



# New filter approaches for feature selection using differential evolution and fuzzy rough set theory

Emrah Hancer<sup>1</sup>

Received: 11 March 2019 / Accepted: 17 January 2020 / Published online: 28 January 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

Nowadays the incredibly advanced developments in information technologies have led to exponential growth in the datasets with respect to both the dimensionality and the sample size. This trend can be easily illustrated in popular online data repositories (e.g., UCI machine learning repository). The more growth in the datasets, the more challenging the data management becomes. This is because such datasets usually comprise a high level of noise as well as the necessary information. Therefore, the elimination of noisy features in the datasets is an important task to discover meaningful knowledge. Although a large number of feature selection approaches have been proposed in the literature to deal with noisy features, the need for the studies based on feature selection has not come to an end. In this paper, we propose differential evolution-based feature selection approaches wrapped around the principles of fuzzy rough set theory. In contrast to well-known feature selection criteria such as standard mutual information, standard rough set and probabilistic rough set, our approaches can also deal with real-valued variables without the requirement of discretization. Moreover, the feature subsets selected by our approaches can profoundly improve the classification performance compared to the recent particle swarm optimization approaches based on probabilistic rough set and the state-of-the-art filter approaches.

**Keywords** Fuzzy rough set · Differential evolution · Feature selection · Classification

## 1 Introduction

Nowadays a real-world dataset is usually represented through many features, causing the existence of noisy features to the target. There are two kinds of noisy features, irrelevant and redundant. An irrelevant feature does not have a direct relationship with the target, but adversely affects the performance of the learning process. A redundant feature does not provide any additional information to the target. Removing such features can improve both the efficiency and the effectiveness of a learning algorithm. It can therefore be possible to get a better insight into a learning problem. However, the selection process of an optimal feature subset in the dataset is not an easy task due to the following challenges. First, an individual feature

may not be significantly correlated with the target. However, if this feature is evaluated with other features, it may have a considerable relation with the target. Second, for a dataset with  $m$  number of features, there exist  $2^m$  candidate feature subsets. Accordingly, the selection process of an optimal feature subset among all candidate subsets may become intensive and impractical, even for a medium number of features. Third and last, the selection process of an optimal feature subset is expected to improve not only the classification performance, but should also consider other important issues, such as interpretability, complexity, scalability and generalization of the learning model.

In the terminology of machine learning, feature selection can be applied using the principles of three fundamental schemes: supervised, unsupervised and semi-supervised [36]. Supervised feature selection requires the labeled information in the data to identify the goodness of features. The labeled information can be a real or order value, or a category depending on the specific task. Semi-supervised feature selection only requires a small amount of the labeled information. In contrast to supervised and

---

✉ Emrah Hancer  
emrahhanc@gmail.com; ehancer@mehmetakif.edu.tr

<sup>1</sup> Department of Computer Technology and Information Systems, Mehmet Akif Ersoy University, 15039 Burdur, Turkey

semi-supervised schemes, unsupervised feature selection does not require any external or labeled information to carry out the selection process of an optimal feature subset and so is treated as more challenging.

In terms of evaluation, both supervised and unsupervised feature selection approaches can be divided into wrapper, embedded and filter approaches. Wrapper approaches first select a candidate feature subset using a predefined search strategy (e.g., sequential forward, sequential floating forward, random search) and then use a specific learning algorithm to evaluate the quality of the chosen feature subset. This procedure is repeated until achieving higher learning performance. However, the selected feature subset is dependent on the specific learning algorithm, i.e., may not perform well through another learning algorithm. Embedded approaches are adopted into learning algorithms through regularization methods or a penalty term in a model-based function. Compared to wrapper approaches, embedded approaches are less time-consuming, but may not perform well in terms of maximizing the performance of a learning algorithm. In contrast to wrapper and embedded approaches, filter approaches do not require a learning algorithm to carry out the selection process of features. Filter approaches are therefore computationally less intensive and more general than wrapper and embedded approaches, but may not perform well due to the independence of learning algorithms. Instead of a learning algorithm, filter approaches use data intrinsic measures such as distance, dependency and consistency to evaluate the quality of the selected subsets [11, 42]. Unfortunately, none of them have been considered as the standard for filter approaches. In our opinion, this is because none of these measures are consistently better than others on all problems, i.e., some metrics works properly in continuous data, while some metrics are suitable for discrete data. Moreover, they detect different kinds of relationships between features and output variables.

Rough set (RS) theory [31] is a mathematical tool that can be used to reduce the dimensionality using intrinsic information within the data without requiring any additional information. For the last two decades, RS theory has gained considerable interest and has been implemented to many fields [49]. For a given dataset represented through discrete-valued features, it is possible to select the most informative features among all available features using RS theory. In the terminology of feature selection, the informative features are expected to be the most predictive to the target. However, RS theory does not consider the degree of overlapping in the data. To deal with this deficiency, Yao and Zhao [45] introduced probabilistic (PRS) rough set theory. Even though PRS theory has proven to perform better than RS theory, PRS theory has still challenges to measure the dependency of features.

Furthermore, both RS and PRS theories encounter problems when the values of features are real-valued. Using the principles of such theories, it is not possible to identify that two different feature values are similar and to what extent they are same. For example, two close values may only be considered as different as a result of noise, but RS theory considers them as two different values with a different order of magnitude. Data discretization may be considered as a solution to this challenge before performing RS-based and PRS-based reduction approaches; however, it is still not adequate since information loss in the data cannot be preventable. In order to alleviate such problems, fuzzy variants of RS-based criteria concerned with fuzzy rough set (FRS) theory were introduced.

In order to enhance the performance of RS-based approaches and its derived criteria, researchers have also implemented such measures in evolutionary computation (EC)-based filter approaches due to the search ability of them to find the global optima. However, there exist only few works on using EC techniques and FRS-based criteria for feature selection in the literature. Moreover, to the best of our knowledge, FRS-based standards have not been adopted in multi-objective EC framework yet. Another open issue concerning feature selection is that the works for filter feature selection concerned with differential evolution (DE) [37] are not adequate in the literature, even though DE has been widely treated as one of the most potent algorithms among a variety of EC techniques.

## 1.1 Goals

In this paper, we propose new filter approaches using DE algorithm and FRS theory to improve both the effectiveness and the efficiency of the classification algorithms. To achieve this objective, we designed FRS-based objective functions for single-objective and multi-objective DE algorithms. Specifically, we aim to investigate:

- the performance of FRS-based approaches versus PRS-based approaches,
- the performance of FRS-based criteria in single-objective design versus in multi-objective design.
- the performance of FRS-based approaches versus the state-of-the-art filter approaches.

## 1.2 Organization

The remainder of the paper is designed as follows. Section 2 presents related work: rough set theory, differential evolution, multi-objective optimization and rough set theory for feature selection. Section 3 introduces the proposed approaches: the design of FRS-based criteria and the integration of the designed FRS-based criteria in single-

objective and multi-objective frameworks. Section 4 gives an outline of the experiment design. Section 5 provides empirical results with analysis. Finally, Sect. 6 draws conclusions and future work.

## 2 Background

### 2.1 Rough set (RS) theory

Proposed by Pawlak [31] in 1982, RS theory is a mathematical tool to deal with uncertainty in an efficient way in the data using a decision table. Let  $U$  be a non-empty finite set of instances (objects) in the data and  $A$  be a non-empty finite set of features such that  $a : U \xrightarrow{V} a$  where  $V_a$  is the set of values that attribute  $a$  can take. The decision table is shown by  $I = (U, A)$ . Moreover, assume that  $P$  and  $Q$  be two subsets of  $A$ , and  $X$  be a subset of  $U$ . The fundamental components of RS theory are given as follows.

1. *Indiscernibility* ( $IND(P)$ ) is a relation between two instances where all values are identical in relation to a subset of considered features ( $P$ ), defined by Eq. (1). If  $(x, y) \in IND(P)$ ,  $x$  and  $y$  instances are indiscernible through feature subset  $P$ . The equivalence classes of the indiscernibility relation of the instance  $x$  in subset  $P$  are denoted by  $[x]_P$ .

$$IND(P) = \{(x, y) \in U | \forall a \in P, a(x) = a(y)\} \tag{1}$$

2. *Set approximation* is investigated in two categories: lower approximation and upper approximation. Lower approximation of  $X$  consists of instances which can be classified with full certainty using feature subset  $P$ , defined by:

$$\underline{P}X = \{x \in U | [x]_P \subseteq X\} \tag{2}$$

Upper approximation consists of instances which may be probably classified using feature subset  $P$ , defined by:

$$\overline{P}X = \{x \in U | [x]_P \cap X \neq \emptyset\} \tag{3}$$

3. *Positive region* comprises all instances of  $U$  that can be uniquely classified into groups of  $U \setminus Q$  through feature subset  $P$ , defined by:

$$POS_P(Q) = \bigcup_{X \in U \setminus Q} \underline{P}X \tag{4}$$

4. *Attribute dependency* measures the correlation between features, defined by Eq. (5). The dependency increases proportionally to the  $k$  value. For  $k = 1$ ,  $Q$  is completely dependent on  $P$ . In an optimal feature subset, features are expected to have the minimum

dependency on each other and maximum correlation on the target.

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{5}$$

*An example:* We further provide an example to show how RS measures are calculated. We provide a dataset in Table 1. For  $P = \{a, b\}$ , the partition of  $U$  created by  $IND(P)$  which is denoted as  $U/IND(P)$  is defined as follows:

$$\begin{aligned} U/IND(P) &= U/IND(\{a\}) \otimes U/IND(\{b\}) \\ &= \{\{0, 3, 4\}, \{1, 7\}, \{2, 5, 6\}\} \otimes \{\{0, 2, 4\}, \{1, 3, 6, 7\}, \{5\}\} \\ &= \{\{0, 4\}, \{3\}, \{1, 7\}, \{2\}, \{6\}, \{5\}\} \end{aligned}$$

Let  $X$  be a subset such that  $X = \{X : e(X) = 1\}$ , and  $P = \{a, b\}$ . Accordingly,  $X = \{2, 4, 5, 7\}$  and  $IND(P) = \{\{0, 4\}, \{1, 7\}, \{2\}, \{3\}, \{5\}, \{6\}\}$ . The lower and upper approximations of  $X$  are defined as follows:

$$\begin{aligned} \underline{P}X &= \{2, 5\} \\ \overline{P}X &= \{0, 4, 1, 7, 2, 5\} \end{aligned}$$

For  $P = \{a, b\}$ , the positive region of the class labels  $e$  is determined as follows:

$$\begin{aligned} U/IND(\{e\}) &= \{\{0\}, \{2, 4, 5, 7\}, \{1, 3, 6\}\} \\ \underline{P}X &= \begin{cases} \emptyset, X = \{X : e(X) = 0\} \\ \{2, 5\}, X = \{X : e(X) = 1\} \\ \{3, 6\}, X = \{X : e(X) = 2\} \end{cases} \\ POS_{a,b}(e) &= \bigcup_{X \in U/e} \underline{P}X = \{2, 3, 5, 6\} \end{aligned}$$

The degree of dependency of the class labels  $e$  from the features  $a, b$  is calculated as follows:

$$\gamma_{a,b}(e) = \frac{|POS_{a,b}(e)|}{|U|} = \frac{|\{2, 3, 5, 6\}|}{|\{0, 1, 2, 3, 4, 5, 6, 7\}|} = \frac{4}{8}$$

**Table 1** An example dataset

$x \in U$	$a$	$b$	$c$	$d$	$e$
0	1	0	2	2	0
1	0	1	1	1	2
2	2	0	0	1	1
3	1	1	0	2	2
4	1	0	2	0	1
5	2	2	0	1	1
6	2	1	1	1	2
7	0	1	1	0	1

## 2.2 Differential evolution (DE)

An evolutionary algorithm developed by Storn and Price [37], DE carries out searching on the solution space on the basis of directional information. DE tries to find the global optima for a given optimization problem by forming a population with candidate solutions and then evolving new candidates through mutation and recombination strategies. In the population, a candidate solution is represented by a vector such that  $X_i = \{x_{i1}, x_{i2}, \dots, x_{id}, \dots, x_{iD}\}$ , where  $i$  denotes the  $i$ th individual and  $D$  denotes the dimensionality of the optimization problem. For each  $X_i$  solution, DE randomly selects three different solutions  $X_{r1}$ ,  $X_{r2}$  and  $X_{r3}$  in the population and then generates a new solution  $U_i$  using the direction of the selected vectors in a probabilistic manner:

$$u_{id} = \begin{cases} x_{r1d} + F \times (x_{r2d} - x_{r3d}) & \text{if } \text{rand}() < CR \text{ or } d = I_{rand} \\ x_{id} & \text{otherwise} \end{cases} \quad (6)$$

where  $F$  is the constant within the range of  $[0,2]$  which controls the rate of mutating solutions;  $\text{rand}()$  is the randomly distributed number between 0 and 1;  $I_{rand}$  is a randomly selected index between 1 and  $D$  which guarantee the evolution of a solution at least one position;  $CR$  is the recombination rate which controls the rate of interacting between solutions.

The general structure of DE is simple to implement and easy to understand, i.e., does not include complex components. Furthermore, it can be applied to a variety of problems such as nonlinear, non-differentiable, discrete and noisy since it does not require assumptions about the problem to be considered to optimize. On the other hand, it does not guarantee the global optima like as any other EC techniques [17, 19].

## 2.3 Multi-objective optimization

Nowadays we can face with many real-world problems with multiple objectives instead of a single objective. Since some objectives are tend to (partially) conflict, the design of multi-objective problems becomes more complex. The general definition of a multi-objective problem with  $k$  objectives,  $n$  inequalities and  $l$  equality constraints is posed as follows:

$$\begin{aligned} \min_x \mathbf{F}(x) &= [F_1(x), F_2(x), \dots, F_k(x)]^T \\ \text{subject to:} & \\ g_j(x) &\leq 0, \quad j = 1, 2, \dots, m \\ h_l(x) &= 0, \quad l = 1, 2, \dots, n \end{aligned} \quad (7)$$

When we consider a multi-objective problem, there exists no such a single global solution typically. Therefore, it is often necessary to find a set of points that all compromise on a predefined definition for an optimum. The general concept for defining an optimal point concerns with Pareto-optimality. A point  $x^*$  is Pareto-optimal iff there is no such another point  $x$  that  $F_i(x) \leq F_i(x^*)$  for all functions and  $F_i(x) < F_i(x^*)$  for at least one function.  $x^*$  is also referred to as a non-dominated solution. Therefore, the non-dominated solution set is called the Pareto-optimal set. For more information concerning multi-objective optimization, please see [30].

## 2.4 Rough set for feature selection

Earlier filter feature selection approaches on the basis of RS-based criteria generally follow traditional search strategies such as sequential forward, sequential floating forward and sequential backward. The most representative one is the Quickreduct algorithm [7] which starts with an empty set and adds a feature to this set for each iteration until reaching the maximum dependency based on the standard RS-based criterion. Prasad and Rao [32] improved the Quickreduct algorithm by a sequential reduction strategy. Another approach that uses a sequential backward elimination strategy was proposed by Gawar [13] by designing a RS-based relative dependency measure. Yong et al. [46] introduced an efficient Quickreduct algorithm by integrating neighborhood RS-based criteria. In summary, all such traditional approaches showed that RS-based criteria could be used as an alternative filter evaluation metric for feature selection. However, all of these conventional approaches tend to converge in local minima points due to their greedy search mechanisms. Accordingly, EC techniques have gained increasing attention due to their approved ability to avoid local minima while investigating the global optima.

Genetic algorithm (GA) [22], the first EC technique in the literature, was developed using the principles of evolution. Besides a variety of successful GA-based applications in many fields, GA has also been designed for RS-based filter approaches. Wroblewski [41] introduced a simple RS-based filter approach using GA. According to the empirical studies, this approach obtained better performance than traditional approaches. However, it was computationally expensive. Bjorvand [3] increased the effectiveness and efficiency of this approach by applying a software system (referred to as the Rough Enough) and a dynamic mutation rate. Jing [24] proposed a hybridized GA filter approach (HGARSTAR) which uses a RS-based local search mechanism to fine-tune parameters. HGARSTAR obtained promising results compared to existing EC-based

filter approaches based on RS-based criteria; however, HGARSTAR was not tested on high-dimensional datasets. Das et al. [8] developed an incremental feature selection (IFS) approach based on GA and RS theory. From the results, it can be inferred that IFS performed better than some representative filter approaches such as ReliefF [26], Correlated Feature Subset Selection (CFS) [15] and Consistency Feature Subset Selection (CON) [10] in both discrete-valued and real-valued datasets. However, the datasets used in experiments included at most 36 features, i.e., it is not possible to make an analysis of IFS on high-dimensional datasets.

Another important EC technique, particle swarm optimization (PSO) [25] mimics the behaviors of natural species which are bird flock or fish school. PSO has also been used to develop RS-based applications. Wang et al. [39] introduced a RS-based approach (PSORSFS) using PSO. In this approach, the dependency between features and the target, and the ratio of the feature subset size to the total number of features were tried to be optimized in a weighted manner. According to the results, PSORSFS obtained great performance compared to the RS-based approaches on the basis of greedy search and GA algorithms. However, the control parameter between the dependency and the feature subset size in the fitness function is difficult to be determined. Abdul-Rahman et al. [1] introduced a two-stage wrapper-filter approach that considers a RS-based reduction approach using PSO as an earlier step to reduce the dimensionality, and a wrapper approach as a latter step to find an optimal feature subset on the reduced dimensionality. Cervante et al. [5] developed a RS-based approach (PSOPRS) using PSO and probabilistic rough set theory (PRS). The results showed that PRS theory could enhance the performance of classifiers better than the standard RS theory. However, like as previous RS-based approaches using PSO, PSOPRS is dependent on the control parameter between the dependency and the feature subset size and cannot be applied to evaluate the dependency of real-valued features without discretization. In order to alleviate the control parameter, Bing et al. [42] considered the PRS-based criterion as a multi-objective problem and introduced a multi-objective PSO filter feature selection approach. Without discretization, this multi-objective approach also cannot be applied to measure real-valued features.

Another well-known EC technique is the differential evolution (DE) algorithm [37], the details of which are presented in Sect. 2.2. Yan and Li [43] proposed a RS-based filter approach using an improved version of DE. However, it is not possible to make a consistent analysis concerning this approach since it was only tested on two datasets. Sangeetha and Kalpana [35] introduced a FRS-based filter approach (FRFSDE) using DE inspired by the fuzzy version of the Quickreduct algorithm. From the

results, it can be revealed that FRFSDE can perform well compared to the fuzzy version of Quickreduct. However, the feature subset size obtained by feature selection approaches was not considered in experiments. Das et al. [9] introduced a RS theory and relational algebra-based filter approach using a multi-objective DE binary algorithm. However, the feature subset size obtained by filter approaches was not investigated and analyzed in the experimental study. Besides GA, PSO and DE, recently developed EC techniques such as artificial bee colony [6], cuckoo search [2], ant lion optimizer [28] have also been used to develop RS-based filter approaches.

In summary, it is observed that RS-based criteria can increase the performance of classification algorithms when they are incorporation with EC techniques. However, there are still some open issues that need to be considered. First, most of the RS-based approaches are based on GA and PSO. In other words, there are only a few RS-based approaches based on DE despite its potential to search the global optimal. Second, most of the RS-based criteria cannot deal with real-valued datasets due to the deficiency of them to quantify continuous variables. Third, adopting FRS-based criteria, fuzzy variants of RS-based criteria, in EC-based frameworks is not common in the literature, to the best of our knowledge. Finally and most importantly, FRS-based criteria have not been modeled as a multi-objective problem in the literature yet.

### 3 Proposed approaches

In this section, we first describe the concepts of fuzzy rough set theory and then give the details of FRS-based approaches using single-objective and multi-objective DE algorithms.

#### 3.1 Fuzzy rough set (FRS) theory

In the terminology of RS theory, instances in the data belong to the lower approximation with certainty or not. In contrast to RS theory, in the terminology of FRS theory, instances may belong to the lower approximation with a membership value in the range of  $[0, 1]$ . This means FRS theory allows flexibility to deal with uncertainty. Let  $U$  be the set of instances (objects) in the data,  $A$  be the features describing instances and  $P$  and  $Q$  are two subsets of  $A$ ; the decision table is defined as  $I = (U, A)$ . The degree of similarity between  $x$  and  $y$  instances through all features in subset  $P$  is defined by:

$$\mu_{R_P}(x, y) = T_{a \in P} \{ \mu_{R_a}(x, y) \} \quad (8)$$

where  $T$  denotes the  $t$ -norm; and  $\mu_{R_a}(x, y)$  is the degree of



similarity between  $x$  and  $y$  instances for feature  $a$ , which can be measured by Eqs. (9) and (10).

$$\mu_{R_a}(x, y) = 1 - \frac{|a(x) - a(y)|}{a_{\max} - a_{\min}} \quad (9)$$

where  $a_{\max}$  and  $a_{\min}$  denote the maximum and minimum values for feature  $a$ .

$$\mu_{R_a}(x, y) = \exp\left(-\frac{(a(x) - a(y))^2}{2\sigma_a^2}\right) \quad (10)$$

where  $\sigma_a^2$  denotes the variance of feature  $a$ .

Using the degree of similarities within the feature subset  $P$ , fuzzy lower and upper approximations are defined by Eqs. (11) and (12).

$$\mu_{R_P X}(x) = \inf_{y \in U} I(\mu_{R_P}(x, y), \mu_X(y)) \quad (11)$$

where  $I$  denotes the fuzzy implicator.

$$\mu_{\overline{R_P X}}(x) = \sup_{y \in U} T(\mu_{R_P}(x, y), \mu_X(y)) \quad (12)$$

The fuzzy positive region is then defined by:

$$\mu_{POS_R(P)}(x) = \sup_{x \in U/Q} \mu_{R_P X}(x) \quad (13)$$

Based on the fuzzy dependency of  $Q$  on  $P$  ( $\gamma'_P(Q)$ ) is defined by Eq. (14). For more information concerning FRS theory, please see [23].

$$\gamma'_P(Q) = \frac{|\mu_{POS_R(P)}(x)|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_R(P)}(x)}{|U|} \quad (14)$$

### 3.2 Proposed FRS-based approaches

In the terminology of RS theory, the information in the data is represented by the decision table in classification tasks. Each instance in the data is considered as an object in RS theory. The target variable in the data, known as the class labels, is regarded as the decision feature  $D$ , and the features building the data are the conditional features  $C$  such that  $A = C \cup D$ . All instances ( $U$ ) in the data can be divided into different classes  $U_1, U_2, \dots, U_n$ , where  $n$  is the total number of classes. The overall aim of a RS-based approach is to eliminate irrelevant, redundant and noisy features, so the remaining feature subset  $P \subset A$  represents the information in the data as well as  $C$ .

(1) DEFERS approach: As described in Sect. 2.1, an

equivalence relation is an essential notion to determine lower and upper approximations of the target in standard RS theory. For the lower approximation, the equivalence class should be a subset of the target. For the upper approximation, the equivalence class should have a non-empty overlap with the target. However, the lack of overlapping degrees limits the application domain of rough sets. To alleviate this deficiency, a number of generalized approximation operators have been developed. One of the generalized variants is PRS theory which is based on the notions of rough membership functions and rough inclusion. Both notions can be interpreted in terms of conditional probabilities or posteriori probabilities. However, PRS theory also limits the application domain due to the dependency of crisp binary relations. Thus, researchers have established fuzzy rough set models where a fuzzy-based similarity relation can be used instead of an equivalence relation. The main advantage of such models is that an instance can belong to more than one class with a membership degree. Based on this fuzzy measure, we propose a new DE-based single-objective approach (DEFERS), where Eq. (15) is designed as the fitness function. DEFERS aims to maximize the dependency of the feature subset to the target and also aims to minimize the feature subset size.

$$Fit_{frs} = \beta * \frac{|\mu_{POS_R(P)}(x)|}{|U|} + (1 - \beta) * \frac{1 - \#subsets\ size}{\#total\ features} \quad (15)$$

where  $\beta \in (0, 1]$  controls the importance of the dependency and the importance of the feature subset size. When  $\beta = 1$ , DEFERS directly evaluates the candidate feature subset using the FRS model. The pseudocode of DEFERS is presented in Algorithm 1. The membership values between features ( $\mu_{R_a}(x, y)$ ) are calculated beforehand and then are considered as the inputs for the evolutionary process. By this way, it is aimed to decrease the computational complexity of the evolutionary process. If any position of a candidate solution is greater than 0.5, its corresponding feature is selected for the feature subset. After the selection process of the feature subset, the quality of the feature subset is evaluated using Eq. (15).

After the single-objective evolutionary process, the final feature subset is evaluated on the testing set with the training set through a predefined classifier to obtain the classification performance of the selected feature subset. Notice that the testing set is transparent to the user during the evolutionary process.

**Algorithm 1:** Pseudo code of DEFERS

---

**Input:** Training set, Testing set,  $\mu_{Ra}(x, y)$ ,  $\beta$   
**Output:**  $g_{best}$ , classification accuracy

```

begin
  Initialize a population of candidate solutions;
  Evaluate the fitness value of each candidate  $X_i$  on the training set by Eq. (15);
  for  $iter \leftarrow 1$  to  $MaxIter$  do
    for  $i \leftarrow 1$  to  $PopSize$  do
      Randomly choose three candidates for candidate  $X_i$ ;
      Evolve a new solution  $U_i$  using the chosen candidates by Eq. (6);
      Evaluate the fitness value of  $U_i$  on the training set by Eq. (15);
      Update candidate  $X_i$  through greedy selection between  $U_i$  and  $X_i$ ;
    end
    Update  $g_{best}$  in current population;
  end
  Evaluate the classification accuracy of the  $g_{best}$  on the testing set through the selected classifiers;
end

```

---

**Algorithm 2:** Pseudo code of MDEFERS

---

**Input:** Training set, Testing set,  $\mu_{Ra}(x, y)$   
**Output:** Pareto Front, classification accuracies

```

begin
  Initialize a population of candidate solutions;
  Evaluate two objective values of each candidate  $X_i$  on the training set;
  for  $iter \leftarrow 1$  to  $MaxIter$  do
    for  $i \leftarrow 1$  to  $PopSize$  do
      Randomly choose three candidates for candidate  $X_i$ ;
      Evolve a new solution  $U_i$  using the chosen candidates by Eq. (6);
      Evaluate two objective values of  $U_i$  on the training set;
      Update candidate  $X_i$  through dominance-based selection between  $U_i$  and  $X_i$ ;
    end
    Identify the Pareto Front;
  end
  Evaluate the classification accuracies of the non-dominated solutions in the Pareto Front on the testing set through the selected classifiers;
end

```

---

(2) MDEFERS approach: DEFERS uses two main objectives in a weighted manner in a single fitness function through the  $\beta$  control parameter which specifies the weight degrees of the objectives. However, the  $\beta$  parameter needs to be predetermined according to the problem, i.e., it is not possible to assign a general  $\beta$  value for all problems. Accordingly, we design a FRS-based multi-objective DE feature selection approach (MDEFERS) to get rid of the  $\beta$  parameter. In MDEFERS, the two objectives which are to maximize the dependency of the feature subset to the target and to minimize the number of features are individually evaluated without using the  $\beta$  parameter.

Due to its recently proposed successful applications for feature selection in the literature [16, 18, 20], we use a simple multi-objective DE variant (MODE) [33] which can be treated as a specified version of the pure DE algorithm for multi-objective problems. Except for the selection scheme which is based on the dominance-based selection

scheme inspired by Lampinen's criterion [27], MODE follows the same procedures as in DE. The pseudocode of MDEFERS is presented in Algorithm 2. As in DEFERS, the threshold value used to select features for the feature subset is set to 0.5. After the multi-objective evolutionary process, the non-dominated solutions in the Pareto-Front are evaluated on the testing set with the training set through a predefined classifier to obtain the classification accuracies. Notice that the testing set is transparent to the user during the multi-objective evolutionary process as in the single-objective evolutionary process.

## 4 Experimental design

Thirty independent runs are carried out for each proposed approach to get meaningful results. All approaches are implemented using Matlab on an Intel Core i7 machine

with 3.6GHz CPU and 16 GB RAM. To verify the effectiveness of the proposed FRS-based approaches, the experiments are conducted on two stages. First, we analyze the performance of the proposed FRS-based approaches by comparing them with PRS-based approaches on six well-known categorical datasets from [12], shown in Table 2. For comparisons, we directly utilize the results of PRS-based approaches reported in [42]. As in [42], we randomly divide each categorical dataset as 7/10 of the instances for training and 3/10 of the instances for testing. For the evaluation of the selected feature subsets, we choose the K-nearest neighbors (KNN) and the decision tree (DT) classification algorithms as in [42], where K is specified as 5 [20].

Second, we analyze the performance of the proposed single-objective FRS-based approach by comparing it with a variety of well-known filter approaches, including feature selection via concave minimization (FSV) [4], local learning-based clustering (LLC-FS) [48], mutual information (MI) [38], Pearson correlation coefficient (PCC) [40], ReliefF [34], Fisher Score [14] and OFS-Density [50] on fourteen real-valued benchmarks from popular repositories [12, 44, 47], shown in Table 3. For comparisons, we directly utilize the results of filter approaches reported in [50]. In order to make consistent and reliable analysis, we follow the same procedure used in [50] such that we apply tenfold cross-validation on each real-valued dataset to divide 9/10 of the instances for training and 1/10 of the instances for testing, and for the evaluation of the selected feature subsets, we choose the support vector machines (SVM) and the KNN classification algorithms.

The parameter values of FRS-based approaches are determined as follows. The population size and the maximum number of cycles are both set to 50;  $F$  and  $CR$  are, respectively, set to 0.8 and 0.7 for single-objective DE framework, and 0.5 and 0.2 for multi-objective DE framework as suggested in [20, 29]. Notice that the maximum number of evaluations for PRS-based approaches using PSO was determined as 6000 in [42], but is determined as only 2500 for FRS-based approaches. It can

**Table 2** Categorical datasets used in experiments

Dataset	Samples	Features	Classes
Lymph	148	18	4
Spect	267	22	2
Dermatology	366	33	6
Soybean large	307	35	19
Chess	3196	36	2
Waveform	5000	40	3

**Table 3** Real-valued datasets used in experiments

Dataset	Samples	Features	Classes
Ionosphere	351	34	2
WBCD	569	30	2
Sonar	208	60	2
Hill	606	100	2
Colon	62	2000	2
Srbct	83	2308	4
Lung2	203	3312	5
Lymphoma	62	4026	3
Glioma	50	4433	4
MLL	72	5848	3
Prostate	102	6033	2
DLBCL	77	6285	2
Leukemia	72	7129	2
Arcene	200	10000	2

therefore be undoubtedly inferred that FRS-based approaches are computationally more efficient than PRS-based approaches. In single-objective FRS-based approach, three different  $\beta$  values (1, 0.9 and 0.5) are used to represent the relative importance of the dependency and the number of features in Eq. (15) (Table 4).

## 5 Experimental results

To verify the effectiveness of the proposed FRS-based approaches, we make comparisons with the PRS-based and the state-of-the-art filter approaches through single-objective design and multi-objective design in this section.

### 5.1 Comparison between DEFERS and PSOPRS

The results of PSOPRS and DEFERS are presented in terms of the average classification accuracy and the average number of selected features through 5NN and DT classifiers in Table 4, where the best values are denoted by bold symbol, and the average number of selected features is represented by ‘NOF.’ In order to further show the difference between PSOPRS and DEFERS in terms of the classification accuracy, the Wilcoxon rank-sum test is applied with the significance level 0.05. The results of the Wilcoxon rank-sum test are presented in Table 5, where ‘+’ denotes DEFERS is significantly superior to PSOPRS, ‘−’ denotes DEFERS is worse than PSOPRS, and ‘=’ denotes DEFERS achieves similar results with PSOPRS.

According to Table 4, it can be indicated that DEFERS obtains higher classification accuracy than PSOPRS in



**Table 4** Results of PSOPRS and DEFERS through 5NN and DT classifiers

Dataset	$\beta$	NOF	PSOPRS			DEFERS	
			5NN	DT	NOF	5NN	DT
Lymph	1	1.77	77.55 + 4.65	72.31 + 6.82	13.4	<b>86.42 + 1.78</b>	<b>77.38 + 10.38</b>
	0.9	5.03	75.31 + 1.1	66.73 + 3.3	6	<b>88.09 + 3E-16</b>	<b>85.71 + 4E-16</b>
	0.5	5	75.51 + 2e-4	67.35 + 3E-4	4	<b>83.33 + 5E-16</b>	<b>85.71 + 4E-16</b>
Spect	1	17.5	<b>81.24 + 1.57</b>	80.04 + 1.96	20.13	78.58 + 0.97	<b>84.16 + 2.86</b>
	0.9	14	<b>83.15 + 39E-4</b>	<b>82.02 + 25E-4</b>	7.03	80.12 + 0.5	80 + 3E-16
	0.5	3.1	<b>84.27 + 3E-4</b>	74.68 + 1.53	1	81.25 + 0	<b>81.25 + 0</b>
Dermatology	1	21	91.89 + 3.25	86.01 + 4.76	30.9	<b>97.21 + 0.24</b>	<b>95.23 + 3E-16</b>
	0.9	8.13	78.17 + 5.59	74.81 + 6.92	18.2	<b>94.28 + 0.43</b>	<b>91.68 + 0.42</b>
	0.5	6.8	73.33 + 5.78	72.87 + 9.48	7.6	<b>89.74 + 2.01</b>	<b>90.92 + 1.34</b>
Soybean	1	21.67	80.70 + 4.22	80.53 + 4.33	25.33	<b>84.16 + 2.25</b>	<b>86.92 + 1.33</b>
	0.9	9.7	68.44 + 3.75	72.36 + 2.87	12.43	<b>77.41 + 2.49</b>	<b>81.79 + 2.67</b>
	0.5	7.67	66.95 + 4.14	72.07 + 4.24	7.83	<b>70.74 + 2.46</b>	<b>77.85 + 5.18</b>
Chess	1	29.97	93.98 + 0.62	98.28 + 0.27	32.86	<b>94.67 + 1.13</b>	<b>99.35 + 6E-3</b>
	0.9	14.43	89.8 + 4.41	97.85 + 0.31	21.93	<b>89.85 + 1.48</b>	<b>99.14 + 0.21</b>
	0.5	5.4	71.82 + 17	93.61 + 0.82	10.9	<b>93.49 + 1.47</b>	<b>97.39 + 0.65</b>
Waveform	1	24.47	75.23 + 2.64	<b>74.79 + 1.88</b>	38.66	<b>79.51 + 0.65</b>	73.91 + 0.44
	0.9	8.03	65.43 + 6.41	68.6 + 5.62	21.53	<b>76.85 + 1.43</b>	<b>73.80 + 1.27</b>
	0.5	7	66.73 + 6.59	<b>70.21 + 2.24</b>	13.6	<b>70.11 + 2.15</b>	68.48 + 1.81

**Table 5** Results of Wilcoxon test between PSOPRS and DEFERS through 5NN and DT classifiers

Dataset	$\beta = 1$	5NN			DT	
		$\beta = 0.9$	$\beta = 0.5$	$\beta = 1$	$\beta = 0.9$	$\beta = 0.5$
Lymph	+	+	+	+	+	+
Spect	-	-	-	+	-	+
Dermatology	+	+	+	+	+	+
Soybean	+	+	+	+	+	+
Chess	=	=	+	+	+	+
Waveform	+	+	+	-	+	=

most cases. Especially in the Lymph, Dermatology and Soybean datasets, the performance of DEFERS is hugely superior to that of PSOPRS. For instance, the average classification accuracy obtained by DEFERS over 5NN in the Dermatology dataset is 94.28%, whereas PSOPRS obtains only 78.17% in the same dataset. Although DEFERS sometimes selects the feature subsets with high dimensionality compared to PSOPRS, this condition may not be considered as a big deal due to the significant classification performance of the feature subsets selected by DEFERS as seen in Table 5. Only in the Spect dataset, DEFERS generally gets worse classification performance than PSOPRS. In summary, the FRS-based filter approach using DE outperforms the PRS-based filter approach using PSO in terms of selecting the most appropriate feature subsets yielding

significantly better classification performance through both the 5NN and DT classifiers.

### 5.2 Comparison between MDEFERS and MPSOPRS

In contrast to PSOPRS and DEFERS, multi-objective approaches (MDEFERS and MPSOPRS) get a set of Pareto-optimal solutions in each independent run. Accordingly, there exist the 30 Pareto-optimal sets over 30 independent runs for each multi-objective approach. In order to analyze and compare multi-objective filter approaches, we use the following metrics [21]:

- Best front: The best front is determined by using the crowding distance metric among the 30 Pareto-optimal sets. The solution sets are first sorted according to the

classification accuracies in ascending order, and then the crowding distance value of a solution is calculated by averaging the distance of its two neighbors. For the solutions with the lowest and highest accuracies which are treated as boundaries, an infinite value is assigned as a crowding distance value. It should be notified that as the classification performance may vary according to the predefined classifier even for the same solution set, different best fronts may be generated for different classifiers.

- Average front: The average front is determined by calculating the average classification accuracy of the 30 Pareto-optimal solution sets having the same number of features. Notice that if there exists only one solution for any number of features, the corresponding solution is also added to the average front.

The results of MDEFRS and MPSOPRS are presented in terms of the best (denoted as ‘-Best’) and the average (denoted as ‘-Avg’) fronts in Figs. 1 and 2 through 5NN and DT, respectively. In charts, the best front is denoted as ‘-Best,’ while the average front is denoted as ‘-Avg.’ In terms of the best front, MDEFRS generally performs better than MPSOPRS in four out of six datasets through both 5NN and DT. For the datasets where MDEFRS sometimes cannot achieve the highest classification accuracy, the space between the lines representing the best fronts is illustrated to be very close. In terms of the average front, MDEFRS outperforms MPSOPRS in all datasets except for the Chess dataset. It should be notified that the space between the lines representing the average fronts is very large. For instance, MDEFRS achieves higher average classification accuracy than MPSOPRS by 6% for the solutions including 11 features. Further, MPSOPRS can

only obtain nearly 85% accuracy at most in the Dermatology dataset in terms of the average front, while there exist lots of solutions obtained by MDEFRS achieving more than 90% classification accuracy in the same dataset. It can therefore be suggested that the proposed MDEFRS approach based on MODE and fuzzy rough set theory produces more consistent and reasonable results without no doubt, i.e., the proposed MDEFRS approach is stable and robust compared to MPSOPRS.

### 5.3 Comparison between DEFERS and MDEFRS

The results of FRS-based approaches are presented through the 5NN and DT classifiers in Fig. 3, where the  $\beta$  value in DEFERS is set to 0.9, and ‘-Avg’ and ‘-Best,’ respectively, denote the average and best fronts, the details of which are presented in Sect. 5.2. Notice that there may be illustrated in charts fewer than 30 distinct points for DEFERS over 30 independent runs since DEFERS can find the same feature subsets in different runs.

According to Fig. 3, MDEFRS has shown to achieve best classification performance in all datasets except for the Chess dataset through both 5NN and DT despite the dimensionality reduction in higher rates. For instance, MDEFRS can obtain nearly 92% accuracy only using 9 features through DT, while DEFERS can generally get between 78 and 82% accuracy in spite of using more than 9 features. Not only in terms of the best fronts but also in terms of the average fronts, a remarkable difference can be observed between MDEFRS and DEFERS. For instance, DEFERS can find 81.25% accuracy through 5NN using 7 features in the Spect dataset, MDEFRS can find the same accuracy using less than 7 features, even using only one feature. For another example, DEFERS can obtain between

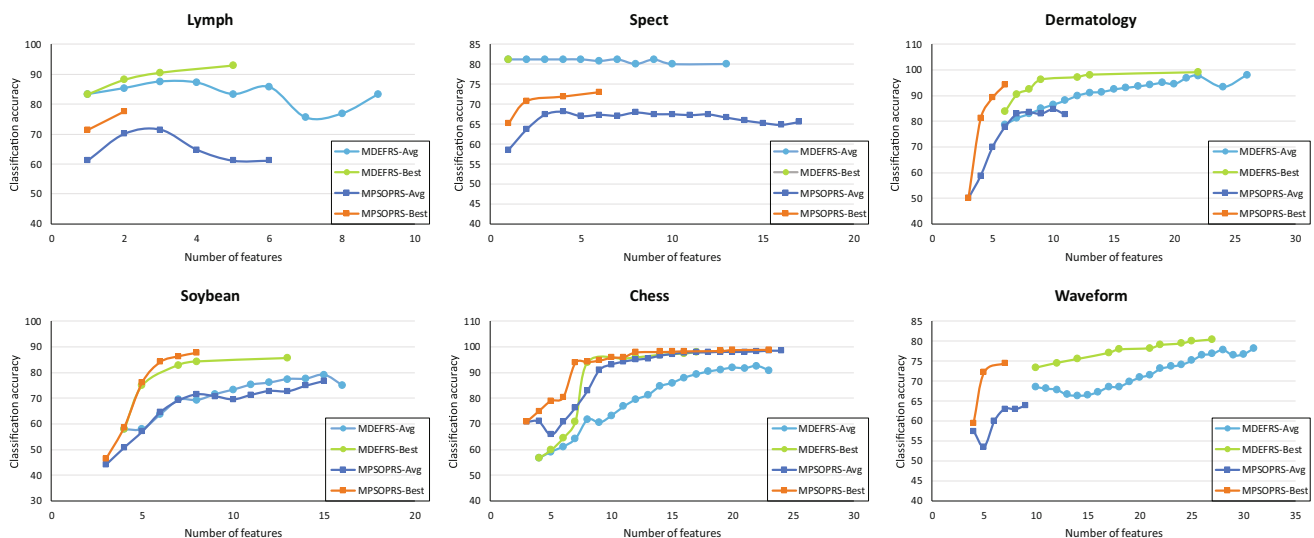


Fig. 1 Results of MPSOPRS and MDEFRS over 5NN classifier

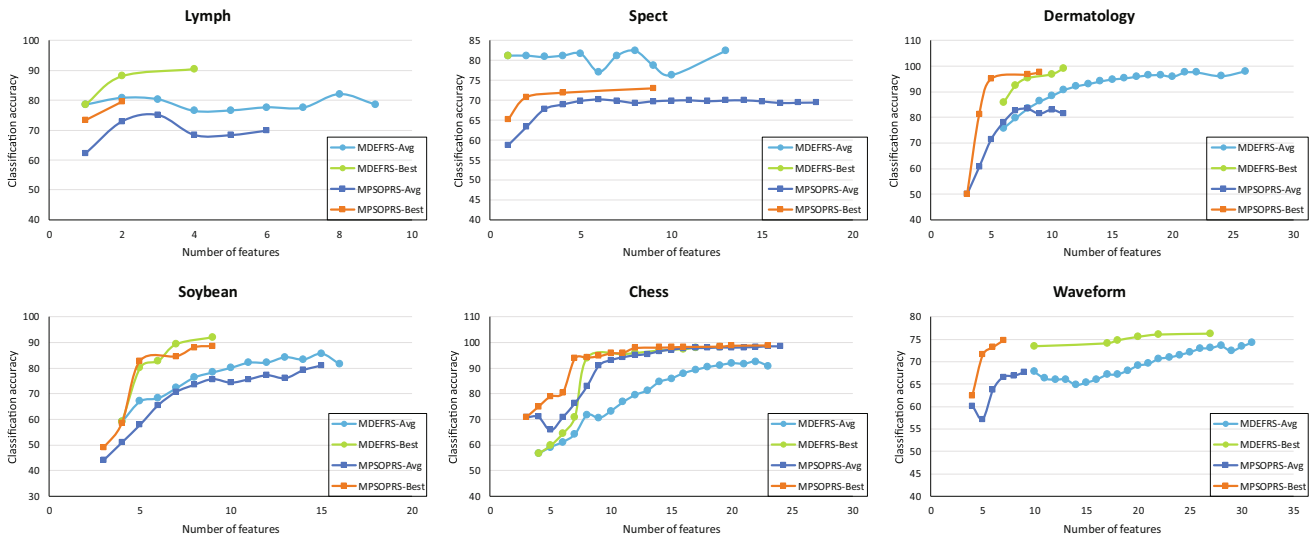


Fig. 2 Results of MPSOPRS and MDEFRS over DT classifier

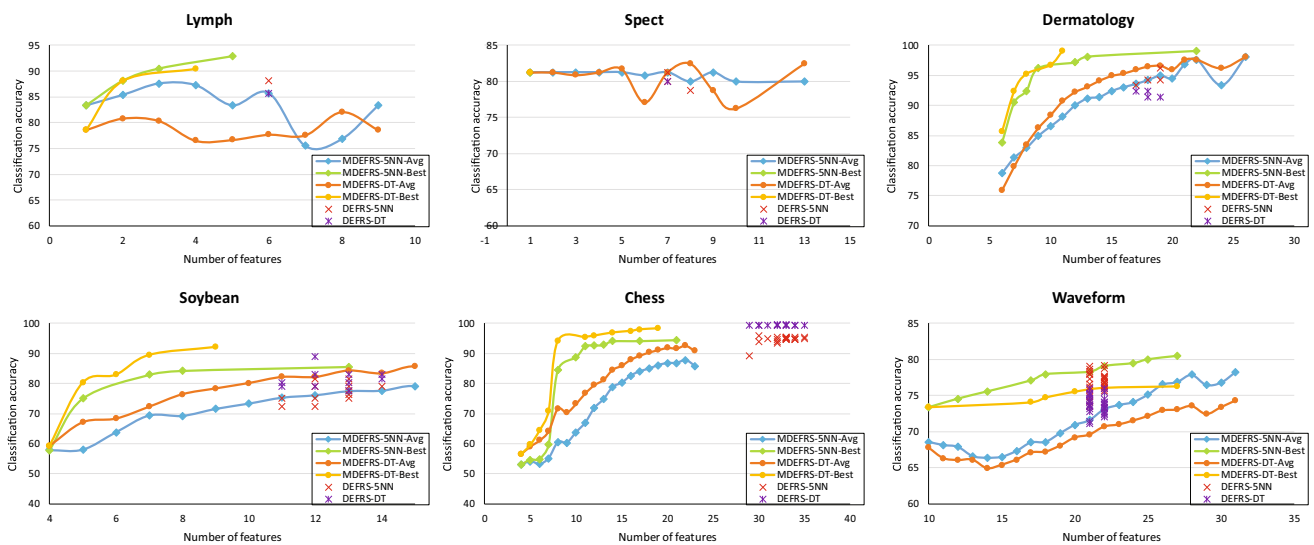


Fig. 3 Results of DEFRS and MDEFRS over 5NN and DT classifiers

93 and 96% accuracy using 17–19 features through 5NN in the Dermatology dataset, MDEFRS obtains similar performance through 5NN by eliminating more than half of the features from the Dermatology dataset. In summary, adopting the principles of fuzzy rough set theory in multi-objective framework can achieve outstanding performance compared to in single-objective design.

### 5.4 Comparison with the state-of-the-art approaches

The results of DEFRS and the state-of-the-art approaches are, respectively, presented on fourteen real-valued benchmark datasets through 5NN and SVM in terms of the average classification accuracy and the Wilcoxon rank-sum

test in Tables 6 and 7, where the best values are highlighted by bold symbol, and  $\beta$  is specified as 0.9 like as Sect. 5.3. Except for DEFRS, all approaches produce a unique feature subset and so have a single classification accuracy. Notice that we could not provide information concerning the number of features obtained by each approach since this was not reported in [50]; therefore, we could not make comparisons through multi-objective framework. In Tables 6 and 7, ‘+’ (‘–’) denotes that DEFRS produces significantly better (worse) classification accuracies than the corresponding filter approach. If there exists no such a significant difference between the proposed and the corresponding approaches, this case is denoted via ‘=.’ Besides the Wilcoxon rank-sum test, we use the Kruskal–Wallis test to illustrate the difference

**Table 6** Results of DEFERS and state-of-the-art approaches through 5NN classifier

Dataset	DEFERS	OFS	Fisher	PCC	ReliefF	MI	Laplacian	INF	LLC-FS	FSV
Ionosphere	88.79 ± 3.38	<b>89.43</b>	88	87.43	82.29	79.43	83.43	81.71	85.43	79.43
		–	+	+	+	+	+	+	+	+
WBCD	93.74 ± 2.19	<b>93.85</b>	93.32	93.32	89.62	90.16	92.44	93.67	90.86	90.16
		=	=	=	+	+	+	=	+	+
Sonar	<b>73 ± 9.05</b>	68.27	71.08	71.08	67.79	54.79	59.55	69.55	66.84	54.79
		+	=	=	+	+	+	+	+	+
Hill	<b>59.5 ± 3</b>	<b>59.5</b>	53.14	53.14	55.7	55.45	49.67	50.33	54.79	55.45
		=	+	+	+	+	+	+	+	+
Colon	<b>98.88 ± 4.22</b>	75	73.33	73.33	71.67	51.67	51.67	56.67	61.67	51.67
		+	+	+	+	+	+	+	+	+
Srbct	<b>97.91 ± 4.73</b>	88.33	88.33	58.33	86.67	63.33	31.67	26.67	38.33	63.33
		+	+	+	+	+	+	+	+	+
Lung2	<b>99.83 ± 0.91</b>	93	79	81.5	80	74.5	76.5	81	79	83.5
		+	+	+	+	+	+	+	+	+
Lymphoma	<b>100 ± 0</b>	<b>100</b>	98.33	95	96.67	78.33	93.33	60	95	78.33
		=	+	+	+	+	+	+	+	+
Glioma	<b>83.33 ± 5.6e-16</b>	68	64	66	26	56	42	40	48	62
		+	+	+	+	+	+	+	+	+
MLL	<b>99.33 ± 2.03</b>	92.86	90	78.57	92.86	64.29	84.29	91.43	87.14	82.86
		+	+	+	+	+	+	+	+	+
Prostate	80.66 ± 7.39	<b>93</b>	91	91	91	63	59	48	72	63
		–	–	–	–	+	+	+	+	+
DLBCL	90.76 ± 4.23	<b>95</b>	79.25	44.25	81.25	60.5	68.5	73.75	76.25	78.75
		–	+	+	+	+	+	+	+	+
Leukemia	<b>99.04 ± 3.62</b>	95.71	87.14	87.14	84.29	60	58.57	47.14	47.14	60
		+	+	+	+	+	+	+	+	+
Arcene	<b>90.31 ± 0.86</b>	86	65.5	65.5	67.5	58.5	70	71.5	65.5	69
		+	+	+	+	+	+	+	+	+
Avg. accuracy	<b>89.64</b>	85.56	80.10	74.39	76.66	64.99	65.75	63.67	69.13	69.44
Avg. ranks	<b>1.28</b>	1.92	4.85	4.92	6.21	7.5	6.5	7.14	7.35	7.28

between feature selection approaches on the classification accuracy through 5NN and SVM. The box plots obtained by the Kruskal–Wallis test for each dataset among all approaches are presented in Fig. 4.

In order to evaluate the general distribution of the classification performance obtained by all feature selection approaches on each real-valued dataset using the Kruskal–Wallis test, we observe from Fig. 4 that feature selection approaches produce a variety of feature subsets for each dataset resulting in a wide range of classification accuracies through 5NN and SVM in all datasets except for the WBCD and Hill datasets. Especially in the Srbct, Lymphoma, Prostate and Leukemia datasets, there exists a huge difference among approaches.

According to Tables 6 and 7, it can be observed that DEFERS outperforms MI, Laplacian, LLC-FS and FSV

through both 5NN and SVM in all datasets except for only one case. When compared with INF, DEFERS also achieves significantly great classification performance in all datasets except for the WBCD dataset. In other words, the difference between DEFERS and these approaches (MI, Laplacian, LLC-FS and FSV) is exceptionally high. For instance, DEFERS obtains nearly 99% classification accuracy in the Leukemia dataset, while these approaches receive the classification accuracy between 47 and 64%. Similar high differences between DEFERS and these approaches can also be observed in other datasets. It can therefore be indicated that the classification performance of such approaches cannot be treated as adequate on real-valued datasets. When compared with the remaining approaches, DEFERS generally performs slightly better or significantly superior performance in ten out of fourteen datasets through both

**Table 7** Results of DEFERS and state-of-the-art approaches through SVM classifier

Dataset	DEFERS	OFS	Fisher	PCC	ReliefF	MI	Laplacian	INF	LLC-FS	FSV
Ionosphere	<b>87.03</b>	81.43	85.71	83.43	69.14	84	69.71	66.86	77.71	84
		+	+	+	+	+	+	+	+	+
WBCD	94.09 ± 1.56	<b>96.14</b>	95.26	95.26	92.09	92.44	92.97	95.26	94.37	92.44
		–	–	–	+	+	+	–	=	+
Sonar	69 ± 5.78	<b>75.04</b>	73.61	73.61	70.08	56.74	61.65	64.56	64.76	56.74
		–	–	–	–	+	+	+	+	+
Hill	<b>78.16 ± 9.91</b>	53.39	50.74	50.74	52.07	50.74	50.99	50.58	51.49	50.74
		+	+	+	+	+	+	+	+	+
Colon	83.88 ± 6.89	81.67	85	85	<b>86.67</b>	65	70	68.33	71.67	65
		+	–	–	–	+	+	+	+	+
Srbct	<b>100 ± 0</b>	80	88.33	66.67	75	68.33	38.33	36.67	30	68.33
		+	+	+	+	+	+	+	+	+
Lung2	<b>95.33 ± 1.26</b>	93.5	84.5	84	85	85	82	86.5	86	85
		+	+	+	+	+	+	+	+	+
Lymphoma	<b>100 ± 0</b>	98.33	93.33	90	91.67	68.33	90	65	91.67	68.33
		+	+	+	+	+	+	+	+	+
Glioma	<b>83.33 ± 5.6e-16</b>	60	58	60	28	60	46	48	42	48
		+	+	+	+	+	+	+	+	+
MLL	<b>100 ± 0</b>	92.86	94.29	90	94.29	70	91.43	94.29	88.57	90
		+	+	+	+	+	+	+	+	+
Prostate	88.66 ± 5.07	<b>94</b>	92	92	<b>94</b>	59	60	47	73	59
		–	–	–	–	+	+	+	+	+
DLBCL	<b>100 ± 0</b>	97.5	86.25	78.75	82.5	69.25	71	68.5	76.25	80
		+	+	+	+	+	+	+	+	+
Leukemia	<b>98.57 ± 4.35</b>	94.29	90	90	88.57	61.43	64.29	58.57	60	61.43
		+	+	+	+	+	+	+	+	+
Arcene	<b>89.84 ± 5.55</b>	80.5	72.5	62.5	65	57.5	63.5	67	66	72.5
		+	+	+	+	+	+	+	+	+
Avg. accuracy	<b>90.56</b>	84.18	82.10	78.71	76.72	67.69	67.99	65.50	69.53	70.10
Avg. ranks	<b>1.5</b>	2.64	4.07	6.35	5.14	6.28	6.92	6.92	7.57	7.57

5NN and SVM. In summary, the proposed DEFERS approach can achieve satisfactory classification performance not only in discrete-valued but also in real-valued datasets.

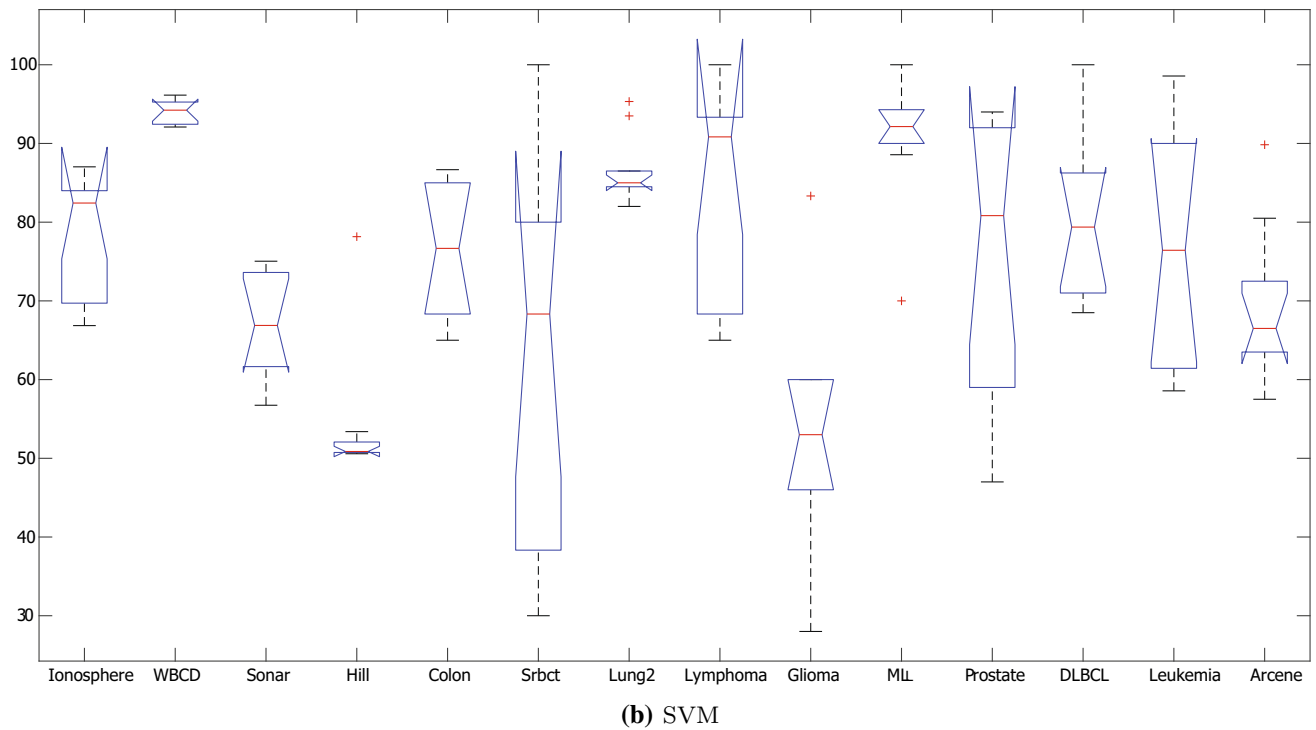
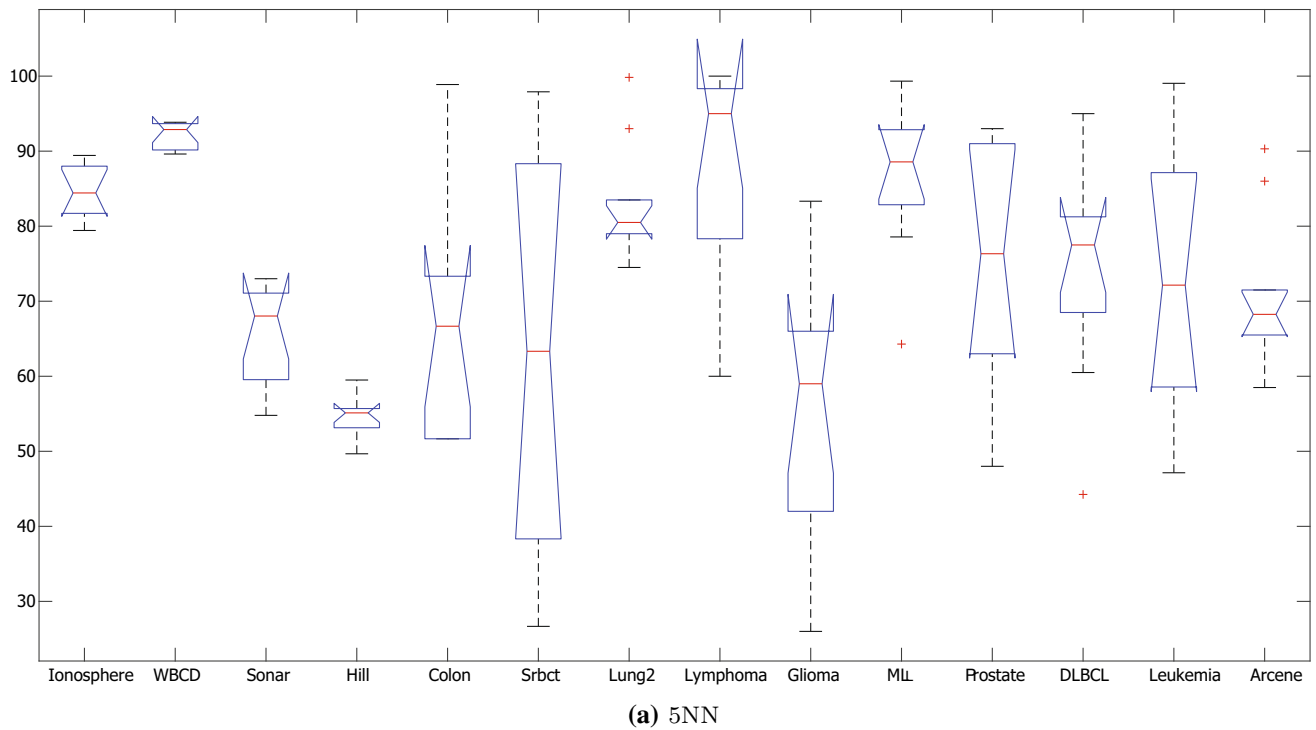
When evaluating approaches in terms of the computational cost, the state-of-the-art approaches generally complete the selection process in a shorter time than DEFERS. This is because most of them are based on deterministic procedures and follow greedy search mechanism to select or rank features for the feature subset rather than evaluating feature combinations. However, due to the dependency of deterministic procedures and greedy search mechanism, these approaches cannot deeply evaluate the possible solution space and so frequently encounter with local-convergence problems. This case can be also observed in terms of the classification performance in Tables 6 and 7.

Furthermore, a GPU-paralleled implementation of DEFERS can be applied to improve the efficiency of DEFERS.

## 6 Conclusions

In this paper, we aim to propose new filter approaches for feature selection that can be used not only in discrete-valued problems but also in real-valued problems. This aim was achieved by designing filter criteria using fuzzy rough set theory and then integrating the criteria in single-objective and multi-objective DE frameworks. While DEFERS tries to optimize the dependency of the feature subset and the feature subset size using a control parameter, MDEFERS optimizes the same objectives without the requirement of a control parameter. The performance of the proposed





**Fig. 4** Box plots obtained by Kruskal–Wallis test

approaches is verified by comparing them with existing approaches on discrete-valued and real-valued datasets. Notice that in order to prove the effectiveness of the proposed approaches, we directly utilize the results of existing

works [42, 50] in the literature. According to the results on discrete-valued datasets, the proposed approaches achieved a much higher classification performance than the existing PSO-based approaches based on probabilistic rough set

theory. Particularly, in many cases, the proposed single-objective approach achieved more than 10% better classification performance than the existing single-objective PSO-based approach. The similar remarkable difference can also be illustrated between the proposed and the existing multi-objective approaches. According to the results on real-valued datasets, the proposed single-objective approach outperforms a variety of conventional feature selection approaches. In particular, DEFERS provided more than 4% better performance than the recently introduced approach (called OFS) which has second best performance in terms of the mean classification accuracy obtained by averaging the classification accuracies in all datasets. Among the proposed approaches, which one is better? The results reveal that considering fuzzy rough set theory in multi-objective design can receive better feature subsets yielding higher classification accuracy and smaller feature subset size.

The proposed approaches based on fuzzy rough set theory sometimes tend to select a large number of features compared to existing approaches. This may increase the computational complexity in feature selection tasks. This issue will be the motivation for our future work.

## Compliance with ethical standards

**Conflict of interest** The author declares that he has no conflict of interest. In addition, this article does not contain any studies with human participants or animals performed by the author. The undersigned author declares that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

## References

- Abdul-Rahman S, Mohamed-Hussein Z, Bakar AA (2010) Integrating rough set theory and particle swarm optimisation in feature selection. In: 10th international conference on intelligent systems design and applications, pp 1009–1014
- Aziz MAE, Hassanien AE (2018) Modified cuckoo search algorithm with rough sets for feature selection. *Neural Comput Appl* 29(4):925–934
- Bjorvand AT, Komorowski J (1997) Finding minimal reducts using genetic algorithms. In: 15th IMACS World Congress on scientific computation, modelling and applied mathematics, pp 601–606
- Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: Proceedings of the fifteenth international conference on machine learning, ICML'98, pp 82–90
- Cervante L, Xue B, Shang L, Zhang M (2012) A dimension reduction approach to classification based on particle swarm optimisation and rough set theory. In: Thielscher M, Zhang D (eds) *Advances in Artificial Intelligence AI2012*. Springer, Berlin Heidelberg, pp 313–325
- Chebrolu S, Sanjeevi SG (2017) Attribute reduction on real-valued data in rough set theory using hybrid artificial bee colony: extended FTSBPSD algorithm. *Soft Comput* 21(24):7543–7569
- Chouchoulas A, Shen Q (2001) Rough set-aided keyword reduction for text categorization. *Appl Artif Intell* 15(9):843–873
- Das AK, Sengupta S, Bhattacharyya S (2018) A group incremental feature selection for classification using rough set theory based genetic algorithm. *Appl Soft Comput* 65:400–411
- Das S, Chang CC, Das AK, Ghosh A (2017) Feature selection based on bi-objective differential evolution. *J Comput Sci Eng* 11(4):130–141
- Dash M, Liu H, Motoda H (2000) Consistency based feature selection. In: Proceedings of the fourth Pacific Asia conference on knowledge discovery and data mining, pp 98–109
- De Silva AM, Leong PHW (2015) Feature selection. Springer, Singapore, pp 13–24
- Dua D, Graff C (2017) UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Gawar A (2014) Performance analysis of quickreduct, quick relative reduct algorithm and a new proposed algorithm. *Glob J Comput Sci Technol C Softw Data Eng* 14(4):1–5
- Gu Q, Li Z, Han J (2011) Generalized fisher score for feature selection. In: Proceedings of the twenty-seventh conference on uncertainty in artificial intelligence, UAI'11, pp. 266–273. AUAI Press, Arlington, VA, USA
- Hall MA (1999) Correlation-based feature selection for machine learning. Tech. rep., University of Waikato
- Hancer E (2018) A multi-objective differential evolution feature selection approach with a combined filter criterion. In: 2nd international symposium on multidisciplinary studies and innovative technologies (ISMSIT2018), pp 1–8
- Hancer E (2019) Differential evolution based multiple kernel fuzzy clustering. *J Fac Eng Archit Gazi Univ* 34(3):1282–1293
- Hancer E (2019) Fuzzy kernel feature selection with multi-objective differential evolution algorithm. *Connect Sci* 31(4):1–19. <https://doi.org/10.1080/09540091.2019.1639624>
- Hancer E (2020) A new multi-objective differential evolution approach for simultaneous clustering and feature selection. *Eng Appl Artif Intell* 87:103307. <https://doi.org/10.1016/j.engappai.2019.103307>
- Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. *Knowl-Based Syst* 140:103–119
- Hancer E, Xue B, Zhang M, Karaboga D, Akay B (2018) Pareto front feature selection based on artificial bee colony optimization. *Inf Sci* 422:462–479
- Holland JH (1984) Genetic algorithms and adaptation. Springer US, Berlin, pp 317–333
- Jensen R, Shen Q (2009) New approaches to fuzzy-rough feature selection. *IEEE Trans Fuzzy Syst* 17(4):824–838
- Jing SY (2014) A hybrid genetic algorithm for feature subset selection in rough set theory. *Soft Comput* 18(7):1373–1382
- Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of international conference on neural networks (ICNN'95), pp 1942–1948
- Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: Bergadano F, De Raedt L (eds) *Machine learning: ECML-94*. Springer, Berlin Heidelberg, pp 171–182
- Lampinen J (2001) Solving problems subject to multiple non-linear constraints by differential evolution. In: 7th international conference on soft computing, pp 50–57
- Mafarja MM, Mirjalili S (2018) Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. *Soft Comput* 23:6249–6265. <https://doi.org/10.1007/s00500-018-3282-y>

29. Marinaki M, Marinakis Y (2014) An island memetic differential evolution algorithm for the feature selection problem, pp 29–42
30. Marler R, Arora J (2004) Survey of multi-objective optimization methods for engineering. *Struct Multidiscip Optim* 26(6):369–395
31. Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11(5):341–356
32. Prasad PSVSS, Rao CR (2009) IQuickReduct: an improvement to Quick Reduct algorithm. In: Sakai H, Chakraborty MK, Hassanien AE, Slezak D, Zhu W (eds) *Rough sets, fuzzy sets, data mining and granular computing*. Springer, Berlin Heidelberg, pp 152–159
33. Price K, Storn RM, Lampinen JA (2005) *Differential evolution: a practical approach to global optimization (natural computing series)*. Springer, Berlin
34. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* 53(1):23–69
35. Sangeetha R, Kalpana B (2013) Enhanced fuzzy roughset based feature selection strategy using differential evolution. *Int J Comput Sci Appl (TIJCSA)* 2(06):13–20
36. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF (2019) A review of unsupervised feature selection methods. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-019-09682-y>
37. Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 11(4):341–359
38. Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. *Neural Comput Appl* 24(1):175–186
39. Wang X, Yang J, Teng X, Xia W, Jensen R (2007) Feature selection based on rough sets and particle swarm optimization. *Pattern Recognit Lett* 28(4):459–471
40. Wasikowski M, Chen X (2010) Combating the small sample class imbalance problem using feature selection. *IEEE Trans Knowl Data Eng* 22(10):1388–1400
41. Wroblewski J (1995) Finding minimal reducts using genetic algorithms. In: *Proceedings of second annual joint conference on information sciences*, pp 186–189
42. Xue B, Cervante L, Shang L, Browne WN, Zhang M (2014) Binary PSO and rough set theory for feature selection: a multi-objective filter based approach. *Int J Comput Intell Appl* 13(02):1450009
43. Yan H, Li X (2010) A novel attribute reduction algorithm based improved differential evolution. In: *Second WRI Global Congress on Intelligent Systems*, vol 3, pp 87–90
44. Yang K, Cai Z, Li J, Lin G (2006) A stable gene selection in microarray data analysis. *BMC Bioinform* 7(1):228
45. Yao Y, Zhao Y (2008) Attribute reduction in decision-theoretic rough set models. *Inf Sci* 178(17):3356–3373
46. Yong L, Wenliang H, Yunliang J, Zhiyong Z (2014) Quick attribute reduct algorithm for neighborhood rough set model. *Inf Sci* 271:65–81
47. Yu L, Ding C, Loscalzo S (2008) Stable feature selection via dense feature groups. In: *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'08*. ACM, New York, pp 803–811
48. Zeng H, Cheung Y (2011) Feature selection and kernel learning for local learning-based clustering. *IEEE Trans Pattern Anal Mach Intell* 33(8):1532–1547
49. Zhang Q, Xie Q, Wang G (2016) A survey on rough set theory and its applications. *CAAI Trans Intell Technol* 1(4):323–333
50. Zhou P, Hu X, Li P, Wu X (2019) OFS-density: a novel online streaming feature selection method. *Pattern Recognit* 86:48–61

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.