement	60.4	66.0	65.6	61.3	64.7	64.3	67.5	66.0	66.4
QSAR	83.2	83.7	82.5	83.5	82.9	83.5	82.3	82.3	82.9
Zoo	91.9	94.0	92.3	92.9	93.4	92.1	94.1	96.0	97.1
Average	83.1	84.1	83.9	83.7	83.5	83.7	83.6	83.1	84.5

6.1.2. F-Measure

Figure 3 shows the F-measure of the compared methods based on the four used classifiers. In NB classifier, FSS-RRD achieved the maximum average F-measure by 88.5%, while MIGM achieved the minimum score by 82.4%. The proposed method outperformed other methods in the range from 1.5% to 6.1%. Similarly, FSS-RRD achieved the maximum average F-measure using SVM by 88.9%, while MIGM achieved the minimum score by 83.5%. The proposed method outperformed other methods in the range from 0.5% to 5.4%. According to KNN classifier, WRFS achieved the maximum average F-measure by 87.7%, while CMIM3 achieved the minimum score by 81.6%. The proposed method achieved the fourth-best position in this case. In DT classifier, FSS-RRD achieved the maximum average F-measure by 87.5%, while MIGM achieved the minimum score by 82.3%. The proposed method outperformed other methods in the range from 0.7% to 5.2%.

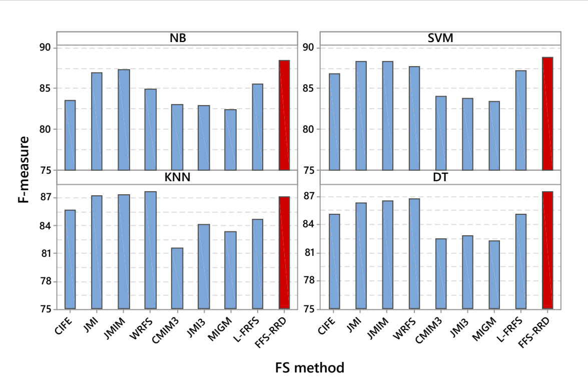

Figure 3. Average F-measure on the four used classifiers. Our proposed method (FFS-RRD) achieved the best result in all cases except KNN.

6.1.3. AUC

It is obvious that the proposed method achieved the highest AUC compared with other methods using all classifiers (Figure 4). According to NB, FSS-RDD achieved the maximum AUC by 87.8%, while CIFE achieved the minimum AUC by 86.2%. The proposed method outperformed other methods in the range from 0.2% to 1.6%. In SVM classifier, FSS-RDD also achieved the maximum AUC by 81.1%, while L-FRFS achieved the minimum score by 78.2%. The proposed method outperformed other methods in the range from 0.9% to 2.9%. Similarly, FSS-RDD also achieved the maximum AUC using KNN by 85.1%, while L-FRFS achieved the minimum score by 83.3%. The proposed method outperformed other methods in the range from 0.4% to 1.8%. Using DT classifier, FSS-RDD kept the best method by 82.2%, L-FRFS kept the worst method by 79.9%. The proposed method outperformed other methods in the range from 0.4% to 2.3%.

Figure 5 shows the average score of the four classifiers in terms of accuracy, F-measure, and AUC. For accuracy term, FFS-RRD achieved the highest accuracy for all classifiers by 84.9%, followed by JMI, JMIM, JMI3, CMIM3, MIGM, WRFS, CIFE, and L-FRFS by 84.4%, 84.3%, 84.2%, 84.1%, 83.9%, 83.8%, 83.3%, and 82.8%, respectively. The proposed method outperformed other methods in the range from 0.5% to 2.1%. Similarly, FFS-RRD achieved the highest average of F-measure by 87.6%. Then, JMIM achieved the second-best method, followed by JMI, WRFS, L-FRFS, CIFE, JMI3, MIGM, CMIM3 with score 87.4%, 87.3%, 86.8%, 85.7%, 85.7%, 85.3%, 83.4%, 83.9%, 82.8%, respectively. The proposed method outperformed other methods in the range from 0.2% to 4.7%. According to AUC, FFS-RRD achieved the highest AUC with a score 84.0%. JMI3 achieved the second-best method, followed by JMI, CMIM3, WRFS, MIGM, JMIM, CIFE, and L-FRFS by 83.6%, 83.5%, 83.4%, 83.3%, 83.2%, 83.0%, 82.4%, and 81.9%, respectively. The proposed method outperformed other methods in the range from 0.4% to 2.1%.

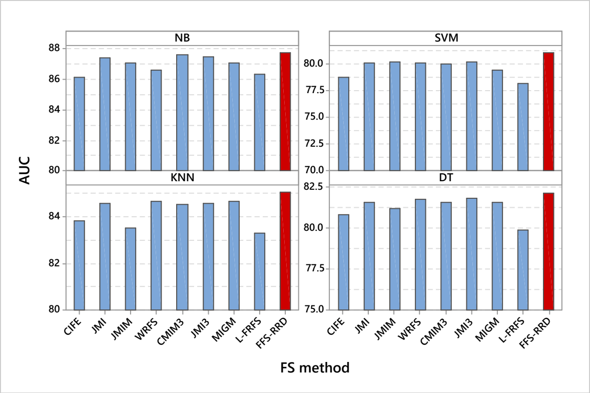

Figure 4. Average AUC on the four used classifiers. Our proposed method (FFS-RRD) achieved the best result in all cases.

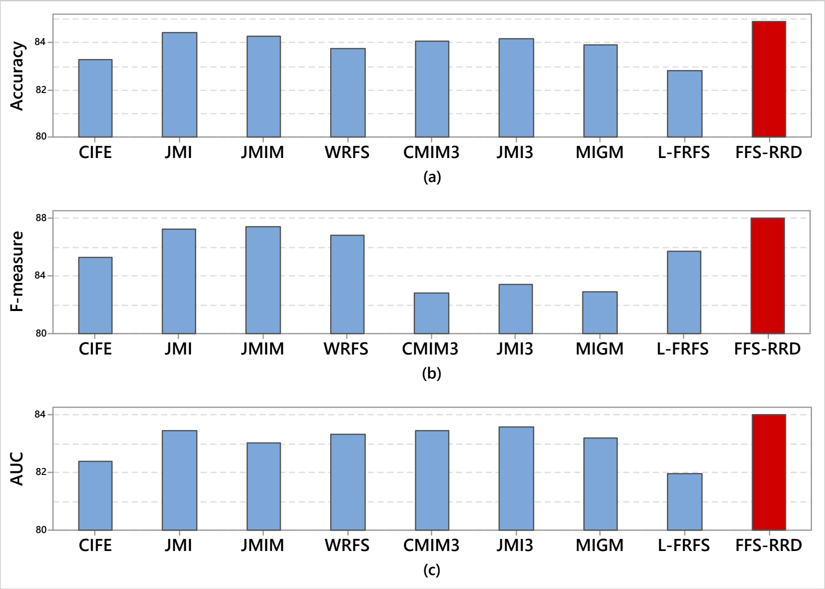

Figure 5. Average score of all classifiers in terms of accuracy, F-measure, and AUC. Our proposed method (FFS-RRD) achieved the best result.

6.2. Stability

Figure 6 shows the average stability across the first half thresholds on all datasets. FFS-RRS achieved the maximum average of stability by 84.3%, while MIGM achieved the minimum score by 67.9%. After that, L-FRFS achieved the second-best method by 78.6%, followed by JMI3, CIFE, JMI, WRFS, CMIM3, and JMIM with an average score 76.0%. 75.8%, 73.5%, 73.0%, 72.3%, and 71.9%, respectively. The proposed method outperformed other methods in the range from 5.7% to 16.4%. Figure 7 shows a box-plot of average stability for all compared methods on the median threshold. In box-plot, the black circle represents the stability median, while the box represents both lower and upper quartiles. As shown in the box-plot, the stability result of the proposed method is better and more consistent than compared methods.

By considering the previous results, it is obvious that FFS-RRD achieved the best experimental results in the term of classification performance and feature stability. This is expected where the proposed method considers the individually and dependency discriminative ability of features. On the other hand, it is obvious that fuzzy-based methods are more stable than probability-based methods. The reason returns to using fuzzy sets to estimate the feature significance without information loss. Consequently, it helps fuzzy-based methods to be more stable against the noise.

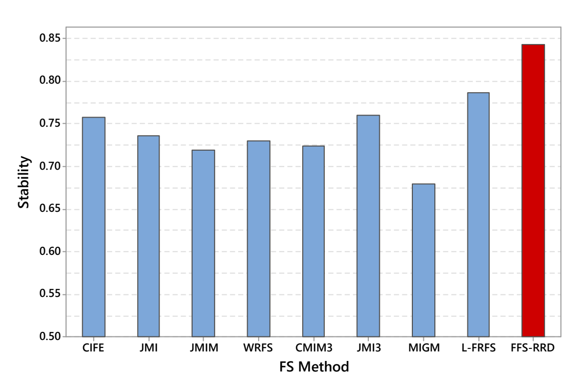

Figure 6. Average stability across the first half thresholds on all datasets. Our proposed method (FFS-RRD) achieved the best result.

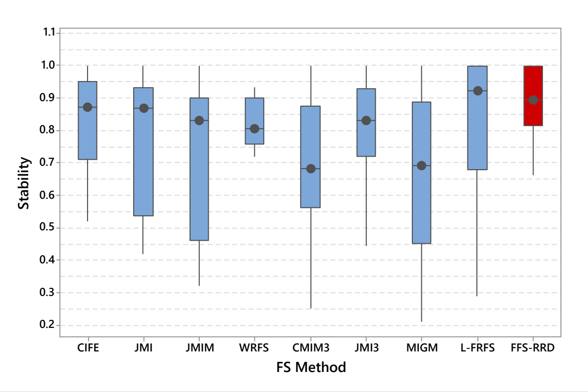

Figure 7. Average stability for all compared methods on the median threshold. Our proposed method (FFS-RRD) achieved the best result.

7. Conclusions

In this paper, we have proposed an ensemble feature selection method, fuzzy feature selection based on relevancy, redundancy, and dependency criteria (FFS-RRD). Unlike the traditional methods, FFS-RRD depends on both individually and dependency discriminative ability. FFS-RRD aims to extract the significant relations from data characteristics to find the best feature subset that improves the performance of classification models. The proposed method consist of combination of two methods: FWRFS, and L-FRFS. FWRFS maximizes the relevancy and minimizes the redundancy relation, while L-FRFS maximizes the dependency relation.

Compared with eight state-of-the-art and conventional FS methods, experiments on 13 benchmark datasets indicate the outperformance of the proposed method in classification performance and stability. Classification performance includes three measures accuracy, F-measure, and AUC. The proposed method FFS-RRD achieved the highest average score of accuracy, and AUC on all datasets, while it achieved the highest average of F-measure on most of the classifiers except KNN classifier. On the other hand, the proposed method achieved the highest average of stability compared with other feature selection methods. In future work, we will extend the proposed method to explore their effect on multi-label classification models.

Author Contributions: Methodology, O.A.M.S., F.L., and Y.-P.P.C.; software, O.A.M.S.; investigation, F.L., X.C., and Y.-P.P.C.; writing—original draft preparation, O.A.M.S., and F.L. ; writing—review and editing, F.L., O.A.M.S., X.C., and Y.-P.P.C.; supervision, F.L., and X.C.; funding acquisition, F.L. All authors read and approved the final version of the manuscript.

Acknowledgments: This research has been supported by the National Natural Science Foundation (61572368).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Macedo, F.; Oliveira, M.R.; Pacheco, A.; Valadas, R. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing* **2019**, *325*, 67–89. [CrossRef]
2. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: a review and future trends. *Inf. Fusion* **2019**, *52*, 1–12. [CrossRef]
3. Lee, S.; Park, Y.T.; d'Auriol, B.J. A novel feature selection method based on normalized mutual information. *Appl. Intell.* **2012**, *37*, 100–120.
4. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; de Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1106–1119. [CrossRef]
5. Imani, M.B.; Keyvanpour, M.R.; Azmi, R. A novel embedded feature selection method: A comparative study in the application of text categorization. *Appl. Artif. Intell.* **2013**, *27*, 408–427. [CrossRef]
6. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
7. Hu, L.; Gao, W.; Zhao, K.; Zhang, P.; Wang, F. Feature selection considering two types of feature relevancy and feature interdependency. *Expert Syst. Appl.* **2018**, *93*, 423–434. [CrossRef]
8. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
9. Oreski, D.; Oreski, S.; Klicek, B. Effects of dataset characteristics on the performance of feature selection techniques. *Appl. Soft Comput.* **2017**, *52*, 109–119. [CrossRef]
10. Bonev, B. *Feature Selection Based on Information Theory*; Universidad de Alicante: Alicante, Spain, 2010.
11. Caballero, Y.; Alvarez, D.; Bello, R.; Garcia, M.M. Feature selection algorithms using rough set theory. In Proceedings of the IEEE Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro, Brazilm, 20–24 October 2007; pp. 407–411.
12. Che, J.; Yang, Y.; Li, L.; Bai, X.; Zhang, S.; Deng, C. Maximum relevance minimum common redundancy feature selection for nonlinear data. *Inf. Sci.* **2017**, *409*, 68–86. [CrossRef]
13. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1991.
14. Hu, Q.; Yu, D.; Xie, Z. Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognit. Lett.* **2006**, *27*, 414–423. [CrossRef]
15. Yu, D.; An, S.; Hu, Q. Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection. *Int. J. Comput. Intell. Syst.* **2011**, *4*, 619–633. [CrossRef]
16. Jensen, R.; Shen, Q. New approaches to fuzzy-rough feature selection. *IEEE Trans. Fuzzy Syst.* **2008**, *17*, 824–838. [CrossRef]
17. Hüllermeier, E. Fuzzy sets in machine learning and data mining. *Appl. Soft Comput.* **2011**, *11*, 1493–1505. [CrossRef]
18. Freeman, C.; Kulić, D.; Basir, O. An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognit.* **2015**, *48*, 1812–1826. [CrossRef]
19. Jensen, R.; Shen, Q. Fuzzy-rough sets for descriptive dimensionality reduction. In Proceedings of the 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE'02, Honolulu, HI, USA, 12–17 May 2002; Proceedings (Cat. No. 02CH37291), Volume 1, pp. 29–34.
20. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
21. Lewis, D.D. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1992; pp. 212–217.
22. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef]
23. Kwak, N.; Choi, C.H. Input feature selection for classification problems. *IEEE Trans. Neural Netw.* **2002**, *13*, 143–159. [CrossRef]

24. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Onpattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
25. Yang, H.; Moody, J. Feature selection based on joint mutual information. In Proceedings of the International ICSC Symposium on Advances in Intelligent Data Analysis, Genova, Italy, 1–4 June 1999; pp. 22–25.
26. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
27. Bennasar, M.; Hicks, Y.; Setchi, R. Feature selection using joint mutual information maximisation. *Expert Syst. Appl.* **2015**, *42*, 8520–8532. [CrossRef]
28. Wang, J.; Wei, J.M.; Yang, Z.; Wang, S.Q. Feature selection by maximizing independent classification information. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 828–841. [CrossRef]
29. Zhang, P.; Gao, W.; Liu, G. Feature selection considering weighted relevancy. *Appl. Intell.* **2018**, *48*, 4615–4625. [CrossRef]
30. Hassanien, A.E.; Suraj, Z.; Slezak, D.; Lingras, P. *Rough Computing: Theories, Technologies and Applications*; IGI Global Hershey: Hershey, PA, USA, 2008.
31. Chouchoulas, A.; Shen, Q. Rough set-aided keyword reduction for text categorization. *Appl. Artif. Intell.* **2001**, *15*, 843–873. [CrossRef]
32. Han, J.; Hu, X.; Lin, T.Y. Feature subset selection based on relative dependency between attributes. In *International Conference on Rough Sets and Current Trends in Computing*; Springer: Berlin, Germany, 2004; pp. 176–185.
33. Zhong, N.; Dong, J.; Ohsuga, S. Using rough sets with heuristics for feature selection. *J. Intell. Inf. Syst.* **2001**, *16*, 199–214. [CrossRef]
34. Jensen, R.; Shen, Q. Fuzzy–rough attribute reduction with application to web categorization. *Fuzzy Sets Syst.* **2004**, *141*, 469–485. [CrossRef]
35. Ziarko, W. Variable precision rough set model. *J. Comput. Syst. Sci.* **1993**, *46*, 39–59. [CrossRef]
36. Zhang, M.; Yao, J. A rough sets based approach to feature selection. In Proceedings of the IEEE Annual Meeting of the Fuzzy Information, Banff, AB, Canada, 27–30 June 2004; Processing NAFIPS'04, Volume 1, pp. 434–439.
37. Ching, J.Y.; Wong, A.K.; Chan, K.C.C. Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 641–651. [CrossRef]
38. Kwak, N.; Choi, C.H. Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1667–1671. [CrossRef]
39. Garcia, S.; Luengo, J.; Sáez, J.A.; Lopez, V.; Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **2012**, *25*, 734–750. [CrossRef]
40. Herman, G.; Zhang, B.; Wang, Y.; Ye, G.; Chen, F. Mutual information-based method for selecting informative feature sets. *Pattern Recognit.* **2013**, *46*, 3315–3327. [CrossRef]
41. Shen, Q.; Jensen, R. Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring. *Pattern Recognit.* **2004**, *37*, 1351–1363. [CrossRef]
42. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2017.
43. Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 68–82.
44. Sechidis, K.; Azzimonti, L.; Pocock, A.; Corani, G.; Weatherall, J.; Brown, G. Efficient feature selection using shrinkage estimators. *Mach. Learn.* **2019**, *108*, 1261–1286. [CrossRef]
45. Wang, X.; Guo, B.; Shen, Y.; Zhou, C.; Duan, X. Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization. *IEEE Access* **2019**, *7*, 151525–151538. [CrossRef]
46. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*; Elsevier: Amsterdam, The Netherlands, 1995; pp. 194–202.
47. Li, Y.; Si, J.; Zhou, G.; Huang, S.; Chen, S. FREL: A stable feature selection algorithm. *IEEE Trans. Neural Networks Learn. Syst.* **2014**, *26*, 1388–1402. [CrossRef]
48. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* **1995**, *14*, 1137–1145.

49. Pes, B.; Dessì, N.; Angioni, M. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Inf. Fusion* **2017**, *35*, 132–147. [CrossRef]
50. Nogueira, S.; Brown, G. Measuring the stability of feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 442–457.
51. Tsai, Y.S.; Yang, U.C.; Chung, I.F.; Huang, C.D. A comparison of mutual and fuzzy-mutual information-based feature selection strategies. In Proceedings of the 2013 IEEE International Conference on Fuzzy Systems (FUZZ), Hyderabad, India, 7–10 July 2013; pp. 1–6.
52. Kuncheva, L.I. A stability index for feature selection. In Proceedings of the 25th IASTED International Multi-Conference Artificial Intelligence and Applications, Innsbruck, Austria, 12–14 Fedbruary 2007; pp. 421–427.