Knowledge-Based Systems 151 (2018) 216–229

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A fuzzy rough set-based feature selection method using representative instances

Xiao Zhang[a,*], Changlin Mei[b], Degang Chen[c], Yanyan Yang[d]

[a] Department of Applied Mathematics, School of Sciences, Xi'an University of Technology, Xi'an, PR China
[b] Department of Statistics, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, PR China
[c] Department of Mathematics and Physics, North China Electric Power University, Beijing, PR China
[d] Department of Automation, Tsinghua University, Beijing, PR China

ARTICLE INFO

Article history:
Received 29 September 2017
Revised 22 March 2018
Accepted 23 March 2018
Available online 27 March 2018

Keywords:
Fuzzy rough set
Representative instance
Feature selection

ABSTRACT

The fuzzy rough set theory has been widely used to deal with uncertainty in real-valued or even complex data, in which one of the most concerned issues is feature selection. Since a real-world data set generally contains redundant data objects (or instances) and errors which lead to the fact that not all the instances are of equal importance, focusing on the representative instances for feature selection can not only acquire more convincing analysis results but also alleviate computational complexity in mining knowledge. At present, however, little attention has been paid on using representative instances to select features. In this paper, the issue of selecting features by using representative instances is investigated based on fuzzy rough sets and a representative instance-based feature selection approach is proposed. First, the fuzzy granular rule is employed to describe the discriminating information of an instance. Then, the representative instances are selected according to the coverage ability of the fuzzy granular rules induced by all of the instances. Furthermore, an implication relationship-preserved reduction is presented to maintain the discriminating information of the selected instances, and then a heuristic algorithm is presented to search for such a feature subset. Finally, a filter-wrapper approach is suggested to select the best subset of the features. Some numerical experiments are further conducted to show the performance of the proposed feature selection method and the results are satisfactory in terms of both efficiency and effectiveness.

1. Introduction

Since the data collected in practice are not purely for some particular application, there always exist the redundant instances, the irrelevant instances and errors during the data collection and recordation. Additionally, it is difficult to mine critical information from the massive data. An effective avenue to deal with such cases is instance selection which chooses a subset of data to achieve the original purpose of a data mining application as if the whole data is used [30].

Via instance selection, one can usually remove redundant data and errors from original data, which makes the mining algorithms only focus on the remaining instances and easily acquire high quality results. More importantly, instance selection can alleviate the time cost of algorithms, and can save much more time especially for the large-scale data. The issue of instance selection is employed in many applications of data mining, such as the condensed nearest neighbor rule [14] (CNN rule) and data reduction by SVM [43] in the classification learning, and the scalable K-means algorithm [3] and CURE [13] in the cluster learning, as well as the sampling-based frequent itemset generation algorithm [47] in the association rule mining.

Feature selection, as another technique for data reduction, has been successfully rendered in both dimensionality reduction and learning performance improvement, and has been provided many relevant algorithms with different search strategies and evaluation functions. Combination between instance selection and feature selection generates the active feature selection [31] which uses partial selective instances rather than the whole instances for the feature selection algorithm to achieve time savings without performance deterioration, and the active feature selection is factually the representative instance-based feature selection.

Rough sets [37] presented by Pawlak in 1982 is an useful tool to deal with the uncertainty in data. One of the main applications

* Corresponding author.
E-mail addresses: zhangxiao@xaut.edu.cn (X. Zhang), clmei@mail.xjtu.edu.cn (C. Mei), wscz973@163.com (D. Chen), hrsinbba@163.com (Y. Yang).

https://doi.org/10.1016/j.knosys.2018.03.031

of rough sets in practice is feature selection (often called attribute reduction). There exist many feature selection approaches based on rough sets [7,9,27–29,33,39,41,54,60,61]. However, less attention was paid on the representative instance-based feature selection. To the best of our knowledge, Chen et al. [5] presented the sample pair selection approach based on rough sets to compress the discernibility function for searching for all the reducts. However, computing one sample pair selection is time-consuming, and in general, only one reduct is needed for practical requirements. Qian et al. [40] proposed the positive approximation-based forward fast feature selection algorithm with traditional rough sets. The critical acceleration of the algorithm is only selecting the instances out of the positive region in each iteration search procedure, which has been empirically shown the significant reduction of the computational time. Nevertheless, the representative instance-based feature selection performs instance selection in advance of the feature selection algorithm. Based on the traditional rough set theory, Dai et al. [9] proposed two quick feature selection algorithms which only consider the equivalence classes containing at least one neighbor inconsistent pair, i.e., the representative objects, and then alleviate the computational time of finding a reduct. Yang et al. [58] selected the critical instances in dynamic environment to achieve the incremental computation of attribute reduction. Additionally, Zhang et al. [62] proposed the confidence-preserved attribute reduction which only considers the instances that induce rules with confidence being not less than a given threshold.

The traditional rough sets can only process the nominal data, and one kind of important extended rough sets is the fuzzy rough sets which can deal with the real-valued or even the mixed data. The fuzzy rough sets have been successfully applied in the fields of both data mining and machine learning [51], in which feature selection is one of the most important issues. The existing attribute reductions may mainly be concluded into the fuzzy positive region-preserved reduction [48] and the fuzzy dependency function-based reduction [17,23–25] as well as the fuzzy information entropy-based reduction [16,63]. In fact, a fuzzy dependency function-based reduct is also a fuzzy positive region-preserved reduct if only the fuzzy lower approximation in the fuzzy dependency function is the same to that in the fuzzy positive region, and the information entropy provided in [63] can equivalently characterize the fuzzy positive region-preserved reduction or the fuzzy dependency function-based reduction. The other attribute reductions of fuzzy rough sets mainly focus on the improvement of the fuzzy lower approximations in the fuzzy dependency function or the fuzzy positive region [4,18–20,52,53,59,64].

It is worth noting that the fuzzy positive region-preserved reduction preserves the information of all of the instances, which may lead to the fact that the reduction algorithms select more features and consume more computational time. However, the data collected in practice usually include redundant instances and errors. Then, it is necessary to consider the issue of instance selection. A preliminary work of instance selection based on fuzzy rough sets can be found in [22] which presented the fuzzy rough instance selection technique (FRIS) that only selects the instances with the membership to the fuzzy positive region being not less than a pre-specified threshold. Based on fuzzy rough sets, Verbiest et al. [50] gave a new instance selection method called fuzzy rough prototype selection (FRPS) which uses a fuzzy rough measure to characterize the quality of the instances and provides a wrapper approach to determine the selected instances, and Tsang et al. [49] designed a weighted sampling technique to select the representative instances for K-nearest neighbor rule (KNN rule).

Besides the research of the fuzzy rough set-based instance selection, some work focuses on the simultaneous feature and instance selection [1,2,11,15,36,44–46]. For example, He et al. [15] presented a neighborhood-based rough set model to select the boundary instances for the classification learning of the Support Vector Machines, in which feature and instance selection is simultaneously conducted. The concept of a bireduct was put forward in [44], which is an extension of the notion of a reduct of rough sets. In fact, a bireduct includes both some instances and attributes, and seems to be a class of classification rules. Thus, the number of all the bireducts is greatly more than that of the reducts. Some analogies and relationships between the decision bireducts and the approximate decision reducts were shown in [45,46]. As the extension of the work in [44], simultaneous feature and instance selection using fuzzy rough sets was investigated in [36], and a corresponding algorithm with a frequency-based approach taking as a heuristic was designed to select features or instances alternatively. Via the Shuffled Frog Leaping Algorithm, Anaraki et al. [1] proposed a novel simultaneous feature and instance selection approach based on fuzzy rough sets. Moreover, Derrac et al. [10] presented such a steady-state genetic algorithm to select instances that is added in a fuzzy rough set-based feature selection process.

Although there have been a lot of research work about feature selection and instance selection as well as simultaneous feature and instance selection based on fuzzy rough sets, little attention has been paid on using representative instances to select features. Only using the representative instances obtained from a given data set, one may alleviate computational time and acquire more convincing analysis results. Therefore, it is meaningful to research the issue of fuzzy rough set-based feature selection by using representative instances.

In this paper, we present a fuzzy rough set-based instance selection approach and a feature selection method by using representative instances, respectively. The concept of a fuzzy granular rule is proposed to describe the discriminating information of an instance. According to the coverage ability of fuzzy granular rules, a representative instance selection method is suggested. The definition of rule implication relationship is given. Then, the instance can preserve the discriminating information of itself if the rule implication relationship holds by removing some features. Afterwards, we put forward an implication relationship-preserved attribute reduction which can maintain the decision discriminating information of all the representative instances, and present a heuristic algorithm for selecting such a feature subset. Furthermore, a filter-wrapper approach is formulated to obtain a best feature subset that includes fewer features and achieves higher classification quality. Finally, the performance of the filter-wrapper feature selection algorithm is empirically evaluated, in which the computational time of searching for a best feature subset, the cardinality of the feature subset and the classification accuracy achieved by the feature subset are compared with other feature selection algorithms.

The remainder of this paper is organized as follows. We briefly review in Section 2 some basic knowledge about fuzzy rough sets in order to facilitate the subsequent discussions. In Section 3, the concepts of a fuzzy granular rule and a minimal fuzzy granular rule set are presented. Then, a representative instance selection approach is suggested and the implication relationship between the fuzzy granular rules is investigated. In Section 4, the implication relationship-preserved attribute reduction is formulated, and a filter-wrapper algorithm is proposed to search for a best feature subset. Some numerical experiments are conducted to show the performance of the proposed filter-wrapper feature selection algorithm in Section 5.

2. Preliminaries

In order to facilitate the subsequent discussions, we first introduce the relevant knowledge and definitions. A data set can often be viewed as a collection of data objects (or instances), and objects are described by a number of attributes (or features) that capture the basic characteristics of objects.

2.1. Fuzzy rough sets

Let $U=\{x_1,x_2,\ldots,x_n\}$ be a nonempty universe of discourse where x_i $(i=1,2,\ldots,n)$ is the object (instance), and $F(U\times U)$ be the fuzzy power set on $U\times U$. R is called a fuzzy relation on $U\times U$ if $R\in F(U\times U)$, where $R(x, y)$ measures the strength of relationship between $x\in U$ and $y\in U$.

Let R be a fuzzy relation on $U\times U$. R is reflexive if $R(x,x)=1$ for any $x\in U$; R is symmetric if $R(x,y)=R(y,x)$ for any $x, y\in U$; and R is T-transitive if $R(x, y)\geq T(R(x, z), R(z, y))$ for a triangular norm T and any $x, y, z\in U$. Furthermore, R is called a T-similarity relation if R is reflexive, symmetric and T-transitive. Specially, if $T=\min$, R is called a fuzzy equivalence relation.

In the pioneering work [12], a pair of lower and upper approximation operators of a fuzzy set X based on a T-similarity relation R is defined, for each $x\in U$, as

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

and

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

to measure the degree of x certainly belonging to X and the degree of x possibly belonging to X, respectively, on which the fuzzy rough set of X is defined by $(\underline{R}X,\overline{R}X)$.

Since the existing research on attribute reduction is mainly based on the fuzzy rough sets in [12], we skip the reviews of other kinds of fuzzy rough sets and one can refer to [21,32,34,35,42,55–57] for the details of the contents. The work of this paper is also based on the fuzzy rough sets in [12].

2.2. Fuzzy information systems and fuzzy decision systems

A fuzzy information system is a pair (U, A) with a mapping $a_t:U\to V_{a_t}$ for each $a_t\in A$, where $U=\{x_1,x_2,\ldots,x_n\}$ is the universe of discourse, $A=\{a_1,a_2,\ldots,a_m\}$ is the attribute set on which a fuzzy relation $R_{\{a_t\}}$ is defined for each attribute $a_t\in A$, and V_{a_t} is the domain of a_t. The fuzzy relation of a subset $B\subseteq A$ is defined by $R_B=\bigcap_{a_t\in B}R_{\{a_t\}}$.

As indicated in [63], it is possible to define the corresponding fuzzy relations for the real-valued attributes. In fact, any monotonic decreasing function with respect to some distance measure can be used to define fuzzy relations for the real-valued attributes. Then, a data set with real-valued attributes can be treated as a fuzzy information system for further analysis.

A fuzzy decision system is a pair $(U, A\cup D)$ with $A\cap D=\emptyset$, where (U, A) is a fuzzy information system, A is called the conditional attribute set and $D=\{d\}$ is called the decision attribute set on which a mapping d: $U\to V_d$ is defined. Here, V_d is the domain of the decision attribute d with nominal values.

It is easily known that a fuzzy information system adding a decision (class or label) attribute is factually one fuzzy decision system. It should be noted that an equivalence relation R_D can be defined for the decision attribute d. R_D partitions the universe U into a family of disjoint subsets $U/R_D=\{D_k:k=1,2,\ldots,l\}$, where D_k is called the decision class. In general, U/R_D is written as U/D for notational simplicity. Given any object $x\in U$, there must exist a decision class D_k such that $x\in D_k$, and the membership function of the decision class D_k is

$$D_k(x)=\begin{cases}1, & x\in D_k,\\ 0, & x\notin D_k.\end{cases}$$

2.3. Attribute reduction in fuzzy decision systems

For a fuzzy decision system $(U, A\cup D)$ with $U=\{x_1,x_2,\ldots,x_n\}$ and $B\subseteq A$, the fuzzy positive region of D with respect to B is defined as

$$\mathrm{Pos}_B(D)=\bigcup_{D_k\in U/D}\underline{R_B}D_k,$$

where $\underline{R_B}D_k(x_i)=\inf_{x_j\in U}\max\{1-R_B(x_i,x_j),D_k(x_j)\}$. The dependency function of D relative to B is

$$\gamma_B(D)=\frac{\sum_{i=1}^{n}\mathrm{Pos}_B(D)(x_i)}{n}.$$

For a fuzzy decision system $(U, A\cup D)$, [16,17,23] and [48] presented the following definitions of a reduct of A relative to D, which we call here the dependency function-based reduct and fuzzy positive region-preserved reduct, respectively.

Definition 1 ([17]). Let $(U, A\cup D)$ be a fuzzy decision system with $U=\{x_1,x_2,\ldots,x_n\}$ and $B\subseteq A$. a is dispensable in B relative to D if $\gamma_{B-\{a\}}(D)=\gamma_B(D)$; otherwise, a is indispensable in B relative to D. $B\subseteq A$ is a dependency function-based reduct in $(U, A\cup D)$ if $\gamma_B(D)=\gamma_A(D)$ and $\gamma_{B-\{a\}}(D)<\gamma_B(D)$ for any $a\in B$.

Definition 2 ([48]). Let $(U, A\cup D)$ be a fuzzy decision system with $U=\{x_1,x_2,\ldots,x_n\}$ and $B\subseteq A$. B is a fuzzy positive region-preserved reduct in $(U, A\cup D)$ if $\mathrm{Pos}_B(D)=\mathrm{Pos}_A(D)$ and $\mathrm{Pos}_{B-\{a\}}(D)\neq\mathrm{Pos}_B(D)$ for any $a\in B$.

As clarified in [63], $B\subseteq A$ is a dependency function-based reduct if and only if B is a fuzzy positive region-preserved reduct in $(U, A\cup D)$.

3. Fuzzy granular rules and a representative instance selection approach for fuzzy decision systems

In this section, a new kind of fuzzy granular rule is given, and an instance selection approach is presented according to the coverage ability of fuzzy granular rules. Furthermore, the implication relationship between the fuzzy granular rules is investigated.

3.1. Fuzzy granular rules

It has been provided in [63] the equivalent characterization of the fuzzy lower approximation Eq.(1) from the viewpoint of granular structures, which is given by the following theorem.

Theorem 1 ([63]). *Let (U, A) be a fuzzy information system with a fuzzy relation R_B for each $B\subseteq A$. For any $X\in F(U)$, we have*

$$\underline{R_B}X(x)=\sup\{\lambda:[x]_B^\lambda\subseteq X,\lambda\in[0,1]\}, \tag{3}$$

where

$$[x]_B^\lambda(y)=\begin{cases}0, & 1-R_B(x,y)\geqslant\lambda;\\ \lambda, & 1-R_B(x,y)<\lambda.\end{cases} \tag{4}$$

It is known from Theorem 1 that the fuzzy granule $[x]_B^\lambda$ with $\lambda=\underline{R_B}X(x)$ is the biggest granule contained in X. Specially, let X be the decision class D_k. Then, the fuzzy granule $[x]_B^\lambda$ can be used to characterize the inner structure of D_k and thus describe the decision knowledge implied by the object x with respect to the attribute subset B, which motivates us to construct a granular rule by means of $[x]_B^\lambda$.

Definition 3. Let $(U, A\cup D)$ be a fuzzy decision system, $B\subseteq A$ and $U/D=\{D_k:k=1,2,\ldots,l\}$. Given an object $x\in D_k$, if $[x]_B^\lambda\subseteq D_k$ $(\lambda>0)$, we say that $[x]_B^\lambda\to D_k$ is a fuzzy granular rule induced by the object x with respect to B.

A semantic explanation of the fuzzy granule rule $[x]_B^{\lambda} \to D_k$ is as follows: Given an object $y \in U$, if $[x]_B^{\lambda}(y) > 0$, then $y \in D_k$, which can be used to classify the objects. According to Eq.(4), $[x]_B^{\lambda}(y) > 0$ if and only if $1 - R_B(x, y) < \lambda$ $(\lambda > 0)$ in which $1 - R_B(x, y)$ seems to be a kind of distance between the objects x and y. That is to say, if the distance $1 - R_B(x, y)$ is smaller than λ, then y can be classified by $[x]_B^{\lambda} \to D_k$ into D_k; otherwise, y cannot be classified by $[x]_B^{\lambda} \to D_k$. Thus, the classification by the fuzzy granular rule $[x]_B^{\lambda} \to D_k$ depends on both the distance $1 - R_B(x, y)$ and λ.

Let $\lambda^* = \underline{R_B^S} D_k(x)$. For any $\lambda \le \lambda^*$ and any $y \in U$, we have $[x]_B^{\lambda}(y) \leqslant [x]_B^{\lambda^*}(y)$. If there exist $\lambda_0 < \lambda^*$ and $y_0 \in U$ such that $[x]_B^{\lambda^*}(y_0) > 0$ and $[x]_B^{\lambda_0}(y_0) = 0$, then $[x]_B^{\lambda^*} \to D_k$ can classify y_0 into D_k whereas $[x]_B^{\lambda_0} \to D_k$ cannot classify y_0. Therefore, the larger the value of λ is, the more powerful classification ability the fuzzy granular rule $[x]_B^{\lambda} \to D_k$ possesses.

We henceforth focus on studying the fuzzy granular rule $[x]_B^{\lambda^*} \to D_k$ $(\lambda^* = \underline{R_B} D_k(x))$ which is denoted by $[x]_B^{\lambda} \to D_k$ $(\lambda = \underline{R_B} D_k(x))$ for notational simplicity. Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$ and $U/D = \{D_k : k = 1, 2, \ldots, l\}$. The set of all these fuzzy granular rules in $(U, B \cup D)$ is denoted by

$$R(B, D) = \{[x_i]_B^{\lambda_i} \to D_k : \; x_i \in D_k, \; \lambda_i = \underline{R_B} D_k(x_i) \in (0, 1]\}, \tag{5}$$

and the set of the objects that induce the fuzzy granular rules in $R(B, D)$ is denoted by

$$U_{R(B,D)} = \{x_i \in U : [x_i]_B^{\lambda_i} \to D_k \in R(B, D), \quad \lambda_i = \underline{R_B} D_k(x_i) \in (0, 1]\} \tag{6}$$

It has been known that the fuzzy granular rule induced by an object x_i can be used to classify objects, which means that the discriminating knowledge of the object x_i can be described by its induced fuzzy granular rules. Moreover, it should be pointed out that, if an object x_{i_0} has $\lambda_{i_0} = \underline{R_A} D_k(x_{i_0}) = 0$, then x_{i_0} cannot induce the corresponding fuzzy granular rule and then does not imply any discriminating knowledge. Therefore, the cardinality of $R(B, D)$ is not greater than n since each object x_i $(i \in \{1, 2, \ldots, n\})$ can induce one corresponding fuzzy granular rule $[x_i]_B^{\lambda_i} \to D_k$ only if $\lambda_i = \underline{R_B} D_k(x_i) \ne 0$.

3.2. Representative instance selection approach

According to the semantic explanation of a fuzzy granular rule $[x_i]_B^{\lambda_i} \to D_k \in R(B, D)$, we introduce the following definition.

Definition 4. Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$, $U/D = \{D_k : k = 1, 2, \ldots, l\}$ and $[x_i]_B^{\lambda_i} \to D_k \in R(B, D)$. For a given object $x_j \in U$, if $[x_i]_B^{\lambda_i}(x_j) > 0$, we say that the object x_j is covered by the fuzzy granular rule $[x_i]_B^{\lambda_i} \to D_k$. The number of the objects covered by $[x_i]_B^{\lambda_i} \to D_k$ is denoted by $\|[x_i]_B^{\lambda_i} \to D_k\|$.

Assume that an object $x_i \in U$ can induce a fuzzy granular rule $[x_i]_B^{\lambda_i} \to D_k$ in $R(B, D)$. If the fuzzy relation defined on each $a_t \in A$ satisfies reflexivity, i.e., $R_{\{a_t\}}(x_j, x_j) = 1$ for any $x_j \in U$, it is easily concluded that the object x_i is at least covered by the rule $[x_i]_B^{\lambda_i} \to D_k$ induced by x_i itself. The reflexivity is held for a variety of fuzzy relations and is always assumed to be true in the subsequent discussions. According to Definition 4, an object covered by a fuzzy granular rule can be factually classified by this granular rule. Therefore, the classification ability of $[x_i]_B^{\lambda_i} \to D_k$ can be measured by $\|[x_i]_B^{\lambda_i} \to D_k\|$. The larger $\|[x_i]_B^{\lambda_i} \to D_k\|$ is, the more powerful classification (or coverage) ability of $[x_i]_B^{\lambda_i} \to D_k$ possesses, which means that the object x_i is of more discriminating knowledge.

Definition 5. Let $(U, A \cup D)$ be a fuzzy decision system, $B \subseteq A$, $U = \{x_1, x_2, \ldots, x_n\}$, $U/D = \{D_k : k = 1, 2, \ldots, l\}$, and $R^*(B, D)$ be a subset of $R(B, D)$. If, for any $x_j \in U_{R(B, D)}$, there exists $[x_i]_B^{\lambda_i} \to D_k \in R^*(B, D)$ such that x_j is covered by $[x_i]_B^{\lambda_i} \to D_k$, we say that $U_{R(B, D)}$ is covered by $R^*(B, D)$. If $U_{R(B, D)}$ is covered by $R^*(B, D)$, but is not covered by $R^*(B, D) \setminus \{[x_i]_B^{\lambda_i} \to D_k\}$ for any $[x_i]_B^{\lambda_i} \to D_k \in R^*(B, D)$, we say that $R^*(B, D)$ is a minimal fuzzy granular rule set of $(U, B \cup D)$.

It can be known from Definition 5 that a minimal fuzzy granular rule set $R^*(B, D)$ possesses the same classification ability as $R(B, D)$. The fuzzy granular rules in $R^*(B, D)$ are induced by the objects in

$$U_{R^*(B,D)} = \{x_i \in U : [x_i]_B^{\lambda_i} \to D_k \in R^*(B, D), \quad \lambda_i = \underline{R_B} D_k(x_i) \in (0, 1]\} \tag{7}$$

which is factually the *representative instance (object) set* determined by the minimal fuzzy granular set $R^*(B, D)$. Therefore, the representative instance set has the same discriminating knowledge as the original instance set that induces all of the fuzzy granular rules. In the following, an algorithm for finding a minimal fuzzy granular rule set and the representative instance set is formulated.

Algorithm 1. Searching for a minimal fuzzy granular rule set and the representative instance set of a fuzzy decision system.

Input: A fuzzy decision system $(U, A \cup D)$ with $U = \{x_1, x_2, \ldots, x_n\}$ and $U/D = \{D_k : k = 1, 2, \ldots, l\}$.
Output: One minimal fuzzy granular rule set $R^*(A, D)$ and the representative instance set $U_{R^*(A,D)}$.

Step 1: Initialize $R^*(A, D) = \emptyset$ and $U_{R^*(A,D)} = \emptyset$.
Step 2: For each $x_i \in D_k$ $(i \in \{1, 2, \ldots, n\}, \; k = 1, 2, \ldots, l)$, compute $\lambda_i = \underline{R_A} D_k(x_i)$ according to Eq. (1) and then obtain $R(A, D)$.
Step 3: For each fuzzy granular rule $[x_i]_A^{\lambda_i} \to D_k \in R(A, D)$, compute $\|[x_i]_A^{\lambda_i} \to D_k\|$.
Step 4: Add into $R^*(A, D)$ the fuzzy granular rule $[x_{i_0}]_A^{\lambda_{i_0}} \to D_k$ satisfying $\|[x_{i_0}]_A^{\lambda_{i_0}} \to D_k\| = \max\limits_{[x_i]_A^{\lambda_i} \to D_k \in R(A,D)} \|[x_i]_A^{\lambda_i} \to D_k\|$ and add into $U_{R^*(A,D)}$ the corresponding instance x_{i_0}, and remove from $R(A, D)$ the rules induced by such objects that are covered by $[x_{i_0}]_A^{\lambda_{i_0}} \to D_k$.
Step 5: *If $R(A, D) \neq \emptyset$, then return to Step 4; otherwise, go to Step 6.*
Step 6: *Output $R^*(A, D)$ and $U_{R^*(A,D)}$.*

It should be pointed out that, in order to search for one minimal granular rule set, we need to previously compute such the similarity relation matrices with respect to each attribute that cost $O(|U|^2|A|)$ and are saved in the computer memory for the succeeding requirements. The time complexity of running both Steps 2 and 3 are $O(|U|(|A| + |U|))$. Running Steps 4 and 5 needs at most $O(|U|)$. Therefore, the time complexity of Algorithm 1 is at most $O(|U|(|A| + |U|))$.

Example 1. Table 1 shows a fuzzy decision system $(U, A \cup D)$, where $U = \{x_1, x_2, \ldots, x_{10}\}$, $A = \{a_1, a_2, a_3, a_4\}$ with $V_{a_t} = [0, 1]$ for each $a_t \in A$, $D = \{d\}$ with $V_d = \{1, 0\}$, and the decision partition $U/D = \{D_1 = \{x_1, x_2, \ldots, x_6\}, D_2 = \{x_7, x_8, x_9, x_{10}\}\}$.

The fuzzy relation for each $a_t \in A$ $(t \in \{1, 2, 3, 4\})$ is defined by

$$R_{\{a_t\}}(x_i, x_j) = 1 - |a_t(x_i) - a_t(x_j)|. \tag{8}$$

With the fuzzy relation $R_A = \bigcap\limits_{a_t \in A} R_{\{a_t\}}$, we obtain the fuzzy relation matrix with each element being $R_A(x_i, x_j)(i, j \in \{1, 2, \ldots, 10\})$

Table 1
A fuzzy decision system $(U, A\cup D)$.

U	a_1	a_2	a_3	a_4	d
x_1	0.7	0.9	0.4	0.6	1
x_2	0.6	0.9	0.3	0.7	1
x_3	0.6	0.8	0.3	0.5	1
x_4	0.3	0.5	0.7	0.2	1
x_5	0.3	0.4	0.8	0.3	1
x_6	0.4	0.5	0.6	0.3	1
x_7	0.9	0.4	0.5	0.9	0
x_8	0.8	0.5	0.4	0.8	0
x_9	0.2	0.6	0.7	1.0	0
x_{10}	0.1	0.7	0.8	0.8	0

as

$$\begin{pmatrix} 1 &&&&&&&&& \\ 0.9 & 1 &&&&&&&& \\ 0.9 & 0.8 & 1 &&&&&&& \\ 0.6 & 0.5 & 0.6 & 1 &&&&&& \\ 0.5 & 0.5 & 0.5 & 0.9 & 1 &&&&& \\ 0.6 & 0.6 & 0.7 & 0.9 & 0.8 & 1 &&&& \\ 0.5 & 0.5 & 0.6 & 0.3 & 0.4 & 0.4 & 1 &&& \\ 0.6 & 0.6 & 0.7 & 0.4 & 0.5 & 0.5 & 0.9 & 1 && \\ 0.5 & 0.6 & 0.5 & 0.2 & 0.3 & 0.3 & 0.3 & 0.4 & 1 & \\ 0.4 & 0.5 & 0.5 & 0.4 & 0.5 & 0.5 & 0.2 & 0.3 & 0.8 & 1 \end{pmatrix}.$$

Initialize $R^*(A,D)=\emptyset$ and $U_{R^*(A,D)}=\emptyset$. According to Eq. (1), it is obtained that

$$\underline{R_A}D_1(x_1)=\underline{R_A}D_1(x_2)=0.4,\ \underline{R_A}D_1(x_3)=0.3,$$
$$\underline{R_A}D_1(x_4)=0.6,\ \underline{R_A}D_1(x_5)=\underline{R_A}D_1(x_6)=0.5,$$
$$\underline{R_A}D_2(x_7)=0.4,\ \underline{R_A}D_2(x_8)=0.3,\ \underline{R_A}D_2(x_9)=0.4,$$
$$\underline{R_A}D_2(x_{10})=0.5.$$

Since each object $x_i\in U$ can induce a fuzzy granular rule $[x_i]_A^{\lambda_i}\to D_k\in R(A,D)$, $R(A, D)$ includes the following ten fuzzy granular rules.

$\mathbf{r}_1$: $[x_1]_A^{0.4}\to D_1$; $\mathbf{r}_2$: $[x_2]_A^{0.4}\to D_1$;
$\mathbf{r}_3$: $[x_3]_A^{0.3}\to D_1$; $\mathbf{r}_4$: $[x_4]_A^{0.6}\to D_1$;
$\mathbf{r}_5$: $[x_5]_A^{0.5}\to D_1$; $\mathbf{r}_6$: $[x_6]_A^{0.5}\to D_1$;
$\mathbf{r}_7$: $[x_7]_A^{0.4}\to D_2$; $\mathbf{r}_8$: $[x_8]_A^{0.3}\to D_2$;
$\mathbf{r}_9$: $[x_9]_A^{0.4}\to D_2$; $\mathbf{r}_{10}$: $[x_{10}]_A^{0.5}\to D_2$.

Then, compute $\|\mathbf{r}_i\|$ for each $\mathbf{r}_i\in R(A, D)$ $(i\in\{1,2,\ldots,10\})$. Specifically, x_1,x_2 and x_3 are covered by $\mathbf{r}_1$, $\mathbf{r}_2$ or $\mathbf{r}_3$; x_1,x_2,x_3,x_4,x_5 and x_6 are covered by $\mathbf{r}_4$ or $\mathbf{r}_6$; x_4,x_5 and x_6 are covered by $\mathbf{r}_5$; x_7 and x_8 are covered by $\mathbf{r}_7$ or $\mathbf{r}_8$; x_9 and x_{10} are covered by $\mathbf{r}_9$ or $\mathbf{r}_{10}$. Therefore, $\|\mathbf{r}_1\|=\|\mathbf{r}_2\|=\|\mathbf{r}_3\|=\|\mathbf{r}_5\|=3$, $\|\mathbf{r}_4\|=\|\mathbf{r}_6\|=6$ and $\|\mathbf{r}_7\|=\|\mathbf{r}_8\|=\|\mathbf{r}_9\|=\|\mathbf{r}_{10}\|=2$. Since $\|\mathbf{r}_4\|=\|\mathbf{r}_6\|=\max_{\mathbf{r}_i\in R(A,D)}\|\mathbf{r}_i\|=6$, we respectively choose $\mathbf{r}_4$ and x_4 to be added into $R^*(A, D)$ and $U_{R^*(A,D)}$, and remove from $R(A, D)$ the rules $\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3,\mathbf{r}_4,\mathbf{r}_5$ and $\mathbf{r}_6$ induced by the objects that are covered by $\mathbf{r}_4$. Since $R(A,D)=\{\mathbf{r}_7,\mathbf{r}_8,\mathbf{r}_9,\mathbf{r}_{10}\}\neq\emptyset$ at present, then Step 4 of Algorithm 1 is run twice and $\mathbf{r}_7$ and $\mathbf{r}_9$ will be stepwise added into $R^*(A, D)$ yielding $R(A,D)=\emptyset$. Therefore, $R^*(A,D)=\{\mathbf{r}_4,\mathbf{r}_7,\mathbf{r}_9\}$ is a minimal fuzzy granular rule set, and $U_{R^*(A,D)}=\{x_4,x_7,x_9\}$ is the representative instance set. Moreover, $\{\mathbf{r}_4, \mathbf{r}_8, \mathbf{r}_9\}$, $\{\mathbf{r}_4, \mathbf{r}_8, \mathbf{r}_{10}\}$ and $\{\mathbf{r}_6, \mathbf{r}_7, \mathbf{r}_9\}$ are also minimal fuzzy granular rule sets of $(U, A\cup D)$. That is to say, the minimal fuzzy granular rule set is of nonuniqueness, and then the determined representative instance set is also of nonuniqueness.

Example 2. The first column of Fig. 1 shows each data set composed by 1,000 random points uniformly distributed into the unit square, and each data set is partitioned into two classes by a circle line, divided into two classes by a peak line, and split into three classes by the hyperbolic line, respectively. The second column of Fig. 1 is the corresponding instance set selected by Algorithm 1. The similarity relation employed in this example is the same to that in Example 1, namely, Eq. (8). According to the results of instance selection, Algorithm 1 do select fewer instance points, and may prefer to select both the central points far from the decision boundary and the points near the decision boundary.

3.3. Implication relationship between the fuzzy granular rules

In general, there are some redundant conditional attributes in a fuzzy decision system $(U, A\cup D)$. If the redundant attributes are removed from $(U, A\cup D)$ without losing the classification ability of the rules $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, we have the reduced fuzzy decision system $(U, B\cup D)$.

Definition 6. Let $(U, A\cup D)$ be a fuzzy decision system, $U=\{x_1,x_2,\ldots,x_n\}$, $B\subseteq A$, $U/D=\{D_k:k=1,2,\ldots,l\}$, and $R^*(A, D)$ be a minimal fuzzy granular rule set of $(U, A\cup D)$. Given a fuzzy granular rule $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, if $[x_i]_B^{\lambda_i}\to D_k$ is a fuzzy granular rule, we say that $[x_i]_A^{\lambda_i}\to D_k$ can be implied by $[x_i]_B^{\lambda_i}\to D_k$ and denote this implication relationship by $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$; otherwise, we say that $[x_i]_A^{\lambda_i}\to D_k$ cannot be implied by $[x_i]_B^{\lambda_i}\to D_k$ and denote this implication relationship by $[x_i]_B^{\lambda_i}\to D_k\not\Rightarrow[x_i]_A^{\lambda_i}\to D_k$

Proposition 1. *Let $(U, A\cup D)$ be a fuzzy decision system, $U=\{x_1,x_2,\ldots,x_n\}$, $B\subseteq A$ and $U/D=\{D_k:k=1,2,\ldots,l\}$. For $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$ if and only if $\underline{R_B}D_k(x_i)=\lambda_i$.*

Proof. $\Rightarrow$) Since $R_A\subseteq R_B$, we have $\underline{D}_{R_{B_k}}\subseteq\underline{D}_{R_{A_k}}$. For $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, we know $\lambda_i=\underline{R_A}D_k(x_i)$ which implies $\underline{D}_{R_{B_k}}(x_i)\leqslant\lambda_i$. If $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$, then $[x_i]_B^{\lambda_i}\to D_k$ is a fuzzy granular rule and $[x_i]_B^{\lambda_i}\subseteq D_k$. According to Eq. (3), we have $\underline{D}_{R_{B_k}}(x_i)\geqslant\lambda_i$. Therefore, $\underline{R_B}D_k(x_i)=\lambda_i$.

$\Leftarrow$) For $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, if $\underline{R_B}D_k(x_i)=\lambda_i$, we have $[x_i]_B^{\lambda_i}\subseteq D_k$ according to Eq. (3), which indicates that $[x_i]_B^{\lambda_i}\to D_k$ is a fuzzy granular rule. Therefore, $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$. □

Similarly, it is easily known from Proposition 1 that, for $[x_i]_A^{\lambda_i}\to D_k\in R^*(A,D)$, $[x_i]_B^{\lambda_i}\to D_k\not\Rightarrow[x_i]_A^{\lambda_i}\to D_k$ if and only if $\underline{R_B}D_k(x_i)<\underline{R_A}D_k(x_i)=\lambda_i$. Furthermore, preserving the implication relationship between the rules is equivalent to keeping the lower approximation membership degree of the object invariant. For the implication relationship $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$, we have $[x_i]_A^{\lambda_i}\subseteq[x_i]_B^{\lambda_i}$ since $R_A\subseteq R_B$. Then, there exists an object $x_j\in U$ such that $0<[x_i]_A^{\lambda_i}(x_j)\leqslant[x_i]_B^{\lambda_i}(x_j)$, which implies that the objects correctly classified by $[x_i]_A^{\lambda_i}\to D_k$ can be also correctly classified by $[x_i]_B^{\lambda_i}\to D_k$. Therefore, it is known from $[x_i]_B^{\lambda_i}\to D_k\Rightarrow[x_i]_A^{\lambda_i}\to D_k$ that removing the attributes in $A\backslash B$ from A does not lose the classification information of $[x_i]_A^{\lambda_i}\to D_k$, and then cannot lose the discriminating information of the instance x_i.

4. Implication relationship preserved-attribute reduction for fuzzy decision systems

Let $(U, A\cup D)$ be a fuzzy decision system, $U=\{x_1,x_2,\ldots,x_n\}$ and $U/D=\{D_k:k=1,2,\ldots,l\}$, and $R^*(A, D)$ be a minimal fuzzy granular rule set of $(U, A\cup D)$. Then, the representative instance set determined by $R^*(A, D)$ is $U_{R^*(A,D)}$ which is denoted by U^* for nota-

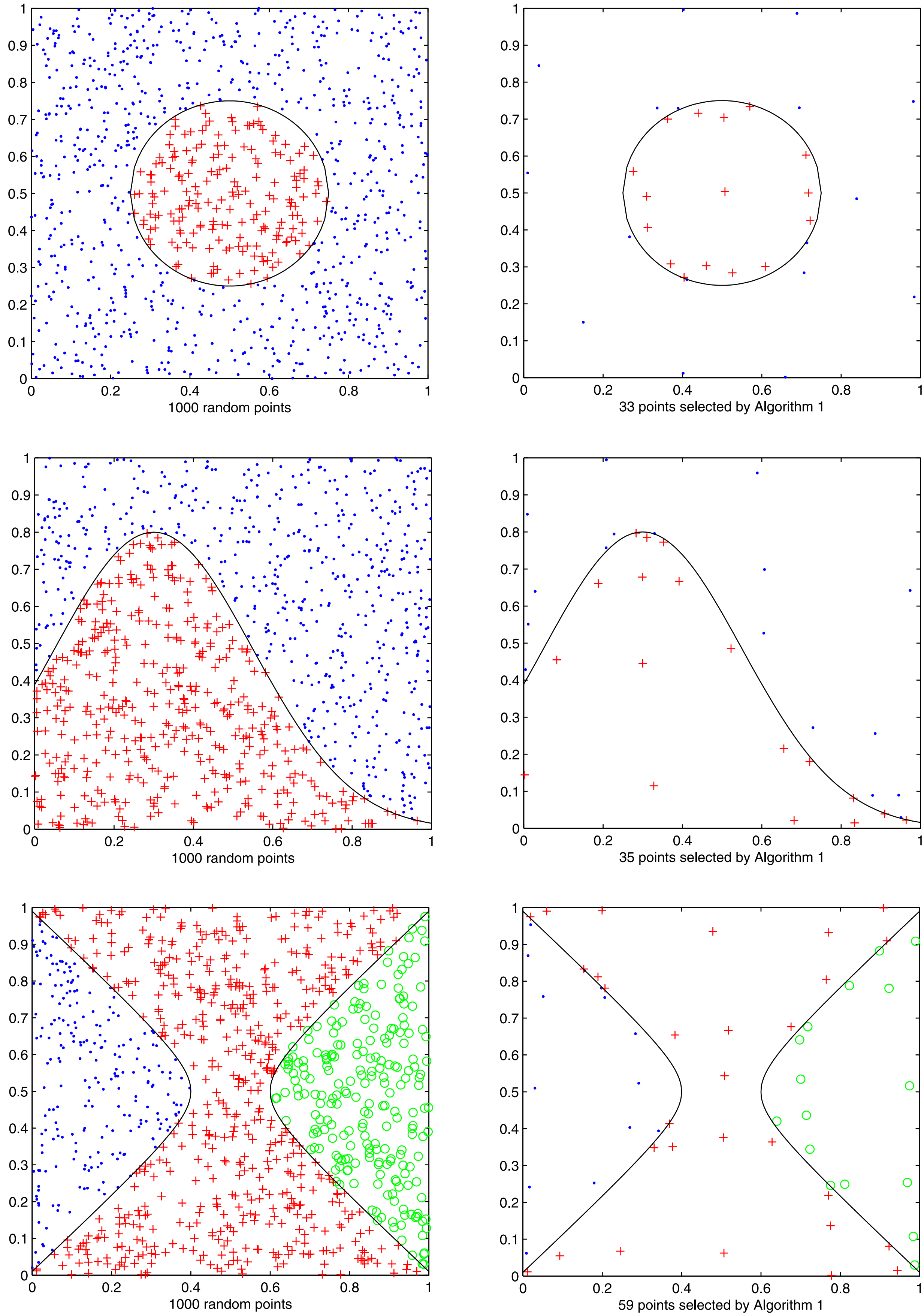

Fig. 1. Example of the instance points selected by Algorithm 1.

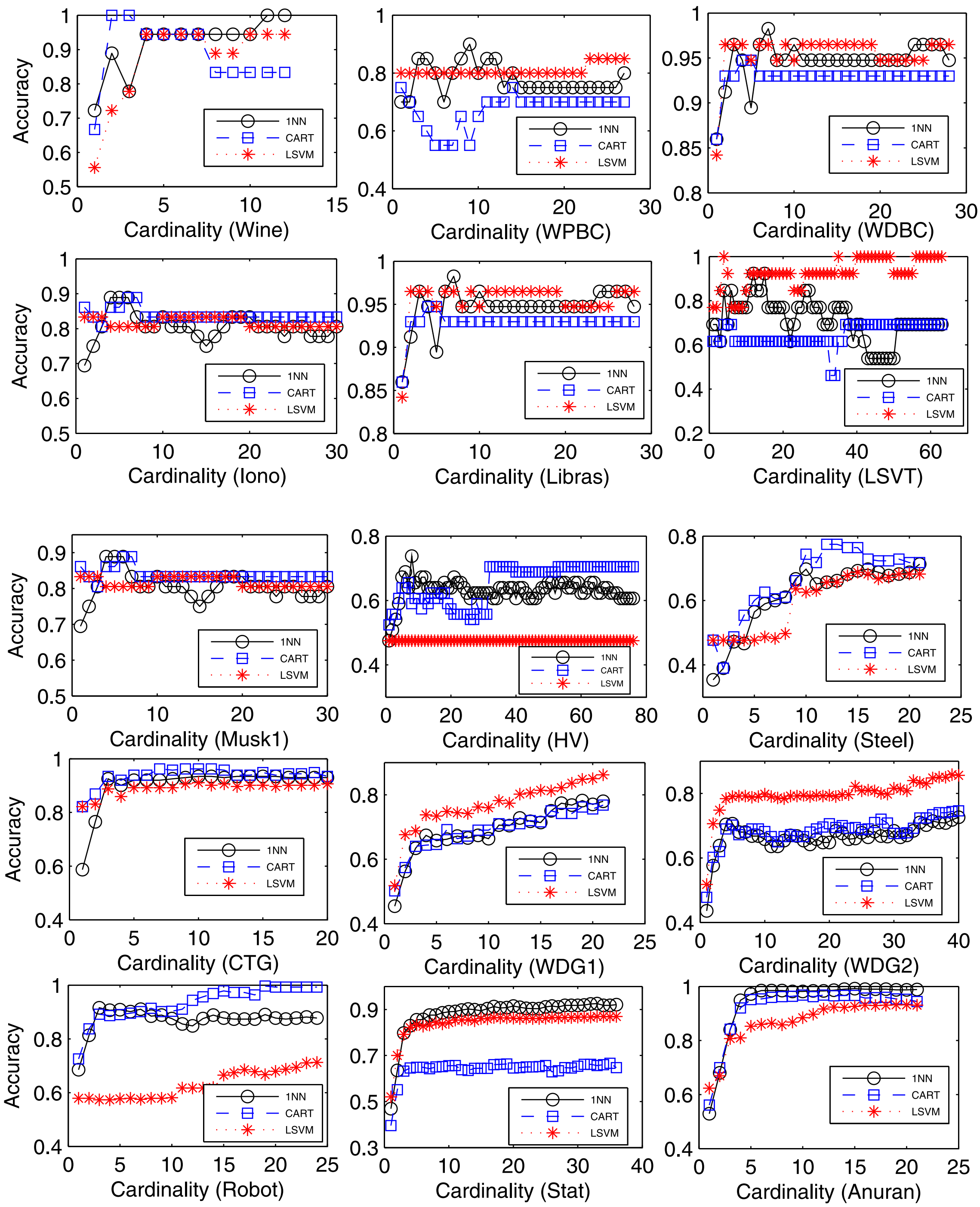

Fig. 2. Classification accuracies of the candidate sequence feature subsets obtained by Algorithm 2.

tional simplicity in the subsequent work if no confusion is made, i.e.,

$$U^* = \{x_i \in U : [x_i]_A^{\lambda_i} \to D_k \in R^*(A, D),$$
$$\lambda_i = \underline{R_A} D_k(x_i) \in (0, 1]\}. \quad (9)$$

Definition 7. Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $U/D = \{D_k : k = 1, 2, \ldots, l\}$, and $R^*(A, D)$ be a minimal fuzzy granular rule set of $(U, A \cup D)$. $B \subseteq A$ is called an implication relationship preserved-consistent set of $(U, A \cup D)$ if, for each $[x_i]_A^{\lambda_i} \to D_k \in R^*(A, D)$, we have $[x_i]_B^{\lambda_i} \to D_k \Rightarrow [x_i]_A^{\lambda_i} \to D_k$. Furthermore, if B is an implication relationship preserved-consistent set and there exists $[x_{i_0}]_A^{\lambda_{i_0}} \to D_k \in R^*(A, D)$ such that $[x_{i_0}]_{B \setminus \{a\}}^{\lambda_{i_0}} \to D_k \not\Rightarrow [x_{i_0}]_A^{\lambda_{i_0}} \to D_k$ for any $a \in B$, then B is said to be an implication relationship preserved-reduct of $(U, A \cup D)$.

Theorem 2. *Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, and $U/D = \{D_k : k = 1, 2, \ldots, l\}$. $B \subseteq A$ is an implication relationship preserved-reduct of $(U, A \cup D)$ if and only if $\underline{R_B} D_k(x_i) = \underline{R_A} D_k(x_i) = \lambda_i$ for each $x_i \in U^*$ and there exists $x_{i_0} \in U^*$ such that $\underline{R_{B \setminus \{a\}}} D_k(x_{i_0}) < \underline{R_A} D_k(x_{i_0}) = \lambda_{i_0}$ for any $a \in B$.*

Proof. $B \subseteq A$ is an implication relationship preserved-consistent set of $(U, A \cup D) \iff$ for each $[x_i]_A^{\lambda_i} \to D_k \in R^*(A, D)$, we have $[x_i]_B^{\lambda_i} \to D_k \Rightarrow [x_i]_A^{\lambda_i} \to D_k \iff \underline{R_B} D_k(x_i) = \underline{R_A} D_k(x_i) = \lambda_i$ for each $x_i \in U^*$, where the last equivalence relation is due to Proposition 1.

Table 2
Description of the data sets.

Data set	Abbreviation of data set	Number of objects	Number of conditional attributes	Number of classes
Wine	Wine	178	13	3
Wisconsin Prognostic Breast Cancer	WPBC	194	33	2
Wisconsin Diagnostic Breast Cancer	WDBC	569	30	2
Ionosphere	Iono	351	34	2
Libras Movement	Libras	360	90	15
LSVT Voice Rehabilitation Data Set	LSVT	126	310	2
Musk(Version 1)	Musk1	476	166	2
Hill-Valley	HV	606	100	2
Steel Plates Faults	Steel	1941	27	7
Cardiotocography	CTG	2126	20	3
Waveform Database Generator (Version 1)	WDG1	5000	21	3
Waveform Database Generator (Version 2)	WDG2	5000	40	3
Wall-Following Robot Navigation Data	Robot	5456	24	4
Statlog(Landsat Satellite)	Stat	6435	36	6
Anuran Calls (MFCCs)	Anuran	7195	21	4

Table 3
Average running time (second) of searching for one feature subset.

Data set	Similarity matrices	FWARA						DFBRA	MQRA	TRA	
		Instance selection		Filter	Wrapper					pretreatment	reduct
		Number	Time		1NN	CART	LSVM				
Wine	0.01	17.5	0.01	0.02	0.05	0.12	1.49	0.16	0.15	0.03	0.20
WPBC	0.01	40.3	0.01	0.33	0.08	0.76	4.76	1.50	0.97	0.11	1.15
WDBC	0.05	43.7	0.04	0.36	0.12	0.60	5.85	4.73	2.97	0.31	4.80
Iono	0.02	78.1	0.03	0.68	0.07	1.22	5.39	2.89	1.91	0.14	3.60
Libras	0.05	43.7	0.04	0.36	0.12	0.60	5.85	4.73	2.97	0.31	4.80
LSVT	0.04	26.8	0.02	13.83	0.31	1.46	17.85	107.04	43.67	0.57	27.77
Musk1	0.02	78.1	0.03	0.68	0.07	1.22	5.39	2.89	1.91	0.14	3.60
HV	0.19	125.3	0.11	25.5	0.59	25.35	43.51	270.92	32.32	0.72	45.12
Steel	0.52	313.7	0.65	4.27	0.85	8.06	12.88	23.69	15.71	0.79	46.33
CTG	0.49	164.0	0.46	1.38	0.21	3.98	6.80	15.73	12.88	0.76	26.95
WDG1	2.85	1013.4	4.97	16.97	3.10	25.60	46.14	74.50	72.59	3.73	154.43
WDG2	9.18	1394.4	14.21	102.90	19.61	194.89	239.93	334.08	302.91	8.47	482.97
Robot	3.67	868.3	5.37	40.52	0.81	20.10	137.12	229.50	241.89	2.94	221.00
Stat	29.24	673.8	78.65	60.70	17.00	72.87	166.84	28407.80	20035.18	9.92	992.30
Anuran	12.76	194.6	27.16	4.58	6.24	7.53	52.74	1059.00	400.74	7.06	421.88

Furthermore, if there exists $[x_{i_0}]_A^{\lambda_{i_0}} \to D_k \in R^*(A, D)$ such that $[x_{i_0}]_{B\setminus\{a\}}^{\lambda_{i_0}} \to D_k \nRightarrow [x_{i_0}]_A^{\lambda_{i_0}} \to D_k$ for any $a \in B \iff$ there exists $x_{i_0} \in U^*$ such that $\underline{R_{B\setminus\{a\}}}D_k(x_{i_0}) < \underline{R_A}D_k(x_{i_0}) = \lambda_{i_0}$ by Proposition 1.

Therefore, it is known from Definition 7 that the conclusion holds. □

According to Theorem 2, an implication relationship preserved-reduct B of $(U, A \cup D)$ is factually a minimal subset of A that preserves $\underline{R_B}D_k(x_i) = \underline{R_A}D_k(x_i) = \lambda_i$ for each $x_i \in U^*$.

Definition 8. Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$, and $U/D = \{D_k : k = 1, 2, \ldots, l\}$. The representative instance-based dependency function of D relative to B is defined by

$$\gamma_B^*(D) = \frac{\sum\limits_{x_i \in U^*} \mathrm{Pos}_B(D)(x_i)}{n}. \tag{10}$$

It should be pointed out that the membership degree of an object x_i belonging to the fuzzy positive region $\mathrm{Pos}_B(D)$ is computed by the following Lemma.

Lemma 1 ([63]). *Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$, and $U/D = \{D_k : k = 1, 2, \ldots, l\}$. Then,*

$$\mathrm{Pos}_B(D)(x_i) = \underline{R_B}D_k(x_i), \ \ x_i \in D_k. \tag{11}$$

As is known, the fuzzy lower approximation satisfies monotonicity, i.e., $\underline{D}_{R_{B_k}}(x_i) \leqslant \underline{D}_{R_{A_k}}(x_i)$ for $B \subseteq A$, which yields that the representative instance-based dependency function satisfies $\gamma_B^*(D) \leqslant \gamma_A^*(D)$.

Theorem 3. *Let $(U, A \cup D)$ be a fuzzy decision system, $U = \{x_1, x_2, \ldots, x_n\}$, and $U/D = \{D_k : k = 1, 2, \ldots, l\}$. $B \subseteq A$ is an implication relationship preserved-reduct of $(U, A \cup D)$ if and only if $\gamma_B^*(D) = \gamma_A^*(D)$ and $\gamma_{B\setminus\{a\}}^*(D) < \gamma_A^*(D)$ for any $a \in B$.*

Proof. $B \subseteq A$ is an implication relationship preserved-consistent set of $(U, A \cup D) \iff \underline{R_B}D_k(x_i) = \underline{R_A}D_k(x_i) = \lambda_i$ for each $x_i \in U^*$ by Theorem 2 $\iff \mathrm{Pos}_B(D)(x_i) = \mathrm{Pos}_A(D)(x_i)$ for each $x_i \in U^* \iff \sum\limits_{x_i \in U^*} \mathrm{Pos}_B(D)(x_i) = \sum\limits_{x_i \in U^*} \mathrm{Pos}_A(D)(x_i) \iff \gamma_B^*(D) = \gamma_A^*(D)$.

Furthermore, if there exists $[x_{i_0}]_A^{\lambda_{i_0}} \to D_k \in R^*(A, D)$ such that $[x_{i_0}]_{B\setminus\{a\}}^{\lambda_{i_0}} \to D_k \nRightarrow [x_{i_0}]_A^{\lambda_{i_0}} \to D_k$ for any $a \in B \iff$ there exists $x_{i_0} \in U^*$ such that $\underline{R_{B\setminus\{a\}}}D_k(x_{i_0}) < \underline{R_A}D_k(x_{i_0}) = \lambda_{i_0}$ for any $a \in B \iff$ there exists $x_{i_0} \in U^*$ such that $\mathrm{Pos}_{B\setminus\{a\}}(D)(x_i) < \mathrm{Pos}_A(D)(x_i)$ for any $a \in B \iff \sum\limits_{x_i \in U^*} \mathrm{Pos}_{B\setminus\{a\}}(D)(x_i) < \sum\limits_{x_i \in U^*} \mathrm{Pos}_A(D)(x_i)$ for any $a \in B \iff \gamma_{B\setminus\{a\}}^*(D) < \gamma_A^*(D)$ for any $a \in$ B.

In conclusion, according to Definition 7, B is an implication relationship preserved-reduct if and only if $\gamma_B^*(D) = \gamma_A^*(D)$ and $\gamma_{B\setminus\{a\}}^*(D) < \gamma_A^*(D)$ for any $a \in B$. □

According to Theorem 3, an implication relationship preserved-reduct B of $(U, A \cup D)$ is also a minimal subset of A that preserves $\gamma_B^*(D) = \gamma_A^*(D)$.

It should be pointed out that, from the viewpoint of the membership degrees of objects belonging to the lower approximation of the decision classes, the implication relationship-preserved reduction is different from the fuzzy dependency function-based reduction in [17,23]. The dependency function-based reduction preserves the membership degrees of all the objects, whereas the implication relationship-preserved reduction keeps the membership degrees of such objects that induce the rules in a minimal fuzzy granular rule set. Given a data set, since the implication relationship-preserved reduction only preserves the membership degrees of the partial objects, it may retain fewer attributes than the fuzzy dependency function-based reduction. Moreover, the implication relationship-preserved reduction is also different from the local reduction in [6] in that the local reduction preserves the membership degrees of the objects in some decision classes.

Based on the aforementioned work, a heuristic algorithm for computing an implication relationship-preserved reduct of a fuzzy decision system is formulated as follows.

Algorithm 2. Computing an implication relationship-preserved reduct of a fuzzy decision system.

Input: A fuzzy decision system $(U, A\cup D)$ with $U=\{x_1, x_2, \dots, x_n\}$, one minimal fuzzy granular rule set $R^*(A, D)$ and the representative instance set U^*.

Output: An implication relationship-preserved reduct B of $(U, A\cup D)$.

Step 1: Initialize $B=\emptyset$ and threshold $=-1$.

Step 2: Compute $\gamma_A^*(D)$.

Step 3: For each $a\in A\backslash B$, compute $\gamma_{B\cup\{a\}}^*(D)$.

Step 4: If $\gamma_{B\cup\{a_{i_0}\}}^*(D)=\max\limits_{a\in A\backslash B}\gamma_{B\cup\{a\}}^*(D)\geqslant threshold$, then update $B=B\cup\{a_{i_0}\}$ and threshold $=\gamma_{B\cup\{a_{i_0}\}}^*(D)$.

Step 5: If threshold $<\gamma_A^*(D)$, return to Step 3; otherwise, output B and terminate the algorithm.

The time complexity of the above algorithm is polynomial. In fact, the complexity of computing $\gamma_A^*(D)$ is $O(|U^*|)$ since the value $\underline{D}_{R_{A_k}}(x_i)$ for each $x_i\in U^*$ is contained in $R^*(A, D)$. The complexity of computing $\gamma_{B\cup\{a\}}^*(D)$ is at most $O(|U^*|(|A|+|U|))$. Then, Step 3 needs at most $O(|U^*||A|(|A|+|U|))$. Carrying out Step 4 needs $O(|A|)$. Totally, the time complexity of Algorithm 2 is at most $O(|U^*||A|(|A|+|U|))$.

Example 3. For the fuzzy decision system $(U, A\cup D)$ in Example 1, we have had $R(A,D)=\{\mathbf{r}_4,\mathbf{r}_7,\mathbf{r}_9\}$ and $U^*=\{x_4,x_7,x_9\}$. First, initialize $B=\emptyset$ and $threshold=-1$. It has been known from Example 1 that $\underline{R_A}D_1(x_4)=0.6$, $\underline{R_A}D_2(x_7)=0.4$ and $\underline{R_A}D_2(x_9)=0.4$, then we obtain $\gamma_A^*(D)=0.14$. Second, for each $a_i\in A$, compute $\gamma_{\{a_1\}}^*(D)=0.04$, $\gamma_{\{a_2\}}^*(D)=0.01$, $\gamma_{\{a_3\}}^*(D)=0.01$, and $\gamma_{\{a_4\}}^*(D)=0.11$, respectively. Since $\gamma_{\{a_4\}}^*(D)=\max\limits_{a_i\in A}\gamma_{\{a_i\}}^*(D)\geqslant threshold$, add a_4 into B and update $threshold=\gamma_{\{a_4\}}^*(D)=0.11$. Because $threshold<\gamma_A^*(D)$ at present, then Steps 3 and 4 in Algorithm 2 are run twice, and a_1 and a_2 are stepwise added into B yielding $threshold=\gamma_{\{a_4,a_1,a_2\}}^*(D)=\gamma_A^*(D)=0.14$. Thus, we obtain an implication relationship-preserved reduct $B=\{a_1,a_2,a_4\}$.

It should be pointed out that the feature subset obtained by Algorithm 2 may be an implication relationship-preserved consistent set rather than the reduct since Algorithm 2 is the forward addition technique. Whether or not there exists an approximate reduct containing fewer features and possessing better classification performance? Let $(U, A\cup D)$ be a fuzzy decision system with $A=\{a_1,a_2,\dots,a_m\}$. Assume that the attributes $a_{i_1}, a_{i_2}, \dots$ are added into the empty set one by one according

Table 4
Experimental results obtained by 1NN.

Data set	Accuracy of original data set	FWARA			DFBRA		MQRA		TRA		Paired t-test(w/t/l)	
		|FS|	$\|\cdot\|$	Accuracy	$\|\cdot\|$	Accuracy	$\|\cdot\|$	Accuracy	$\|\cdot\|$	Accuracy	Cardinality	Accuracy
Wine	96.05 ± 3.78	11.8	3.6	98.89 ± 3.51	12.8	96.05 ± 3.78	10.0	96.08 ± 4.59	6.7	94.90 ± 3.30	3/0/0	3/0/0
WPBC	75.18 ± 7.92	28.6	4.6	88.66 ± 5.75	32.2	74.68 ± 7.37	16.2	69.39 ± 11.31	8.5	72.16 ± 8.86	3/0/0	3/0/0
WDBC	95.08 ± 2.72	26.8	6.3	98.07 ± 1.74	29.9	95.08 ± 2.72	14.6	94.90 ± 3.16	10.2	95.08 ± 2.31	3/0/0	3/0/0
Iono	86.90 ± 4.50	30.8	3.6	94.32 ± 4.82	32.0	86.90 ± 4.50	16.1	89.18 ± 5.82	13.0	88.59 ± 5.57	3/0/0	3/0/0
Libras	95.08 ± 2.72	26.8	6.3	98.07 ± 1.74	29.9	95.08 ± 2.72	14.6	94.90 ± 3.16	10.2	95.08 ± 2.31	3/0/0	3/0/0
LSVT	75.64 ± 12.36	61.5	5.8	94.42 ± 6.59	96.2	75.64 ± 12.88	45.4	68.91 ± 9.82	6.1	75.64 ± 13.53	2/1/0	3/0/0
Musk1	86.90 ± 4.50	30.8	3.6	94.32 ± 4.82	32.0	86.90 ± 4.50	16.1	89.18 ± 5.82	13.0	88.59 ± 5.57	3/0/0	3/0/0
HV	58.06 ± 6.33	72.9	12.4	66.30 ± 8.95	90.0	57.40 ± 6.53	23.6	57.24 ± 7.69	14.0	56.10 ± 3.44	2/1/0	3/0/0
Steel	71.92 ± 2.80	20.5	12.8	77.28 ± 2.10	21.2	71.82 ± 2.80	13.2	69.91 ± 3.05	20.3	69.30 ± 3.34	2/1/0	3/0/0
CTG	90.97 ± 2.50	20.0	5.5	93.09 ± 1.92	20.0	90.97 ± 2.50	13.5	91.06 ± 2.54	16.6	91.11 ± 2.51	3/0/0	3/0/0
WDG1	77.24 ± 1.38	21.0	15.2	80.20 ± 1.79	21.0	77.24 ± 1.38	18.9	76.30 ± 2.15	18.7	75.38 ± 1.43	3/0/0	3/0/0
WDG2	73.86 ± 2.20	40.0	30.7	79.48 ± 2.74	40.0	73.86 ± 2.20	32.0	68.08 ± 1.85	14.0	66.46 ± 5.52	1/1/1	3/0/0
Robot	88.53 ± 1.65	24.0	3.6	93.37 ± 00.98	24.0	88.53 ± 1.65	21.6	88.42 ± 1.68	24.0	88.53 ± 1.65	3/0/0	3/0/0
Stat	90.32 ± 1.62	36.0	25.7	92.29 ± 1.45	36.0	90.32 ± 1.62	22.9	89.87 ± 1.09	35.9	90.38 ± 1.51	2/1/0	3/0/0
Anuran	99.35 ± 0.32	21.0	11.6	99.65 ± 0.26	21.0	99.35 ± 0.32	13.1	99.12 ± 0.40	18.3	99.31 ± 0.29	3/0/0	3/0/0

Table 5
Experimental results obtained by CART.

Data set	Accuracy of original data set	FWARA			DFBRA		MQRA		TRA		Paired t-test (w/t/l)	
		$\lvert FS\rvert$	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	Cardinality	Accuracy
Wine	88.73 ± 5.97	11.8	2.3	96.63 ± 3.91	12.8	88.73 ± 5.97	10.0	88.73 ± 5.97	6.7	88.20 ± 6.11	3/0/0	3/0/0
WPBC	66.95 ± 8.87	28.6	3.5	84.03 ± 6.59	32.2	66.95 ± 8.87	16.2	67.08 ± 7.80	8.5	70.63 ± 11.26	3/0/0	3/0/0
WDBC	92.08 ± 3.56	26.8	3.6	96.49 ± 2.34	29.9	92.08 ± 3.56	14.6	91.56 ± 3.10	10.2	93.31 ± 3.90	3/0/0	3/0/0
Iono	88.34 ± 5.37	30.8	4.6	95.17 ± 4.43	32.0	88.34 ± 5.37	16.1	90.34 ± 4.57	13.0	90.33 ± 6.70	3/0/0	3/0/0
Libras	92.08 ± 3.56	26.8	3.6	96.49 ± 2.34	29.9	92.08 ± 3.56	14.6	91.56 ± 3.10	10.2	93.31 ± 3.90	3/0/0	3/0/0
LSVT	70.96 ± 12.06	61.5	2.4	84.17 ± 5.28	96.2	73.14 ± 9.53	45.4	71.54 ± 16.30	6.1	71.54 ± 13.02	2/1/0	3/0/0
Musk1	88.34 ± 5.37	30.8	4.6	95.17 ± 4.43	32.0	88.34 ± 5.37	16.1	90.34 ± 4.57	13.0	90.33 ± 6.70	3/0/0	3/0/0
HV	59.88 ± 7.13	72.9	6.6	70.79 ± 8.74	90.0	59.88 ± 7.13	23.6	56.91 ± 8.01	14.0	57.58 ± 6.14	3/0/0	3/0/0
Steel	73.98 ± 2.18	20.5	11.5	79.08 ± 3.81	21.2	73.93 ± 2.28	13.2	72.75 ± 1.76	20.3	72.54 ± 3.66	3/0/0	3/0/0
CTG	93.27 ± 1.38	20.0	7.5	95.30 ± 1.73	20.0	93.27 ± 1.38	13.5	93.04 ± 1.63	16.6	93.13 ± 1.57	3/0/0	3/0/0
WDG1	75.82 ± 2.31	21.0	15.1	79.46 ± 1.71	21.0	75.82 ± 2.31	18.9	75.24 ± 1.26	18.7	75.50 ± 1.67	3/0/0	3/0/0
WDG2	74.46 ± 1.74	40.0	22.9	80.46 ± 1.23	40.0	74.46 ± 1.74	32.0	72.50 ± 2.34	14.0	67.94 ± 4.76	2/0/1	3/0/0
Robot	99.34 ± 0.38	24.0	7.3	99.58 ± 0.30	24.0	99.34 ± 0.38	21.6	99.30 ± 0.42	24.0	99.34 ± 0.38	3/0/0	3/0/0
Stat	66.14 ± 2.49	36.0	15.9	68.35 ± 2.28	36.0	66.14 ± 2.49	22.9	65.59 ± 2.20	35.9	66.06 ± 2.48	3/0/0	3/0/0
Anuran	95.75 ± 1.03	21.0	9.2	97.30 ± 0.82	21.0	95.75 ± 1.03	13.1	95.79 ± 1.04	18.3	95.77 ± 1.26	3/0/0	3/0/0

Table 6
Experimental results obtained by LSVM.

Data set	Accuracy of original data set	FWARA			DFBRA		MQRA		TRA		Paired t-test(w/t/l)	
		$\lvert FS\rvert$	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	$\lvert\cdot\rvert$	Accuracy	Cardinality	Accuracy
Wine	96.01 ± 3.91	11.8	3.2	97.75 ± 2.91	12.8	96.01 ± 3.91	10.0	94.90 ± 4.22	6.7	92.09 ± 7.10	3/0/0	3/0/0
WPBC	76.71 ± 7.66	28.6	5.8	78.29 ± 7.38	32.2	76.71 ± 7.66	16.2	75.74 ± 5.57	8.5	76.26 ± 6.13	3/0/0	2/1/0
WDBC	97.18 ± 2.23	26.8	3.3	98.42 ± 1.54	29.9	97.18 ± 2.23	14.6	96.66 ± 2.93	10.2	95.95 ± 3.22	3/0/0	3/0/0
Iono	84.63 ± 5.02	30.8	5.1	88.90 ± 6.88	32.0	84.63 ± 5.02	16.1	84.60 ± 7.05	13.0	83.76 ± 3.57	3/0/0	3/0/0
Libras	97.18 ± 2.23	26.8	3.3	98.42 ± 1.54	29.9	97.08 ± 2.23	14.6	96.66 ± 2.93	10.2	95.95 ± 3.22	3/0/0	3/0/0
LSVT	86.47 ± 7.40	61.5	3.7	93.65 ± 7.12	96.2	87.31 ± 9.77	45.4	84.94 ± 7.65	6.1	75.58 ± 12.27	3/0/0	3/0/0
Musk1	84.63 ± 5.02	30.8	5.1	88.90 ± 6.88	32.0	84.63 ± 5.02	16.1	84.60 ± 7.05	13.0	83.76 ± 3.57	3/0/0	3/0/0
HV	46.36 ± 3.18	72.9	7.2	47.18 ± 4.42	90.0	46.36 ± 3.08	23.6	45.70 ± 3.08	14.0	45.70 ± 3.08	2/1/0	2/1/0
Steel	70.02 ± 2.22	20.5	13.4	72.08 ± 2.82	21.2	69.81 ± 2.77	13.2	68.47 ± 4.07	20.3	67.03 ± 3.86	2/1/0	3/0/0
CTG	89.04 ± 2.87	20.0	10.3	90.36 ± 2.68	20.0	89.04 ± 2.87	13.5	88.67 ± 2.59	16.6	88.90 ± 2.52	3/0/0	3/0/0
WDG1	86.80 ± 0.98	21.0	16.5	88.32 ± 1.01	21.0	86.80 ± 0.98	18.9	85.46 ± 1.30	18.7	86.00 ± 1.42	3/0/0	3/0/0
WDG2	86.46 ± 1.23	40.0	29.7	88.46 ± 1.27	40.0	86.46 ± 1.23	32.0	82.40 ± 1.32	14.0	77.18 ± 4.24	1/1/1	3/0/0
Robot	71.06 ± 1.96	24.0	18.2	74.65 ± 1.34	24.0	71.06 ± 1.96	21.6	69.21 ± 2.67	24.0	71.06 ± 1.96	3/0/0	3/0/0
Stat	86.73 ± 1.73	36.0	24.6	87.77 ± 1.76	36.0	86.73 ± 1.73	22.9	86.62 ± 1.98	35.9	86.73 ± 1.73	2/1/0	3/0/0
Anuran	94.33 ± 0.97	21.0	13.8	94.73 ± 0.84	21.0	94.33 ± 0.97	13.1	93.26 ± 1.19	18.3	94.08 ± 1.11	2/1/0	3/0/0

to Steps 3 and 4 in Algorithm 2. The process continues until there exists some $t \in \{1, 2, \ldots, m\}$ such that $\gamma^*_{\{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}}(D) = \gamma^*_A(D)$, then Algorithm 2 is terminated. It is obtained that $\gamma^*_{\{a_{i_1}\}}(D) \leqslant \gamma^*_{\{a_{i_1}, a_{i_2}\}}(D) \leqslant \ldots \leqslant \gamma^*_{\{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}}(D)$. If $\{a_{i_1}\}$, $\{a_{i_1}, a_{i_2}\}$, $\{a_{i_1}, a_{i_2}, a_{i_3}\}, \ldots, \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$ are considered as the *candidate sequence feature subsets* which are used to build a classifier, respectively, the performance of the classifier built by some candidate feature subset with fewer features may be better than that built by $\{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$, which is caused by the fact that practical induction algorithms may benefit from the omission of features including strongly relevant features [26]. Denote $S_1 = \{a_{i_1}\}$, $S_2 = \{a_{i_1}, a_{i_2}\}$, ..., and $S_t = \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$. Some classifier is employed to compute the classification accuracies achieved by S_1, S_2, ..., and S_t, respectively. Assume that S_k $(1 \leqslant k \leqslant t)$ achieves the highest accuracy. Then, S_k is taken as the candidate best feature subset, and a backward elimination method is applied in S_k to select a best feature subset. Specifically, remove the jth $(j = 1, 2, \ldots, k)$ feature from S_k and denote the obtained feature subset as $S_k^{(j)}$. Then, compute the classification accuracies achieved by $S_k^{(1)}$, $S_k^{(2)}$,..., and $S_k^{(k)}$, respectively. If there exists j_0 $1 \leqslant j_0 \leqslant k$ such that $S_k^{(j_0)}$ acquires highest accuracy and the accuracy is not less than the accuracy achieved by S_k, $S_k^{(j_0)}$ is selected as the candidate best feature subset. The above procedure is repeated until no gain in either classification accuracy improvement or feature dimension reduction, and the backward elimination technique terminates. In conclusion, the process of selecting a best feature subset includes the following steps.

Step 1. Use Algorithm 2 to stepwise select the attributes $a_{i_1}, a_{i_2}, \ldots, a_{i_t}$.

Step 2. For the candidate sequence feature subsets $S_1 = \{a_{i_1}\}$, $S_2 = \{a_{i_1}, a_{i_2}\}, \ldots,$ $S_t = \{a_{i_1}, a_{i_2}, \ldots, a_{i_t}\}$, some classifier is employed to compute $acc(S_1)$, $acc(S_2)$, ..., and $acc(S_t)$, in which $acc(\cdot)$ is the accuracy obtained by some feature subset. Choose S_k satisfying $acc(S_k) = \max_{1 \leqslant k \leqslant t}\{acc(S_1), acc(S_2), \ldots, acc(S_t)\}$ as a candidate best feature subset, and let $S = S_k$ and $acc = acc(S_k)$.

Step 3. Remove the jth $(j = 1, 2, \ldots, |S|)$ feature from S and obtain the feature subsets $S^{(1)}$, $S^{(2)}$, ..., $S^{(|S|)}$, where $|S|$ is the cardinality of S. Compute $acc(S^{(1)})$, $acc(S^{(2)})$, ..., $acc(S^{(|S|)})$.

Step 4. If $acc(S^{(j_0)}) = \max_{1 \leqslant j_0 \leqslant |S|}\{acc(S^{(1)}), acc(S^{(2)}), \ldots, acc(S^{(|S|)})\} \geqslant acc$, update $S = S^{(j_0)}$ and $acc = acc(S^{(j_0)})$, and return to Step 3; otherwise, output $S^{(j_0)}$ as a best feature subset.

Step 1 and Steps 2–4 are factually the filter procedure and the wrapper procedure, respectively. Then, the above procedure for selecting a best approximate reduct is a filter-wrapper feature selection method which is also called a filter-wrapper approximate reduction algorithm (FWARA).

5. Numerical experiments

In this section, some numerical experiments are conducted to show the performance of FWARA. The experiments mainly focus on selecting a best feature subset by the proposed filter-wrapper approach, and comparing with other feature selection algorithms in terms of the computational time, the cardinality of the selected feature subsets, and the classification performance of the feature subsets. In order to achieve these tasks, we downloaded fifteen data sets from UCI Repository of machine learning databases. The data sets are briefly described in Table 2.

5.1. Pretreatment of the data sets and design of the experiments

For each data set, we denote the object set, conditional attribute set and decision attribute set by U, A and D, respectively. For each real-valued attribute $a \in A$, the attribute value of each object is normalized as

$$\tilde{a}(x_i) = \frac{a(x_i) - \min_j a(x_j)}{\max_j a(x_j) - \min_j a(x_j)}, \; x_i \in U, \tag{12}$$

so that $\tilde{a}(x_i) \in [0, 1]$ for each $x_i \in U$. Here, we still use a to denote the corresponding normalized conditional attribute for notational simplicity.

The experiments were designed as follows. Given one of the pretreated data sets, the ten-fold cross validation approach was used. Specifically, the instances were randomly divided into ten approximately equal parts. One of the ten parts was chosen as a testing data set and the remainder was taken as the training data set. Here, we denoted by U' the universe of discourse generated by the training data set. Then, a fuzzy relation for each normalized conditional attribute a is defined as

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

where x_i, $x_j \in U'$. In this way, a fuzzy decision system $(U', A \cup D)$ is formed for the training data set. We used Algorithm 1 to select representative instances from the training data set and used Algorithm 2 to filter a feature subset which then yielded the candidate sequence feature subsets, and then used the backward elimination technique to obtain a best feature subset. Here, the k-Nearest Neighbor Classifier with k = 1(1NN), the Classification and Regression Tree (CART) and Linear Support Vector Machine (LSVM) were respectively taken to evaluate the classification accuracies achieved by feature subsets, in which all the parameters of the classifiers are default. This process was repeated for each of the ten parts. Moreover, it should be pointed out that the classification accuracy (%) for each classifier was reported in the form of $v \pm \sigma$ in which v and σ are respectively the mean and the standardized error of the ten classification accuracies from the ten-fold cross validation experiment.

The experiments of comparison with other feature selection algorithms were performed on the same training data sets and the same testing data sets to ours. Finally, a paired t-test was performed to ensure that the experimental results were significantly different, where the significance level was specified to be 0.05.

5.2. Feature selection by FWARA

Let $(U', A \cup D)$ be the decision system formed by a given training data set in the ten-fold cross validation experiment. Firstly, we need to search for a minimal fuzzy granular rule set of $(U', A \cup D)$ and the representative instance set. In practice, the collected data usually contain noise which makes the induced rules have weak covering ability. Nevertheless, one minimal fuzzy granular rule set covers the decision discriminating information of all the instances. From the viewpoint of dealing with noise, we searched for such a minimal fuzzy granular rule set by Algorithm 1 that contains the rules with the covering instance numbers being greater than 1, and then obtained the corresponding representative instance set. Afterwards, we filtered features by Algorithm 2, and obtained the candidate sequence feature subsets. Then, we computed the classification accuracies achieved by the candidate sequence feature subsets. Specifically, for each of the candidate sequence feature subset, we only retained for both the training data set and the corresponding testing data set the features in the candidate sequence feature subset. The reduced training data was taken to build 1NN, CART and LSVM, respectively, and the reduced testing data set was classified by the built classifiers and then the rate of correct classification of the testing data, i.e., the classification accuracy, was computed. It should be noticed that the succeeding classification accuracies achieved by feature subsets were all computed in this way.

Table 7
Experimental results obtained by mRMR-wrapper algorithm.

Data set	1NN		CART		LSVM		Paired t-test(w/t/l)	
	$|\cdot|$	Accuracy	$|\cdot|$	Accuracy	$|\cdot|$	Accuracy	Cardinality	Accuracy
Wine	3.8	98.89 ± 2.34	2.6	93.82 ± 6.71	3.1	98.89 ± 2.34	0/3/0	0/3/0
WPBC	7.1	87.58 ± 8.55	3.4	84.47 ± 8.13	5.4	78.29 ± 7.38	1/2/0	0/3/0
WDBC	4.6	98.60 ± 1.61	2.7	96.66 ± 3.14	3.3	98.59 ± 1.62	0/1/2	0/3/0
Iono	6.5	96.59 ± 3.23	4.0	95.44 ± 3.61	7.5	90.31 ± 5.08	1/2/0	0/2/1
Libras	4.6	98.60 ± 1.61	2.7	96.66 ± 3.14	3.3	98.59 ± 1.62	0/1/2	0/3/0
LSVT	8.2	97.69 ± 3.72	2.1	91.22 ± 6.03	3.2	95.26 ± 5.47	0/3/0	0/2/1
Musk1	6.5	96.59 ± 3.23	4.0	95.44 ± 3.61	7.5	90.31 ± 5.08	1/2/0	0/2/1
HV	29.9	69.64 ± 4.03	7.6	71.08 ± 9.49	12.0	48.00 ± 4.92	0/3/0	0/2/1
Steel	13.5	77.02 ± 2.12	12.8	79.34 ± 2.28	14.3	72.90 ± 2.48	1/2/0	0/2/1
CTG	6.2	93.51 ± 2.55	9.5	95.44 ± 1.75	10.9	91.06 ± 2.24	1/2/0	0/3/0
WDG1	13.8	82.30 ± 1.07	12.6	79.50 ± 1.75	15.5	88.08 ± 0.99	0/2/1	1/1/1
WDG2	14.0	81.84 ± 1.07	12.3	79.18 ± 1.32	15.5	88.40 ± 1.50	0/0/3	1/1/1
Robot	4.0	93.33 ± 0.61	7.6	99.60 ± 0.30	16.9	73.48 ± 2.08	0/3/0	1/2/0
Stat	27.6	92.35 ± 1.61	22.5	69.00 ± 2.74	21.3	87.65 ± 1.68	1/2/0	0/3/0
Anuran	12.1	99.61 ± 0.31	12.1	97.29 ± 1.11	15.1	94.98 ± 0.90	1/2/0	0/2/1

The classification results of the candidate sequence feature subsets were depicted in Fig. 2. Here, the results depicted in Fig. 2 were obtained from one training data set and the corresponding testing data set in the ten-fold cross validation experiment, and the horizontal axis and the vertical axis of each subgraph express the cardinalities and the accuracies of the candidate sequence feature subsets, respectively.

It is seen from Fig. 2 that, with the increase of the cardinality of the candidate sequence feature subset, the classification accuracies of almost all the data sets increase significantly from the beginning to some value. Afterwards, the accuracies of the data sets WDBC, Iono, Libras, Steel, CTG, WDG1, WDG2, Robot, Stat and Anuran increase slowly or keep invariant, and the accuracies of the data sets Wine, WPBC, LSVT, HV and Musk1 fluctuate on a range. Therefore, the feature subset directly filtered by Algorithm 2 may not be the best, and then the wrapper procedure is conducted.

The candidate sequence feature subset with highest classification accuracy was denoted by S for convenient description and was taken to select a best feature subset by the backward elimination technique. Specifically, each feature of S was removed to yield a new feature subset and the classification accuracy of the new feature subset was computed, in which the feature subset achieving the highest accuracy that is not less than the accuracy obtained by S was used to update S. The iteration procedure was repeated until the termination condition was satisfied, and a best feature subset was obtained. It should be pointed out that different classifiers may get diverse best feature subsets since the wrapper procedure depends on the used classifiers.

5.3. Comparison with other reduction methods

In this subsection, the computational time and the effectiveness of the feature subset obtained by FWARA are compared with those of the feature subsets respectively obtained by the dependency function-based reduction algorithm (DFBRA) in [23,25], the modified quick reduction algorithm (MQRA) [8] and the traditional reduction algorithm (TRA).

As is well known, the traditional rough set theory is powerful in discovering knowledge in a data set with nominal attributes. Therefore, it is necessary to perform discretization on the real-valued conditional attributes before TRA is used. To achieve this task, the values of the real-valued attributes in each training data set was discretized into three nominal values by the fuzzy C-means approach. The forward addition algorithm was then used to search for a dependency function-based reduct for each discretized training data set. DFBRA in [23,25] and MQRA in [8] were used to search for one reduct of each fuzzy decision system, respectively. Here, the evaluation measure in MQRA is $\gamma_B = |\mathrm{Pos}_B|/|\mathrm{Pos}_A|$ and the degree threshold $\alpha = 0.95$ since the measure γ needs less time to be computed and possesses better performance, and $\alpha = 0.95$ is a suitable overall choice as claimed in [8]. Additionally, it should be pointed out that the fuzzy lower approximations in [23,25] and [8] are taken the same to that of this paper for convenient comparison.

5.3.1. Comparison on computational time

We list in Table 3 the average running time of searching for one reduct by each reduction algorithm in the ten-fold cross validation experiment. It should be noted that, for each fuzzy decision system, the similarity relation matrices with respect to each attribute were previously computed and saved in the computer memory for FWARA, DFBRA and MQRA, and the average running time is listed in 2nd column of Table 3. Besides, both the discretization and the computation of the equivalence relation matrices are deemed as the pretreatment process for TRA, and the average running time is reported in the last 2nd column of Table 3. Moreover, in the ten-fold cross validation experiment, both the average number of the representative instances and the average running time of selecting instances are listed in the 3rd and 4th columns of Table 3, respectively. The experiments were performed by Matlab on a personal computer with Intel(R) Core(TM) i7-4510U CPU @2.00 GHz configuration, 8G Memory and the 64-bit Windows 7 system.

The average running time of searching for a best feature subset by FWARA is the summation of the average running time of the instance selection and the filter procedure as well as the wrapper procedure. It can be seen form Table 3 that the average time of the instance selection is less than the average time of the filter procedure for almost all the data sets. Furthermore, the average time of the filter procedure is less than the average running time of the other three reduction algorithms, which is mainly caused by the fact that the filter process only concerns the representative instances rather than all the instances concerned by the other reduction algorithms. Especially for the larger data sets WDG1, WDG2, Robot, Stat and Anuran, the average filter time is greatly less than the average running time of the other three reduction algorithms. Therefore, the way to select representative instances may provide an approach to deal with large data. Besides, the wrapper time depends on the classifiers where 1NN costs the lest time and LSVM spends the most time. In conclusion, the average running time of FWARA with the classifier of the wrapper procedure being 1NN is less than or even greatly less than that of the other three reduction algorithms for all of the data sets, and FWARA with the classifier in the wrapper procedure being either CART or LSVM costs more time on some data sets.

5.3.2. Comparison on cardinality and accuracy

The classification accuracies achieved by the obtained feature subsets were computed, and the classification results of 1NN, CART and LSVM are reported in Tables 4–6, respectively. In Tables 4–6, it should be pointed out that the notation |FS| indicates the average cardinality of the feature subset acquired by the filter procedure (i.e., Algorithm 2) of FWARA, and $|\cdot|$ represents the average cardinality of the feature subset obtained by the corresponding algorithm. Furthermore, both the cardinality and the accuracy of the feature subset obtained by FWARA were statistically compared with those acquired by DFBRA, MQRA and TRA by using the paired t-test, respectively. The comparison results are listed in the last two columns of Tables 4–6, respectively. It should be indicated that "w" is the number of win achieved by our FWARA, in which win means that the cardinality (or accuracy) of the feature subset obtained by FWARA is significantly fewer (or higher) than that of DFBRA, MQRA or TRA; "t" is the number of tie achieved by our FWARA, in which tie means that the results obtained by FWARA have no statistically difference with that of DFBRA, MQRA or TRA; similarly, "l" is the number of lose achieved by our FWARA.

Since DFBRA, MQRA and TRA are factually the filter algorithms, the results of the filter procedure (Algorithm 2) of FWARA are firstly taken to be compared. Here, only the results in Table 4 are used to be elaborated since the similar conclusions can also be obtained from Tables 5 and 6. The features obtained by Algorithm 2 are fewer than or equal to those selected by DFBRA, which can be known from the comparison between the 3rd and 6th columns of Table 4. The reason is that both Algorithm 2 and DFBRA are the forward addition algorithms, whereas the feature subset acquired by Algorithm 2 preserves the fuzzy lower approximation values of the representative instances rather than all of the instances. Additionally, the features obtained by MQRA are fewer than those obtained by Algorithm 2 due to the threshold control of the evaluation measure of MQRA. Nevertheless, the fewer features obtained by MQRA cannot guarantee the accuracy (See the results of the data sets WPBC, LSVT and WDG2 in Table 4). The features obtained by TRA are not more than those obtained by Algorithm 2, and even for the data set WDG2 the features obtained by TRA are obviously few but the accuracy is low. In conclusion, Algorithm 2 has more advantage in computational time but less advantage in feature numbers compared with the other three reduction algorithms. Then, adding the wrapper procedure into Algorithm 2 yields FWARA, and the results obtained by FWARA are listed in the 4th and 5th columns of Tables 4–6, respectively.

It can be seen clearly from Tables 4–6 that, for almost all the data sets, FWARA outperforms the other three reduction algorithms in terms of both cardinality and accuracy of the feature subset. Specifically, in Table 4, FWARA achieves significantly fewest features and highest accuracy for the whole data sets except the data sets LSVT, HV, Steel, WDG2 and Stat. For each of the data sets LSVT, HV, Steel and Stat, the cardinality of the feature subset got by FWARA is not significantly different from that obtained by MQRA or TRA, whereas the accuracy achieved by FWARA is significantly higher than MQRA or TRA. Moreover, for the data set WDG2, FWARA obtains significantly more features than TRA and gets no significantly different feature subset cardinality from MQRA, but FWARA achieves significantly higher accuracy. Similar conclusions can be easily obtained from Tables 5 and 6. Therefore, FWARA is of effectiveness in terms of both acquiring few features and achieving high accuracy, in which the effectiveness may mainly contribute to the wrapper procedure.

5.3.3. Comparison with mRMR-wrapper algorithm

In this subsection, both the cardinality and the classification accuracy of the feature subset obtained by the proposed filter-wrapper algorithm are compared with those acquired by the mRMR-wrapper algorithm [38] which is a state-of-the-art feature selection method including the wrapper procedure. The mRMR needs to pre-specify such the number of the candidate features that was set to be the number of the conditional attributes in each original data set. The backward wrapper technique in [38] was taken in the experiment. It should be pointed out that the mRMR method is factually a feature permutation approach and thus the computational time is extremely little. Moreover, both the wrapper technique in [38] and that of ours are backward elimination techniques. Therefore, the average running time of searching for a best feature subset is not compared here. We report in Table 7 that the average cardinality and accuracy of the feature subset obtained by the mRMR-wrapper algorithm for each data set. Moreover, the paired t-test was used to compare the statistical differences between the results of our FWARA and the mRMR-wrapper algorithm under the same classifier, where "w/t/l" indicates the number of win/tie/lose achieved by FWARA, respectively.

It is obtained from Table 7 that the average numbers of win/tie/lose achieved by FWARA for feature subset cardinality and accuracy are 0.5/2.0/0.5 and 0.2/2.3/0.5, respectively. Then, FWARA and the mRMR-wrapper algorithm nearly make a draw with respect to the feature subset cardinality, and the mRMR-wrapper algorithm has a little advantage to possess higher accuracy for some data sets. Therefore, FWARA is of competitiveness compared with the mRMR-wrapper algorithm. In the meantime, it can be known from the whole numerical experiments that the wrapper procedure should be properly taken in feature selection since the wrapper procedure does work for both feature dimension reduction and classification accuracy improvement and also needs more computational time.

6. Summary

In this paper, we present a representative instance-based feature selection approach with fuzzy rough sets. The concept of a fuzzy granular rule is put forward to describe the discriminating information of an instance for fuzzy decision systems. Via acquiring a minimal fuzzy granular rule set, the corresponding representative instance set is obtained. The implication relationship between the fuzzy granular rules is investigated and an implication relationship-preserved reduction is formulated to preserve the discriminating information of the representative instances while removing some attributes. Then, a representative instance-based feature selection algorithm with the forward addition procedure is provided. Furthermore, by adding the backward elimination procedure, the feature selection algorithm becomes a filter-wrapper approach (i.e., FWARA) which is suggested to obtain a best feature subset. The results of numerical experiments shown that the representative instance-based feature selection algorithm costed the least computational time of finding a feature subset for each data set, and FWARA has significant advantages in both the cardinality and accuracy of the feature subset.

One of the highlights of this paper is that a novel instance selection approach is presented according to the coverage ability of the fuzzy granular rules. The instance selection method may have some other applications besides feature selection. For example, the representative instances may be considered to directly build some classifiers to alleviate computational time, and the instance selection may be used to deal with dynamic data environment in which a new coming instance is compared with the representative instances rather than the whole instances to acquire dynamic information. Moreover, the scalability of the instance selection approach for large data sets is needed to be investigated. In our future work, the applications of representative instance will be further investigated, and the fuzzy granular rules will be considered to build a rule-classifier.

Acknowledgments

The authors thank the reviewers and the associate editor for their valuable comments and suggestions which lead to much improvement on the paper. This work was supported by the National Natural Science Foundation of China (Nos. 61602372, 71471060 and 61572019) and the Ph.D Research Startup Foundation of Xi'an University of Technology (No. 109-256081504).

References

[1] J.R. Anaraki, S. Samet, J.H. Lee, C.W. Ahn, SUFFUSE: simultaneous fuzzy-rough feature-sample selection, J. Adv. Inf. Technol. 6 (2015) 103–110.
[2] M.J. Benítez-Caballero, J. Medina, E. Ramírez-Poussa, D. Ślęzak, Bireducts with tolerance relations, Inf. Sci. 435 (2018) 26–39.
[3] P.S. Bradley, U. Fayyad, C. Reina, Scaling clustering algorithms to large databases, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 1998, pp. 9–15.
[4] D.G. Chen, Q.H. Hu, Y.P. Yang, Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets, Inf. Sci. 181 (2011) 5169–5179.
[5] D.G. Chen, S.Y. Zhao, L. Zhang, Y.P. Yang, X. Zhang, Sample pair selection for attribute reduction with rough set, IEEE Trans. Knowledge Data Eng. 24 (2012) 2080–2093.
[6] D.G. Chen, S.Y. Zhao, Local reduction of decision system with fuzzy rough sets, Fuzzy Sets Syst. 161 (2010) 1871–1883.
[7] Y.M. Chen, D.Q. Miao, R.Z. Wang, A rough set approach to feature selection based on ant colony optimization, Pattern Recognit. Lett. 31 (2010) 226–233.
[8] C. Cornelis, R. Jensen, G. Hurtado, D. Ślęzak, Attribute selection with fuzzy decision reducts, Inf. Sci. 180 (2010) 209–224.
[9] J.H. Dai, Q.H. Hu, H. Hu, D.B. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, IEEE Trans. Syst.10.1109/TFUZZ.2017.2698420
[10] J. Derrac, C. Cornelis, S. García, F. Herrera, Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection, Inf. Sci. 186 (2012) 73–92.
[11] J. Derrac, N. Verbiest, S. Garcia, C. Cornelis, F. Herrera, On the use of evolutionary feature selection for improving fuzzy rough set based prototype selection, Soft Comput. 17 (2013) 223–238.
[12] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. General Syst. 17 (1990) 191–209.
[13] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, in: Proceedings of ACM-SIGMOD, ACM Press, 1998, pp. 73–84.
[14] P.E. Hart, The condensed nearest neighbor rule, IEEE Trans. Inf. Theory 14 (1968) 515–516.
[15] Q. He, Z.X. Xie, Q.H. Hu, C.X. Wu, Neighborhood based sample and feature selection for SVM classification learning, Neurocomputing 74 (2011) 1585–1594.
[16] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximations spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2006) 191–201.
[17] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (2006) 414–423.
[18] Q.H. Hu, L. Zhang, S. An, D. Zhang, D.R. Yu, On robust fuzzy rough set models, IEEE Trans. Fuzzy Syst. 20 (2012) 636–651.
[19] Q.H. Hu, L. Zhang, D.G. Chen, W. Pedrycz, D.R. Yu, Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications, Int. J. Approx. Reason. 51 (2010) 453–471.
[20] Q.H. Hu, L.J. Zhang, Y.C. Zhou, P. Witold, Large-scale multi-modality attribute reduction with multi-kernel fuzzy rough sets, IEEE Trans. Fuzzy Syst. (2017), doi:10.1109/TFUZZ.2017.2647966.
[21] D.C. Li, W.Z. Wu, On the characterization of fuzzy rough sets based on a pair of implications, Int. J. Mach. Learn. Cybern. (2017) 1–12.
[22] R. Jensen, C. Cornelis, Fuzzy-rough instance selection, in: WCCI 2010 IEEE World Congress on Computational Intelligence, 2010, pp. 1–7. Barcelona
[23] R. Jensen, Q. Shen, Fuzzy-rough attributes reduction with application to web categorization, Fuzzy Sets Syst. 141 (2004) 469–485.
[24] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute reduction, IEEE Trans. Fuzzy Syst. 15 (2007) 73–89.
[25] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (2009) 824–837.
[26] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[27] M. Kryszkiewicz, Comparative study of alternative type of knowledge reduction in inconsistent systems, Int. J. Intell. Syst. 16 (2001) 105–120.
[28] J.Y. Liang, Z.B. Xu, The algorithm on knowledge reduction in incomplete information systems, Int. J. Uncertainty Fuzziness Knowl. Based Syst. 10 (2002) 95–103.
[29] J.Y. Liang, F. Wang, C.Y. Dang, Y.H. Qian, A group incremental approach to feature selection applying rough set technique, IEEE Trans. Knowl. Data Eng. 26 (2014) 294–308.
[30] H. Liu, H. Motoda, On issues of instance selection, Data Mining Knowl. Discov. 6 (2002) 115–130.
[31] H. Liu, H. Motoda, L. Yu, A selective sampling approach to active feature selection, Artif. Intell. 159 (2004) 49–74.
[32] J.S. Mi, Y. Leung, H.Y. Zhao, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Inf. Sci. 178 (2008) 3203–3213.
[33] J.S. Mi, W.Z. Wu, W.X. Zhang, Approaches to knowledge reduction based on variable precision rough set model, Inf. Sci. 159 (2004) 255–272.
[34] J.S. Mi, W.X. Zhang, An axiomatic characterization of a fuzzy generalization of rough sets, Inf. Sci. 160 (2004) 235–249.
[35] N.N. Morsi, M.M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets Syst. 100 (1998) 327–342.
[36] N.M. Parthaláin, R. Jensen, Simultaneous feature and instance selection using fuzzy-rough bireducts, in: In IEEE International Conference on Fuzzy Systems, 2013, pp. 1–8.
[37] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[38] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[39] Y.H. Qian, J.Y. Liang, C.Y. Dang, D.W. Tang, Set-valued ordered information systems, Information Sciences 179 (2009) 2809–2832.
[40] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artif. Intell. 174 (2010) 597–618.
[41] Y.H. Qian, Q. Wang, H.H. Cheng, J.Y. Liang, C.Y. Dang, Fuzzy-rough feature selection accelerator, Fuzzy Sets Syst. 258 (2015) 61–78.
[42] A.M. Radzikowska, E.E. Kerre, A comparative study of fuzzy rough sets, Fuzzy Sets Syst. 126 (2002) 137–155.
[43] B. Schölkopf, C. Burges, V. Vapnik, Extracting support data for a given task, in: U. Fayyad, R. Uthurusamy (Eds.), Proceedings of the First International Conference on Knowledge Discvoery and Data Mining, AAAI Press, Menlo Park, CA, 1995, pp. 252–257.
[44] A. Janusz, D. Ślęzak, Ensembles of bireducts: Towards robust classification and simple representation, in: International Conference on Future Generation Information Technology, Springer, Berlin, Heidelberg, 2011, pp. 64–77.
[45] S. Stawicki, S. Widz, Decision bireducts and approximate decision reducts: Comparison of two approaches to attribute subset ensemble construction, in: Proceedings of the Federated Conference on Computer Science and Information Systems, 2012, pp. 331–338.
[46] S. Stawicki, D. Ślęzak, A. Janusz, S. Widz, Decision bireducts and decision reducts—a comparison, Int. J. Approx. Reason. 84 (2017) 75–109.
[47] H. Toivonen, Sampling large databases for association rules, in: Proceedings of the 22nd VLDB conference, Mumbai, India, 1996, pp. 134–145.
[48] E.C.C. Tsang, D.G. Chen, D.S. Yeung, X.Z. Wang, J.W.T. Lee, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (2008) 1130–1141.
[49] E.C.C. Tsang, Q.H. Hu, D.G. Chen, Feature and instance reduction for PNN classifiers based on fuzzy rough sets, Int. J. Mach. Learn. Cybern. 7 (2016) 1–11.
[50] N. Verbiest, C. Cornelis, F. Herrera, FRPS: a fuzzy rough prototype selection method, Pattern Recognit. 46 (2013) 2770–2782.
[51] S. Vluymans, L. D'eer, Y. Saeys, C. Cornelis, Applications of fuzzy rough set theory in machine learning: a survey, Fundamenta Informaticae 142 (2015) 53–86.
[52] C.Z. Wang, Y.L. Qi, M.W. Shao, Q.H. Hu, D.G. Chen, Y.H. Qian, Y.J. Lin, A fitting model for feature selection with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 25 (2017) 741–753.
[53] C.Z. Wang, M.W. Shao, Q. He, Y.H. Qian, Y.L. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl. Based Syst. 111 (2016) 173–179.
[54] G.Y. Wang, H. Yu, D.C. Yang, Decision table reduction based on conditional information entropy, Chin. J. Comput. 25 (2002) 759–766.
[55] W.Z. Wu, Y. Leung, M.W. Shao, Generalized fuzzy rough approximation operators determined by fuzzy implicators, Int. J. Approx. Reason. 54 (2013) 1388–1409.
[56] W.Z. Wu, J.S. Mi, W.X. Zhang, Generalized fuzzy rough sets, Inf. Sci. 151 (2003) 263–282.
[57] W.Z. Wu, W.X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Inf. Sci. 159 (2004) 233–254.
[58] Y.Y. Yang, D.G. Chen, H. Wang, Active sample selection based incremental algorithm for attribute reduction with rough set, IEEE Trans. Fuzzy Syst. 25 (2017) 825–838.
[59] Y.Q. Yao, J.S. Mi, Z.J. Li, A novel variable precision (θ, σ)-fuzzy rough set model based on fuzzy granules, Fuzzy Sets Syst. 236 (2014) 58–72.
[60] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Inf. Sci. 178 (2008) 3356–3373.
[61] A.P. Zeng, T.R. Li, D. Liu, J.B. Zhang, H.M. Chen, A fuzzy rough set approach for incremental feature selection on hybrid information systems, Fuzzy Sets Syst. 258 (2015) 39–60.
[62] X. Zhang, C.L. Mei, D.G. Chen, J.H. Li, Multi-confidence rule acquisition oriented attribute reduction of covering decision systems via combinatorial optimization, Knowl. Based Syst. 50 (2013) 187–197.
[63] X. Zhang, C.L. Mei, D.G. Chen, J.H. Li, Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[64] S.Y. Zhao, E.C.C. Tsang, D.G. Chen, The model of fuzzy variable precision rough sets, IEEE Trans. Fuzzy Syst. 17 (2009) 451–467.