Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 61–78

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy-rough feature selection accelerator ☆

Yuhua Qian [a,b,*], Qi Wang [a], Honghong Cheng [a], Jiye Liang [a], Chuangyin Dang [b]

[a] *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006, Shanxi, China*

[b] *Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong*

Received 6 August 2013; received in revised form 8 April 2014; accepted 10 April 2014

Available online 22 May 2014

Abstract

Fuzzy rough set method provides an effective approach to data mining and knowledge discovery from hybrid data including categorical values and numerical values. However, its time-consumption is very intolerable to analyze data sets with large scale and high dimensionality. Many heuristic fuzzy-rough feature selection algorithms have been developed however, quite often, these methods are still computationally time-consuming. For further improvement, we propose an accelerator, called forward approximation, which combines sample reduction and dimensionality reduction together. The strategy can be used to accelerate a heuristic process of fuzzy-rough feature selection. Based on the proposed accelerator, an improved algorithm is designed. Through the use of the accelerator, three representative heuristic fuzzy-rough feature selection algorithms have been enhanced. Experiments show that these modified algorithms are much faster than their original counterparts. It is worth noting that the performance of the modified algorithms becomes more visible when dealing with larger data sets.

Keywords: Rough sets; Fuzzy rough sets; Feature selection; Forward approximation; Accelerator; Granular computing

1. Introduction

There are many factors that motivate the inclusion of a feature selection step in a variety of fields, such as data mining, machine learning and pattern recognition, which addresses the problem of selecting those input features that are most predictive of a given outcome [30,33,34,41]. Databases expand quickly not only in the rows (objects) but also in the columns (features) nowadays [3]. In recent several years, big data analysis has become a new hot topic. For a task of data analysis, a given data set is called big data if it cannot be efficiently processed via existing methods. In some tasks of data analysis, some of features are irrelevant to the learning or problem solving. It is likely that the

☆ This is an extended version of the paper presented at 2011 International Conference of Rough Sets and Knowledge Technology, Banff, 2011, Canada.

* Corresponding author at: Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006, Shanxi, China. Tel./fax: +86 0351 7010566.

E-mail addresses: jinchengqyh@126.com (Y. Qian), wqking@163.com (Q. Wang), chhsxdx@163.com (H. Cheng), ljy@sxu.edu.cn (J. Liang), mecdang@cityu.edu.hk (C. Dang).

http://dx.doi.org/10.1016/j.fss.2014.04.029
0165-0114/

omission of some features will not seriously increase the error probability. In such cases, the loss of optimality may not only be tolerable but even desirable relative to the costs involved.

In the framework of rough set theory, feature selection is also called attribute reduction [8,43,44], which preserves the original meaning of the features after reduction [45]. The classical rough set model, proposed by Pawlak [22,23], is based on crisp equivalence relations and crisp equivalence classes. It is only applicable to categorical attribute reduction and knowledge discovery. In order to deal with numerical and categorical data (or a mixture of both) in data sets, fuzzy rough set model was first proposed by Dübois and Prade [5], which combines rough set and fuzzy set together. The lower/upper approximation in these fuzzy rough set models tries to give a membership function of each object to a set. As Dübois and Prade defined, if a fuzzy set is approached by a family of crisp sets in the same universe, then the corresponding lower/upper approximation pair is called a rough fuzzy set; and if a crisp/fuzzy set is approached by a family of fuzzy sets in the same universe, then the corresponding lower/upper approximation pair is called a fuzzy rough set. To widely apply the fuzzy rough set method, many extended versions and relative applications have been developed, cf. [11,12,14,20,21,24,25,31,35,36,39,40,42,46–48]. In particular, to keep the same form as classical rough set by Pawlak, Hu et al. [11] proposed a novel fuzzy rough model with a crisp lower/upper approximation. In fact, in the new model, the lower approximation and the upper approximation can be seen as the 1-cut/strong 0-cut of original counterparts in Dübois's model, respectively. Taking the same idea into account, Wang et al. [36] developed a generalized fuzzy rough model in which a β-cut is used to define its lower/upper approximation. These two methods have a consistent form with Pawlak's rough set, and their lower/upper approximations induced by a given cut are crisp approximations rather than fuzzy approximations. According to Dübois and Prade's definition, each of these rough set models is a fuzzy rough set.

Attribute reduction using fuzzy rough sets is often called fuzzy-rough feature selection. To support efficient feature selection, many heuristic algorithms have been developed in fuzzy rough set theory, cf. [2,4,10,11,13,15–17,36]. Each of these feature selection methods can extract a single reduct from a given decision table. For convenience, from the viewpoint of heuristic functions, we classify these feature selection methods into two categories: fuzzy positive region reduction and fuzzy information entropy reduction. Hence, we only review two kinds of representative heuristic fuzzy-rough feature selection methods.

(1) Fuzzy positive region reduction

The concept of positive region was proposed by Pawlak in [22], which is used to measure the significance of a condition attribute in a decision table. Then, Hu and Cercone [9] proposed a heuristic attribute reduction method, called positive region reduction, which remains the positive region of target decision unchanged. Under Dübois's fuzzy rough set model, Jensen and Shen [15–17] developed a series of heuristic fuzzy-rough feature selection algorithms based on fuzzy positive region. Bhatt and Gopal [2] proposed a modified version to improve computational efficiency. Under Hu's fuzzy rough set model and Wang's fuzzy rough set model, Hu et al. [11] extended the method from the literature [9] to select a feature subset from hybrid data. Owing to the consistency of ideas and strategies of these methods, we regard the method from [11] as their representative.

(2) Fuzzy information entropy reduction

The entropy reducts have first been introduced in 1993/1994 by Skowron in his lectures at Warsaw University. Wang et al. [37] used conditional entropy of Shannon's entropy to calculate the relative attribute reduction of a decision information system. Hu et al. extended the entropy to measure the information quantity in fuzzy sets and applied its conditional entropy to feature selection from hybrid data [13]. This reduction method remains the conditional entropy of a target decision unchanged. The fuzzy information entropy is an important approach to characterizing the uncertainty of a fuzzy binary relation, which can be used to select a feature subset from a given big data set [13,14].

Each of these above methods preserves a particular property of a given decision table. However, these above methods are still computationally very expensive, which are intolerable for dealing with large-scale data sets with high dimensions. So, this kind of attribute reduction problems can be regarded as data analysis of big data. The objective of this study is to focus on how to improve the time efficiency of a heuristic fuzzy-rough feature selection algorithm.

In a recent published paper in Artificial Intelligence, to overcome the shortcoming of computationally time-consuming of all heuristic attribute reduction algorithms, Qian et al. [27] proposed an accelerator for attribute reduction in rough set theory, which is based on a theoretic framework called positive approximation. Using the experience of the method for reference, in this paper, we wish to develop an extended version of the accelerator for accelerating fuzzy-rough feature selection. Its motivation is mainly caused by the three issues: (1) a fuzzy rough set including fuzzy-rough feature selection is a kind of very important models in rough set theory; (2) the accelerator proposed by Qian et al. cannot be used to accelerate feature selection for hybrid data but that for symbolic data; and (3) heuristic functions in fuzzy-rough feature selection are constructed by the membership of each object, which brings different methods of feature selection. Taking these three issues into account, one needs to develop an extended version of the accelerator for accelerating fuzzy-rough feature selection algorithms. The main advantage of this approach stems from the fact that the new accelerator can improve the time efficiency of a heuristic fuzzy-rough feature selection, which provides a vehicle of making algorithms of fuzzy-rough set based feature selection techniques faster. By incorporating the new accelerator into each of the above two kinds of representative heuristic attribute reduction methods, we construct their modified versions. Numerical experiments show that each of the modified methods can greatly reduce computing time while obtaining an attribute reduct. We would like to stress that the improvement becomes more profoundly visible when the data sets under discussion get larger.

The study is organized as follows. Several fuzzy rough set models are briefly reviewed in Section 2. In Section 3, we establish the forward approximation framework and investigate some of its main properties. In Section 4, we develop a modified attribute reduction algorithm based on the forward approximation. Experiments on six public data sets show that these modified algorithms are much faster than their original counterparts in terms of computational time. Finally, Section 5 concludes this paper by bringing some remarks and discussions.

2. Review on fuzzy rough set models

In this section, we review three representative fuzzy-rough set models and some related concepts.

Given a nonempty finite set U, $\widetilde{R}$ is a fuzzy binary relation over U, denoted by a matrix

$$M(\widetilde{R}) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}, \tag{1}$$

where $r_{ij} \in [0, 1]$ is the relation value between x_i and x_j. Some operations of relation matrices are defined as

1) $\widetilde{R}_1 = \widetilde{R}_2 \Leftrightarrow \widetilde{R}_1(x, y) = \widetilde{R}_2(x, y)$;
2) $\widetilde{R} = \widetilde{R}_1 \cup \widetilde{R}_2 \Leftrightarrow \widetilde{R} = max\{\widetilde{R}_1(x, y), \widetilde{R}_2(x, y)\}$;
3) $\widetilde{R} = \widetilde{R}_1 \cap \widetilde{R}_2 \Leftrightarrow \widetilde{R} = min\{\widetilde{R}_1(x, y), \widetilde{R}_2(x, y)\}$;
4) $\widetilde{R}_1 \subseteq \widetilde{R}_2 \Leftrightarrow \widetilde{R}_1(x, y) \leq \widetilde{R}_2(x, y)$.

If using the terms of granular computing, we denote the coarseness/fineness relationship between any two fuzzy binary relations by $\widetilde{R}_1 \preceq \widetilde{R}_2$, which is equivalent to $\widetilde{R}_1 \subseteq \widetilde{R}_2$ and $\widetilde{R}_1(x, y) \leq \widetilde{R}_2(x, y)$ for any x and y. It can be said that the fuzzy binary relation $\widetilde{R}_1$ is much finer than the fuzzy binary relation $\widetilde{R}_2$. Symmetrically, it also can be written as $\widetilde{R}_2 \succeq \widetilde{R}_1$.

The granular structure of the universe generated by a fuzzy binary relation $\widetilde{R}$ is defined as

$$\langle U, \widetilde{R} \rangle = \left([x_1]_{\widetilde{R}}, [x_2]_{\widetilde{R}}, \cdots, [x_n]_{\widetilde{R}}\right), \tag{2}$$

where $[x_i]_{\widetilde{R}} = r_{i1}/x_1 + r_{i2}/x_2 + \cdots + r_{in}/x_n$. $[x_i]_{\widetilde{R}}$ is the fuzzy neighborhood of x_i and r_{ij} is the degree of x_i equivalent to x_j. Here, "+" means the union of elements. The cardinality of $[x_i]_{\widetilde{R}}$ can be calculated with

$$\left|[x_i]_{\widetilde{R}}\right| = \sum_{j=1}^{n} r_{ij}, \tag{3}$$

which appears to be a natural generalization of the cardinality of a crisp set.

In this case, $[x_i]_{\widetilde{R}}$ is a fuzzy set and the family of $[x_i]_{\widetilde{R}}$ forms a fuzzy concept system of the universe. This system will be used to approximate the object subset of the universe.

Let $\widetilde{X}$ be a fuzzy set. Then, it can be represented as

$$\widetilde{X} = \mu_{\widetilde{X}}(x_1)/x_1 + \mu_{\widetilde{X}}(x_2)/x_2 + \cdots + \mu_{\widetilde{X}}(x_n)/x_n, \tag{4}$$

where $\mu_{\widetilde{X}}(x_j)$ denotes the membership degree of the object x_j in $\widetilde{X}$.

It is well known that, a categorical attribute can induce a crisp equivalence relation on the universe and generate a family of crisp information granules, whereas a numerical attribute will give a fuzzy binary relation and form a set of fuzzy information granules [13]. As crisp information granules are a special case of fuzzy ones, we will consider all of them as fuzzy ones in the following. Given an information system $S = (U, C \cup D)$, $B, B_1, B_2 \subseteq C$. We mean $\widetilde{R}_B$ as the fuzzy binary relation induced by the attribute subset B. Then we have

1) $\widetilde{R}_B = \bigcap_{a\in B} \widetilde{R}_a$;
2) $\widetilde{R}_{B_1\cup B_2} = \widetilde{R}_{B_1} \cap \widetilde{R}_{B_2}$.

The first fuzzy rough set model was introduced by Dübois and Prade [5]. By their definition, a universe of objects $U = \{x_1, x_2, \ldots, x_n\}$ is described by a fuzzy binary relation $\widetilde{R}$. Given $X \subseteq U$ a crisp subset of objects, the memberships of an object x_i in a fuzzy rough set $(\underline{\widetilde{R}}(X), \overline{\widetilde{R}}(X))$ of fuzzy sets on U are described as

$$\begin{cases} \mu_{\underline{\widetilde{R}}(X)}(x_i) = \inf\limits_{x_j\in U} \max\{1 - \widetilde{R}(x_i, x_j), \mu_X(x_j)\}, \\ \mu_{\overline{\widetilde{R}}(X)}(x_i) = \sup\limits_{x_j\in U} \min\{\widetilde{R}(x_i, x_j), \mu_X(x_j)\}, \end{cases}$$

where U is a nonempty universe and R is a fuzzy binary relation on U.

To keep the same form as classical rough set model, Wang et al. [36] proposed a new fuzzy rough set model (for simplification, called Wang's fuzzy rough set), which is explicitly expressed as follows.

Let $\langle U, \widetilde{R}\rangle$ be a fuzzy approximation space, and $X \subseteq U$ a crisp subset of objects. Wang's fuzzy lower and upper approximation of X can be defined as

$$\begin{cases} \underline{\widetilde{R}}_\beta(X) = \{x_i \in X \mid \widetilde{R}(x_i, x_j) \le 1 - \beta,\ \forall x_j \in U - X\}, \\ \overline{\widetilde{R}}_\beta(X) = \{x_i \in U \mid \exists x_j \in X,\ \text{such that } \widetilde{R}(x_i, x_j) \ge \beta\}, \end{cases}$$

where $\widetilde{R}(x_i, x_j)$ is the similarity degree between x_i and x_j with respect to $\widetilde{R}$. The order pair $\langle \underline{\widetilde{R}}_\beta \widetilde{X}, \overline{\widetilde{R}}_\beta \widetilde{X}\rangle$ is called a β-fuzzy rough set, in which a β-cut is used to define its lower/upper approximation.

In a recent paper, Hu et al. [13] gave another definition of a fuzzy rough set in the context of hybrid data (for simplification, called Hu's fuzzy rough set), which is shown as follows.

Let $\langle U, \widetilde{R}\rangle$ be a fuzzy approximation space and $\widetilde{X}$ a fuzzy subset of U. The lower approximation $\underline{\widetilde{R}}\widetilde{X}$ and upper approximation $\overline{\widetilde{R}}\widetilde{X}$ are defined as [11]

$$\begin{cases} \underline{\widetilde{R}}\widetilde{X} = \{x_i \mid [x_i]_{\widetilde{R}} \subseteq \widetilde{X},\ x_i \in U\}, \\ \overline{\widetilde{R}}\widetilde{X} = \{x_i \mid [x_i]_{\widetilde{R}} \cap \widetilde{X} \ne \emptyset,\ x_i \in U\}, \end{cases}$$

where $[x_i]_{\widetilde{R}} \subseteq \widetilde{X}$ means $\mu_{[x_i]_{\widetilde{R}}}(x_i) \le \mu_{\widetilde{X}}(x_i)$, and $[x_i]_{\widetilde{R}} \cap \widetilde{X} \ne \emptyset$ implies that $min\{\mu_{[x_i]_{\widetilde{R}}}(x_i), \mu_{\widetilde{X}}(x_i)\} \ne 0$, $\emptyset = \{\frac{0}{x_1} + \frac{0}{x_2} + \cdots + \frac{0}{x_n}\}$. The order pair $\langle \underline{\widetilde{R}}\widetilde{X}, \overline{\widetilde{R}}\widetilde{X}\rangle$ is called a fuzzy rough set. In fact, in the new model, the lower approximation and the upper approximation can be seen as the 1-cut/strong 0-cut of original counterparts in Dübois's model, respectively.

It is easy to see that given a lower approximation in Dübois's fuzz rough set model, one easily obtains the corresponding crisp lower approximation with an α-cut, and given an upper approximation in Dübois's fuzzy rough set model, one also easily gets its crisp upper approximation with a β-cut. Hence, one can obtain crisp approximation of an object set according to a user's requirement. Without loss of generality, we will select Hu's fuzzy rough model as their representative in this study.

A decision table is an information system $S = (U, C \cup D)$, where C is called a condition attribute set and D is called a decision attribute set [11]. In practical decision-making issues, in general, the decision attribute set D can induce an equivalence partition, i.e., a crisp classification. In this paper, we only focus on this kind of decision tables. Assume the objects are partitioned into r mutually exclusive crisp subsets $\{Y_1, Y_2, \cdots, Y_r\}$ by the decision attribute D. Given a decision table $S = (U, C \cup D)$ and a subset $B \subseteq C$, and $\widetilde{R}_B$ the fuzzy similarity relation induced by B, one can define the lower and upper approximations of the decision attribute D as

$$\begin{cases} \underline{\widetilde{R}_B} D = \{\underline{\widetilde{R}_B} Y_1, \underline{\widetilde{R}_B} Y_2, \cdots, \underline{\widetilde{R}_B} Y_r\}, \\ \overline{\widetilde{R}_B} D = \{\overline{\widetilde{R}_B} Y_1, \overline{\widetilde{R}_B} Y_2, \cdots, \overline{\widetilde{R}_B} Y_r\}. \end{cases}$$

Denoted by $POS_B(D) = \bigcup_{i=1}^{r} \underline{\widetilde{R}_B} Y_i$, it is called the positive region of D with respect to the condition attribute set B. The crisp positive region in traditional rough set theory is defined as the union of the lower approximations. In the original fuzzy rough set, the membership of an object belonging to the fuzzy positive region can be defined by

$$\mu_{POS_B(D)}(x) = \sup_{X \in U/D} \mu_{\underline{B}(X)}(x),$$

where $\mu_{\underline{B}(X)}(x) = \inf_{x_j \in U} \max\{1 - \widetilde{B}(x, x_j), \mu_X(x_j)\}$.

In the above three fuzzy rough set models, many efficient fuzzy-rough feature selection algorithms have been developed [2,10,11,13,15–17,36]. However, these algorithms are still computationally very expensive, which is intolerable for dealing with large-scale data sets with high dimensions. The objective of this study is to focus on how to improve the time efficiency of a heuristic fuzzy-rough feature selection algorithm. In a recent published paper in Artificial Intelligence, to overcome the shortcoming of computationally time-consuming of all heuristic attribute reduction algorithms, Qian et al. [27] proposed an accelerator for attribute reduction in rough set theory, which is performed on a gradually reduced universe. Using the experience of the method for reference, in this paper, we wish to develop an extended version of the accelerator for accelerating fuzzy-rough feature selection. From the point of view, in next study, we decide to use three representative fuzzy-rough feature selection algorithms for explaining and verifying the mechanism and efficiency of the extended accelerator.

3. Forward approximation (FA): an accelerator to fuzzy-rough feature selection

In this section, we introduce a new set-approximation approach called forward approximation and investigate some of its important properties, in which a given set (also called a target concept in rough set theory) is approximated by a forward granulation world [28]. Given a decision table $S = (U, C \cup D)$, $U/D = \{Y_1, Y_2, \cdots, Y_r\}$ is called a target decision, in which each equivalence class Y_i $(i \leq r)$ can be regarded as a target concept. These concepts and properties will be helpful to understand the notion of a granulation order and set approximation under a granulation order.

Definition 1. Let $P = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_n\}$ be a family of fuzzy binary relations with $\widetilde{R}_1 \succeq \widetilde{R}_2 \succeq \cdots \succeq \widetilde{R}_n$, and X a crisp set. Given $P_i = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_i\}$, we define P_i-lower approximation $\underline{P_i}(X)$ and P_i-upper approximation $\overline{P_i}(X)$ of P_i-positive approximation of X as

$$\begin{cases} \underline{P_i}(X) = \bigcup_{k=1}^{i} \underline{\widetilde{R}_k} X_k, \\ \overline{P_i}(X) = \overline{\widetilde{R}_i} X, \end{cases}$$

where $X_1 = X$ and $X_k = X - \bigcup_{j=1}^{k-1} \underline{\widetilde{R}_j} X_j$, $k = 2, 3, \cdots, n$, $i = 1, 2, \cdots, n$.

Correspondingly, the boundary of X is given as

$$BN_{P_i}(X) = \overline{P_i}(X) - \underline{P_i}(X).$$

Theorem 1. *Let $P = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_n\}$ be a family of fuzzy binary relations with $\widetilde{R}_1 \succeq \widetilde{R}_2 \succeq \cdots \succeq \widetilde{R}_n$, and X a crisp set. Given $P_i = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_i\}$, then $\forall P_i$ $(i = 1, 2, \cdots, n)$, we have*

$$\underline{P_i}(X) \subseteq X \subseteq \overline{P_i}(X),$$
$$\underline{P_1}(X) \subseteq \underline{P_2}(X) \subseteq \cdots \subseteq \underline{P_i}(X).$$

Theorem 2. *Let $P = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_n\}$ be a family of fuzzy binary relations with $\widetilde{R}_1 \succeq \widetilde{R}_2 \succeq \cdots \succeq \widetilde{R}_n$, and X a crisp set. Given $P_i = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_i\}$, then $\forall P_i$ $(i = 1, 2, \cdots, n)$, we have*

$$\alpha_{P_1}(X) \leq \alpha_{P_2}(X) \leq \cdots \leq \alpha_{P_i}(X),$$

where $\alpha_{P_i}(X) = \frac{|\underline{P_i}(X)|}{|\overline{P_i}(X)|}$ is the approximation measure of X with respect to P_i.

Definition 2. Let $P = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_n\}$ be a family of fuzzy binary relations with $\widetilde{R}_1 \succeq \widetilde{R}_2 \succeq \cdots \succeq \widetilde{R}_n$ and $U/D = \{Y_1, Y_2, \cdots, Y_r\}$. Lower approximation and upper approximation of D with respect to P_i are defined as

$$\begin{cases} \underline{P_i} D = \left\{\underline{P_i}(Y_1), \underline{P_i}(Y_2), \cdots, \underline{P_i}(Y_r)\right\}, \\ \overline{P_i} D = \left\{\overline{P_i}(Y_1), \overline{P_i}(Y_2), \cdots, \overline{P_i}(Y_r)\right\}. \end{cases}$$

$\underline{P_i} D$ is also called the positive region of D with respect to the granulation order P_i, denoted by $POS^U_{P_i}(D) = \bigcup_{k=1}^{r} \underline{P_i} Y_k$.

Theorem 3 *(Recursive expression principle). Let $P = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_n\}$ be a family of fuzzy binary relations with $\widetilde{R}_1 \succeq \widetilde{R}_2 \succeq \cdots \succeq \widetilde{R}_n$ and $U/D = \{Y_1, Y_2, \cdots, Y_r\}$. Given $P_i = \{\widetilde{R}_1, \widetilde{R}_2, \cdots, \widetilde{R}_i\}$, we have*

$$POS^U_{P_{i+1}}(D) = POS^U_{P_i}(D) \cup POS^{U_{i+1}}_{\widetilde{R}_{i+1}}(D),$$

where $U_1 = U$ and $U_{i+1} = U - POS^U_{P_i}(D)$.

The dependency function is used to characterize the dependency degree of an attribute subset with respect to a given decision [7,8,26,29]. Given a decision table $S = (U, C \cup D)$, the dependency function of condition attribute set C with respect to the decision attribute set D is formally defined as $\gamma_C(D) = |POS^U_C(D)|/|U|$. Using this notation, we give the definition of dependency function of a granulation order P with respect to D in the following.

Definition 3. A dependency function involving a granulation order P and D is defined as

$$\gamma_P(D) = \frac{|POS^U_P(D)|}{|U|},$$

where $|\cdot|$ denotes the cardinality of a set and $0 \leq \gamma_P(D) \leq 1$.

The dependency function reflects the granulation order P's power to dynamically approximate D. This dependency function can be used to measure the significance of attributes relative to the decision and construct a heuristic function for designing an attribute reduction algorithm.

4. Fuzzy-rough feature selection based on forward approximation

4.1. Fuzzy-rough feature selection algorithms

In fuzzy-rough feature selection, to support efficient attribute reduction, many heuristic attribute reduction methods have been developed, in which a forward greedy search strategy is usually employed, cf. [2,10,11,13,15–17,36]. In this kind of attribute reduction approach, important measures of attributes are used for heuristic functions, which can be used in a forward feature selection. It is deserved to point out that each kind of attribute reduction tries to preserve a particular property of a given decision table.

In each forward greedy attribute reduction approach, we take the attribute with the maximal significance into the attribute subset in each loop until this feature subset satisfies the stopping criterion, and then we can get an attribute

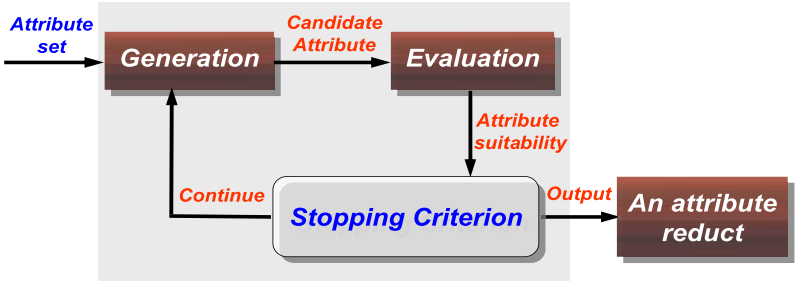


Fig. 1. The process of forward greedy attribute reduction algorithm.

reduct. In this algorithm framework, we denote the evaluation function (stop criterion) by $EF^U(B, D) = EF^U(C, D)$. For example, if one adopts Shannon's conditional entropy, then the evaluation function is $H^U(B, D) = H^U(C, D)$. That is to say, if $EF^U(B, D) = EF^U(C, D)$, then B is said to be an attribute reduct. Formally, a forward greedy attribute reduction algorithm can be written as follows.

Algorithm 1 A general forward greedy attribute reduction algorithm.

Input: Decision table $S = (U, C \cup D)$;
Output: One reduct *red*.
Step 1: $red \leftarrow \emptyset$; //*red* is the pool to conserve the selected attributes
Step 2: While $EF(red, D) \neq EF(C, D)$ Do //This provides a stopping criterion.
{
$B \leftarrow C - red$,
Select $a_0 \in B$ which satisfies $Sig(a_0, red, D, U) = max\{Sig(a_k, red, D, U), a_k \in B\}$,
If $Sig(a_0, red, D, U) > 0$, then $red \leftarrow red \cup \{a_0\}$
};
Step 3: Return *red* and end.

This algorithm can obtain an attribute reduct from a given decision table. Fig. 1 displays the process of attribute reduction based on the forward greedy attribute reduction algorithm in rough set theory, which is helpful for more clearly understanding of the mechanism of the algorithm.

4.2. Three representative significance measures of attributes

For efficient attribute reduction, many heuristic attribute reduction methods have been developed in fuzzy rough set theory, see [2,10,11,13,15–17,36]. For convenience, as was pointed out in the introduction part of this paper, we only focus on the three representative fuzzy-rough feature selection methods here.

Given a decision table $S = (U, C \cup D)$, one can obtain $\langle U, \widetilde{R}_C \rangle = ([x_1]_{\widetilde{R}_C}, [x_2]_{\widetilde{R}_C}, \cdots, [x_n]_{\widetilde{R}_C})$ and the decision $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. Through these notations, in what follows we review three representative significance measures of attributes.

For attribute reduction, Hu and Cercone [9] proposed a heuristic attribute reduction method, called positive region reduction (PR), which remains the positive region of target decision unchanged. Hu, Xie and Yu [11] extended this method to fuzzy-rough feature selection, called fuzzy positive region reduction (FPR). In this method, the significance measures of attributes are defined as follows.

Definition 4. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$$Sig_1(a, B, D, U) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D),$$

where $\gamma_B(D) = \frac{|POS_B^U(D)|}{|U|} = \frac{|\bigcup_{i=1}^r \underline{\widetilde{R}_B} Y_i|}{|U|}$.

Shannon's information entropy [32] was introduced to search reducts in classical rough set model. In fact, several authors also have used variants of Shannon's entropy to measure uncertainty in rough set theory and construct heuris-

tic algorithm of attribute reduction [1,6,18,19,31,38]. Wang et al. used its conditional entropy to calculate the relative attribute reduction of a decision information system [37]. Hu, Xie and Yu [13] proposed a so-called fuzzy information entropy to fuzzy rough set model and used its fuzzy conditional entropy to design a heuristic feature selection algorithm. This reduction method remains the fuzzy conditional entropy of target decision unchanged, denoted by FSCE, in which the fuzzy conditional entropy reads as

$$H(D|B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}.$$

Using the fuzzy conditional entropy, the definitions of the significance measures are expressed in the following way.

Definition 5. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$$Sig_2(a, B, D, U) = H(D|B) - H\big(D|B \cup \{a\}\big).$$

In the original fuzzy rough set, Jensen and Shen [16] extended this method to fuzzy-rough feature selection. In this method, the significance measures of attributes can be formally written as follows.

Definition 6. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$$Sig_3(a, B, D, U) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D),$$

where $\gamma_B(D) = \frac{|POS_B^U(D)|}{|U|} = \frac{\sum_{x\in U}\mu_{POS_B^U(D)}(x)}{|U|}$.

In a heuristic fuzzy-rough feature selection algorithm, based on the above definitions, one can find an attribute reduct by gradually adding selected attributes.

4.3. Rank preservation principle

As mentioned above, each of significance measures of attributes provides some heuristics to guide the mechanism of forward searching a feature subset. Unlike the discernibility matrix, the computational time of the heuristic algorithms has been largely reduced when only one attribute reduct is needed. Nevertheless, these algorithms still could be very time consuming. To introduce an improved strategy of heuristic attribute reductions, we concentrate on the rank preservation of the four significance measures of attributes based on the positive approximation encountered in a decision table.

Firstly, we investigate the rank preservation of significance measures of attributes based on the dependency measure. For more clear representation, we denote the significance measure of an attribute by $Sig_\Delta^{outer}(a, B, D, U)$ ($\Delta = \{1, 2, 3, 4\}$), which denotes the value of the significance measure on the universe U. One can prove the following theorem of rank preservation.

Theorem 4. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS_B^U(D)$. For $\forall a, b \in C - B$, if $Sig_1(a, B, D, U) \geq Sig_1(b, B, D, U)$, then $Sig_1(a, B, D, U') \geq Sig_1(b, B, D, U')$.*

Secondly, we research the rank preservation of significance measures of attributes based on the Shannon's conditional entropy. The following theorem elaborates on the rank preservation of this measure.

Theorem 5. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS_B^U(D)$. For $\forall a, b \in C - B$, if $Sig_2(a, B, D, U) \geq Sig_2(b, B, D, U)$, then $Sig_2(a, B, D, U') \geq Sig_2(b, B, D, U')$.*

Finally, we research the rank preservation of significance measures of attributes based on the original fuzzy positive region defined by Jensen and Shen [16]. The following theorem elaborates on the rank preservation of this measure.

Theorem 6. *Let* $S = (U, C \cup D)$ *be a decision table,* $B \subseteq C$ *and* $U' = U - \{x | \mu_{POS_B^U(D)}(x) = 1, x \in U\}$. *For* $\forall a, b \in C - B$, *if* $Sig_3(a, B, D, U) \geq Sig_3(b, B, D, U)$, *then* $Sig_3(a, B, D, U') \geq Sig_3(b, B, D, U')$.

From these theorems, one can see that the rank of attributes in the process of attribute reduction will remain unchanged after reducing the lower approximation of positive approximation. This mechanism can be used to improve the computational performance of a heuristic attribute reduction algorithm, while retaining the same selected feature subset.

4.4. Accelerated versions

The objective of rough set-based feature selection is to find a subset of attributes which retains some particular properties as the original data and without redundancy. In fact, there may be multiple reducts for a given decision table. It has been proven that finding the minimal reduct of a decision table is an NP hard problem. When only one attribute reduct is needed, based on the significance measures of attributes, some heuristic algorithms have been proposed, most of which are greedy and forward search algorithms. These search algorithms start with a nonempty set, and keep adding one or several attributes of high significance into a pool each time until the dependence has not been increased.

From the discussion in the previous subsection, we can construct an improved forward search algorithm based on the forward approximation, which is formulated as follows.

Algorithm Q3 An improved feature selection algorithm based on the forward approximation (FA).

Input: Decision table $S = (U, C \cup D)$;
Output: One feature subset *red*.
Step 1: $red \leftarrow \emptyset$, $i \leftarrow 1$, $R_1 \leftarrow red$, $P_1 \leftarrow \{R_1\}$ and $U_1 \leftarrow U$; //*red* is the pool to conserve the selected attributes
Step 2: While $EF(red, D) \neq EF(C, D)$ Do //This provides a stopping criterion.
{
Compute the positive region of forward approximation $POS_{P_i}^U(D)$,
$U_{i+1} \leftarrow U - POS_{P_i}^U(D)$,
$i \leftarrow i + 1$,
$B \leftarrow C - red$,
Select $a_0 \in B$ which satisfies $Sig(a_0, red, D, U_i) = max\{Sig(a_k, red, D, U_i), a_k \in B\}$,
If $Sig(a_0, red, D, U_i) > 0$, then $red \leftarrow red \cup \{a_0\}$,
$R_i \leftarrow R_i \cup \{a_0\}$,
$P_i \leftarrow \{R_1, R_2, \cdots, R_i\}$;
}
Step 3: Return *red* and end.

It deserves to point out that the feature subset obtained by Algorithm Q3 from a given data set may not be a reduct as commonly used in rough set literature. The result of Algorithm Q3 may still contain some superfluous attributes.

Computing the significance measure of an attribute $Sig^{inner}(a_k, C, D, U)$ is one of the key steps in FA, whose time complexity is $O(|C||U|^2)$. In Step 2, we begin with the empty set and add an attribute with the maximal significance into the set in each stage until finding a reduct. This process is called a forward reduction algorithm whose time complexity is $O(\sum_{i=1}^{|C|}(|C| - i + 1)|U_i|^2)$. However, the time complexity of the original heuristic algorithm is $O(\sum_{i=1}^{|C|}(|C| - i + 1)|U|^2)$. Obviously, the time complexity of FA is much lower than that of each of classical heuristic attribute reduction algorithms. Hence, one can draw a conclusion that the modified feature selection algorithm based on the forward approximation (FA) may significantly reduce the computational time for fuzzy-rough feature selection.

To support the substantial contribution of the improved attribute reduction algorithm based on the forward approximation, we summarize two factors of speedup of this accelerator as follows.

(1) One only reserves a much smaller similarity matrix in each iterative loop via gradually decreasing the size of data set. This is an important factor of the improved algorithm.
(2) Computational time of significance measures of attributes is significantly reduced, which is because that it is only considered on the gradually reduced universe. It is the other factor of the accelerated algorithm.

Table 1
Data sets description

	Data sets	Samples	Features	Classes
1	Image Segmentation	2310	19	7
2	Sonar, mines vs. rocks	208	60	2
3	Wisconsin diagnostic breast cancer (Cancer1)	569	30	2
4	Ionosphere	351	34	2
5	Wisconsin prognostic breast cancer (Cancer2)	198	33	2
6	Wine recognition	178	13	3

Based on the above two speedup factors, we draw such a conclusion that: the modified algorithm can significantly reduce the computational time of each existing attribute reduction algorithm.

4.5. Time efficiency analysis of algorithms

Some heuristic attribute reduction methods have been developed for hybrid data, cf. [2,10,11,13,15–17,36]. The three heuristic algorithms mentioned in Section 4.2 are very representative. The objective of the following experiments is to show the performance of time reduction of the proposed framework for selecting a feature subset. The data used in the experiments are outlined in Table 1, which were all downloaded from UCI Repository of machine learning databases.

For numeric data, we normalize the numerical attribute a into the interval [0, 1] with

$$a' = \frac{a - a_{min}}{a_{max} - a_{min}}.$$

The value of the fuzzy similarity degree r_{ij} between objects x_i and x_j with respect to numerical attribute a is computed as

$$r_{ij} = \begin{cases} 1 - 4 \times |x_i - x_j|, & |x_i - x_j| \leq 0.25, \\ 0, & \text{otherwise.} \end{cases}$$

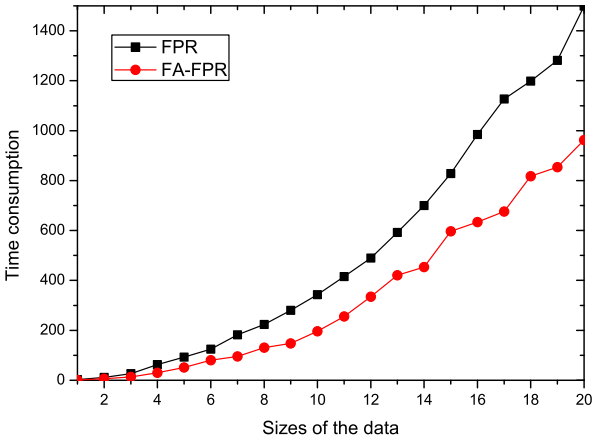
As $r_{ij} = r_{ji}$ and $r_{ii} = 1$, $0 \leq r_{ij} \leq 1$, the matrix $M = (r_{ij})_{n \times n}$ is a fuzzy similarity relation.

From the definition of attribute reduction based on fuzzy rough sets, we know that each modified attribute reduction algorithm must select an attribute reduct from original attributes. Therefore, in the following experiments, we only consider attribute reducts obtained and computational time.

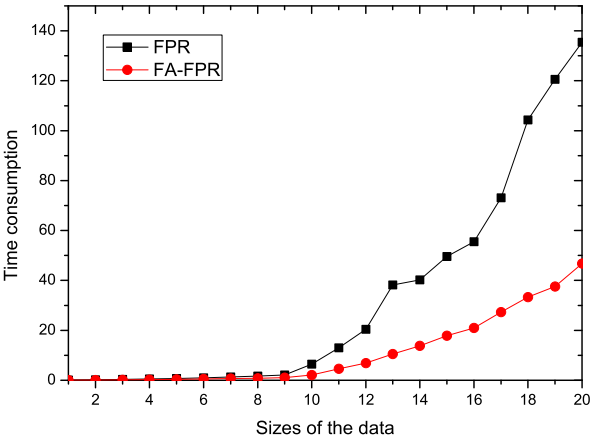
In what follows, we apply each of the original algorithms along with its modified version for searching attribute reducts. To distinguish the computational times, we divide each of these nine data sets into twenty parts of equal size. The first part is regarded as the 1st data set, the combination of the first part and the second part is viewed as the 2nd data set, the combination of the 2nd data set and the third part is regarded as the 3rd data set, $\cdots$, the combination of all twenty parts is viewed as the 20th data set. These data sets can be used to calculate time used by each of the original attribute reduction algorithms and the corresponding modifications and show it vis-a-vis the size of universe. These algorithms are run on a personal computer with Windows XP and Inter(R) Core(TM)2 Quad CPU Q9400, 2.66 GHz and 3.37 GB memory. The software being used is Microsoft Visual Studio 2005 and Visual *C#*.

4.5.1. FPR and FA-FPR

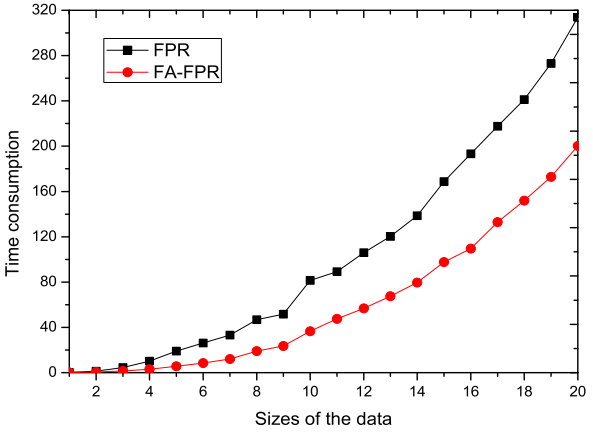
In the sequence of experiments, we compare FPR with FA-FPR on the six real world data sets shown in Table 1. The experimental results of these six data sets are shown in Table 2 and Fig. 2. In each of these sub-figures, the x-coordinate pertains to the size of the data set (the 20 data sets starting from the smallest one), while the y-coordinate concerns the computing time. Table 2 shows the comparisons of selected features and computational time with original algorithm FPR and the accelerated algorithm FA-FPR on six data sets. While Fig. 2 displays more detailed change trend of each of two algorithms with size of data set becoming increasing.



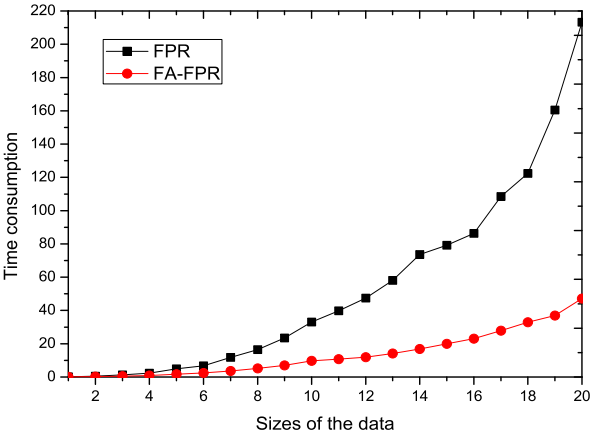
(a) Image Segmentation



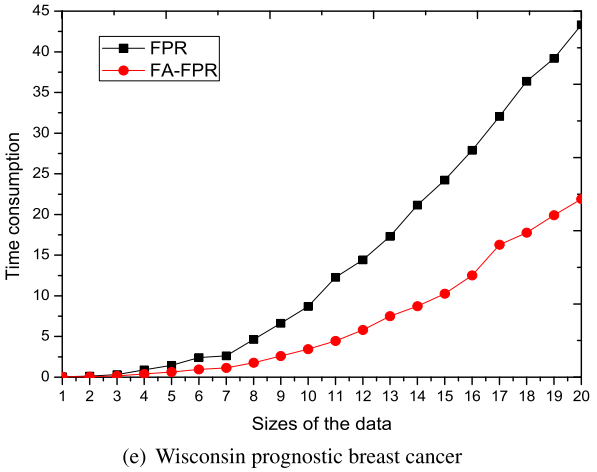
(b) Sonar, mines vs. rocks



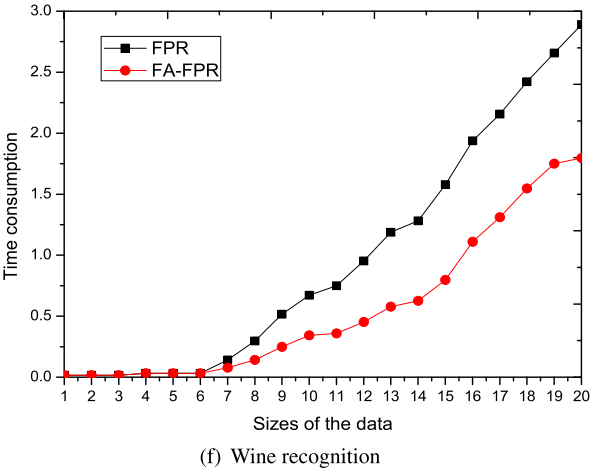
(c) Wisconsin diagnostic breast cancer



(d) Ionosphere



(e) Wisconsin prognostic breast cancer



(f) Wine recognition

Fig. 2. Times of FPR and FA-FPR versus the size of data.

Table 2
The time and attribute reduction of the algorithms FPR and FA-FPR.

Data sets	Original features	FPR algorithm		FA-FPR algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Image Segmentation	19	15	1499.2031	15	962.3594
Sonar, mines vs. rocks	60	20	135.4218	20	46.7187
Cancer1	30	22	313.7968	22	200.1875
Ionosphere	34	24	213.2187	24	47.1250
Cancer2	33	24	43.3281	24	21.8906
Wine recognition	13	13	2.8906	13	1.7968

Table 3
The time and attribute reduction of the algorithms FSCE and FA-FSCE.

Data sets	Original features	FSCE algorithm		FA-FSCE algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Image Segmentation	19	17	1258.0468	17	900.5781
Sonar, mines vs. rocks	60	41	300.5625	41	50.0000
Cancer1	30	27	228.9218	27	171.8750
Ionosphere	34	24	137.0468	24	43.9218
Cancer2	33	29	44.6562	29	26.8593
Wine recognition	13	13	2.8906	13	2.0937

It is easy to note from Table 2 and Fig. 2 that the computing time of each of these two algorithms increases with the increase of the size of data. As one of the important advantages of the FA, as shown in Table 2 and Fig. 2, we see that the modified algorithms are much more faster than their original counterparts on the basis of obtaining an attribute reduct. Sometimes, the effect of this reduction can reduce over two thirds of the computational time. For example, the reduced time achieves 88.7032 seconds on the data set (Sonar, mines vs. rock), while the reduced time is 166.0938 seconds on the data set (Ionosphere). Furthermore the differences are profoundly larger when the size of the data set increases.

4.5.2. FSCE and FA-FSCE

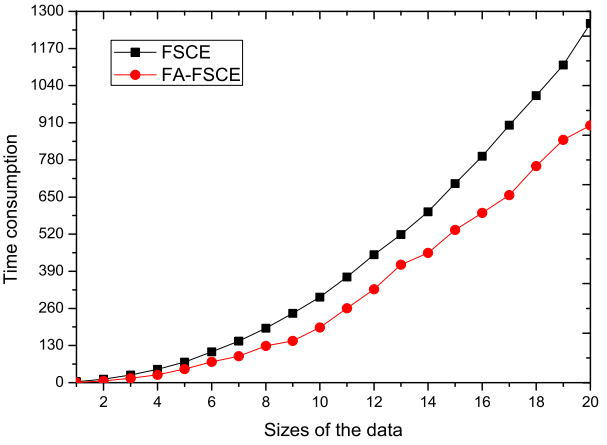
It is well known that, the attribute reduct induced by fuzzy information entropy keeps the fuzzy condition entropy of original data set, which is based on a more strict definition of attribute reduct. Hence, the attribute reduct obtained by this approach is often much longer than the one induced by the fuzzy positive region reduction.

In what follows, we compare FSCE with FA-FSCE on those six real world data sets shown in Table 1 from computational time and selected feature subsets. Table 3 presents the comparisons of selected features and computational time with original algorithm FSCE and the accelerated algorithm FA-FSCE on six data sets, while Fig. 3 gives more detailed change trendline of each of two algorithms with size of data set becoming increasing.

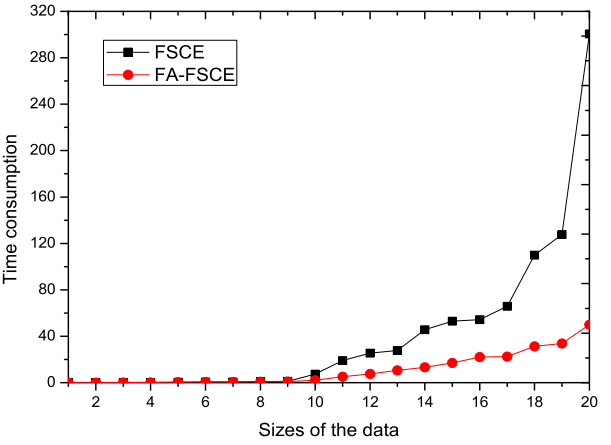
From Table 3 and Fig. 3, it is easy to see that the modified algorithms are consistently faster than their original counterparts. Sometimes, the reduced time can almost achieve five-fifths of the original computational time. For example, the reduced time achieves 250.5625 seconds on the data set (Sonar, mines vs. rocks), and the reduced time achieves 93.1250 seconds on the data set (Ionosphere). Furthermore the differences are profoundly larger when the size of the data set increases. Hence, attribute reduction based on the accelerator should be a good solution.

In addition, the fuzzy-rough feature selection algorithm induced by the original fuzzy rough set also can be correspondingly modified. Similar to Sections 4.5.1 and 4.5.2, its improved version is also much faster than the original one. Hence, we omit its relative experimental analysis.

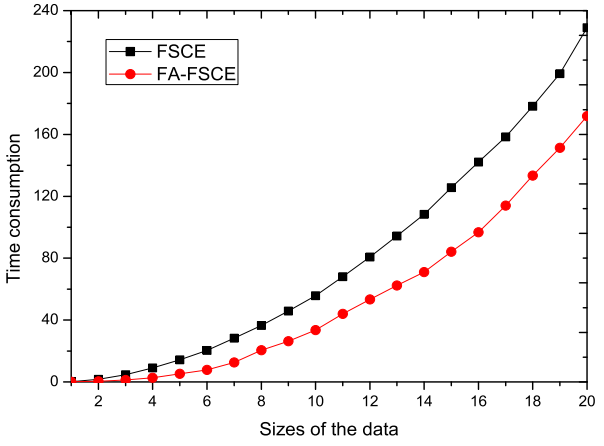
Remark. In rough set literature, there are three main control structures for constructing an attribute reduct with a heuristic strategy [40]. They include addition, deletion, and addition+deletion. The proposed method in this study is successful for the addition strategy only. In fact, how to accelerate those feature selection algorithms with deletion



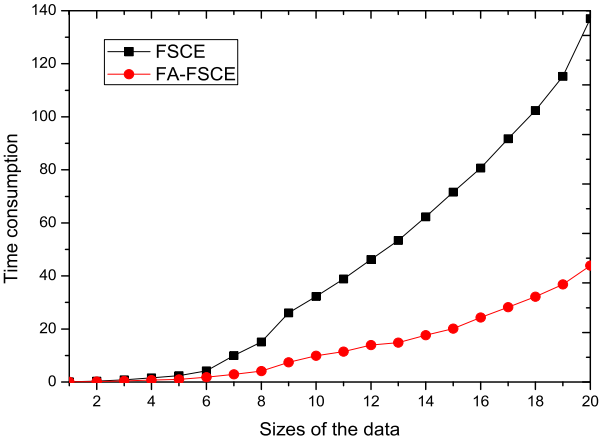
(a) Image Segmentation



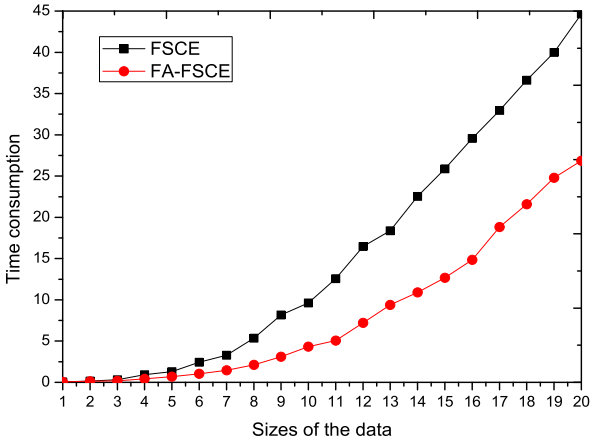
(b) Sonar, mines vs. rocks



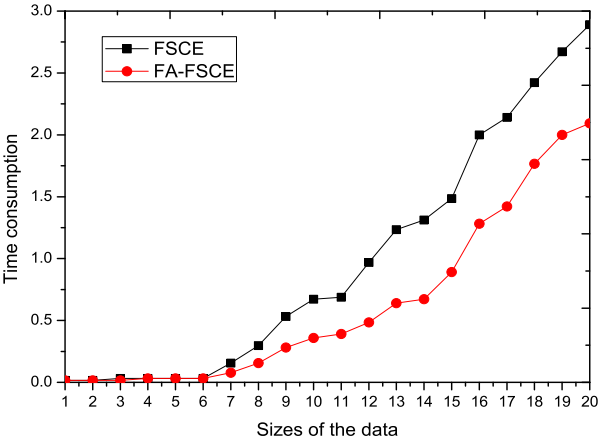
(c) Wisconsin diagnostic breast cancer



(d) Ionosphere



(e) Wisconsin prognostic breast cancer



(f) Wine recognition

Fig. 3. Times of FSCE and FA-FSCE versus the size of data.

strategy and addition+deletion strategy are also very interesting issues. However, these are beyond the scope of this study. We will address them in future work.

5. Conclusions

In this study, a theoretic framework based on rough set theory has been proposed, called the forward approximation, which can be used to accelerate algorithms of heuristic attribute reduction. Based on this framework, an improved heuristic feature selection algorithm (FSPA) has been presented. Several representative heuristic attribute reduction algorithms encountered in rough set theory have been revised and modified. Experimental studies pertaining to six UCI data sets show that the modified algorithms can significantly reduce computing time of attribute reduction. The results show that the attribute reduction based on the forward approximation is an effective accelerator and can efficiently obtain an attribute reduct.

In the conclusion section, we summarize the advantages of the accelerator-forward approximation for attribute reduction and offer some explanatory comments. Based on the theoretical analysis and experimental evidence, we can affirm that:

- From the stop criterion of the algorithm, it follows that one must obtain an attribute reduct of the decision table. This provides a restriction of keeping the approximation ability of the decision.

From the definition of each of attribute reduction using fuzzy rough sets and the stop criterion of the algorithm, one can know that one must obtain an attribute reduct of a given decision table when the algorithm is stopped. Hence, each of the accelerated algorithms does not affect the approximation ability of the attribute reduct induced by the corresponding method.

- Each of the accelerated algorithms usually comes with a substantially reduced computing time when compared with amount of time used by the corresponding original algorithm.

Through using the accelerator-forward approximation, the size of data set could be reduced in each loop of each of modified algorithms. Therefore, the computational time for determining similarity matrix and significance measures of attributes in the reduced data set would be much smaller than that encountered for the entire data set. Evidently, these modified algorithms are much faster than the previous methods for the time consumption.

- The performance of these modified algorithms is getting better in presence of larger data sets; the larger the data set, the more profound computing savings.

The stopping criterion of attribute reduction will be stricter when the data set becomes larger, and the number of attributes in the reduct induced by a heuristic attribute reduction algorithm usually is much bigger. In this situation, each of the modified algorithms can delete much more objects from the data set in all loops, and hence can take far less time for attribute reduction. The greater the size of the data set is, the larger the number of attributes selected, and the better the performance of these modified algorithms becomes when it comes to computing time. Hence, these accelerated algorithms are particularly suitable for dealing with attribute reduction in large-scale data sets with high dimensions.

Acknowledgements

The authors wish to thank research assistant Nannan Ma for large numeric experiments on this study.

This work was supported by National Natural Science Foundation of China (Nos. 61322211, 71031006, 61202018, 61303008), Program for New Century Excellent Talents in University (No. NCET-12-1031), National Key Basic Research and Development Program of China (973) (Nos. 2013CB329404, 2013CB329502), the Research Fund for the Doctoral Program of Higher Education (No. 20121401110013), MOE Project of Humanities and Social Sciences (No. 12YJC630174), Program for the Innovative Talents of Higher Learning Institutions of Shanxi, China (No. 20120301).

Appendix A. Related proof

Theorem 4. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS_B^U(D)$. For $\forall a, b \in C - B$, if $Sig_1(a, B, D, U) \geq Sig_1(b, B, D, U)$, then $Sig_1(a, B, D, U') \geq Sig_1(b, B, D, U')$.*

Proof. From the definition of $Sig_1(a, B, D, U) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$, we know that its value only depends on the dependency function $\gamma_B(D) = \frac{|POS_B(D)|}{|U|}$. Since $U' = U - POS_B^U(D)$, one can know $POS_B^{U'}(D) = \emptyset$ and $POS_{B\cup\{a\}}^{U'}(D) = POS_{B\cup\{a\}}^U(D) - POS_B^U(D)$. Therefore, we have

$$\begin{aligned}\frac{Sig_1(a, B, D, U)}{Sig_1(a, B, D, U')} &= \frac{\gamma_{B\cup\{a\}}^U(D) - \gamma_B^U(D)}{\gamma_{B\cup\{a\}}^{U'}(D) - \gamma_B^{U'}(D)}\\ &= \frac{|U'|}{|U|}\frac{|POS_{B\cup\{a\}}^U(D)| - |POS_B^U(D)|}{|POS_{B\cup\{a\}}^{U'}(D)| - |POS_B^{U'}(D)|}\\ &= \frac{|U'|}{|U|}\frac{|POS_{B\cup\{a\}}^U(D)| - |POS_B^U(D)|}{|POS_{B\cup\{a\}}^U(D)| - |POS_B^U(D)|}\\ &= \frac{|U'|}{|U|}.\end{aligned}$$

Because $\frac{|U'|}{|U|} \geq 0$ and if $Sig_1(a, B, D, U) \geq Sig_1(b, B, D, U)$, then $Sig_1(a, B, D, U') \geq Sig_1(b, B, D, U')$. This completes the proof. □

Theorem 5. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS_B^U(D)$. For $\forall a, b \in C - B$, if $Sig_2(a, B, D, U) \geq Sig_2(b, B, D, U)$, then $Sig_2(a, B, D, U') \geq Sig_2(b, B, D, U')$.*

Proof. Without any of generality, we suppose that $POS_B^U(D) = \{x_1, x_2, \cdots, x_p\}$ and $U' = U - POS_B^U(D) = \{x_{p+1}, x_{p+2}, \cdots, x_{|U|}\}$. From the definition of positive region, one has $r_{ij}^B \leq r_{ij}^D$ when $x_i \in POS_B^U(D)$, $\forall j \leq n$. In addition, it follows from the definition of similarity matrix that $r_{ij}^B = r_{ji}^B$ and $r_{ij}^D = r_{ji}^D$. Hence, we have that

$$\begin{aligned}H^U(D|B) &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}\\ &= -\frac{1}{|U|}\left(\sum_{i=1}^{p}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|} + \sum_{i=p+1}^{|U|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}\right)\\ &= -\frac{1}{|U|}\left(\sum_{i=1}^{p}\log\frac{|[x_i]_{\widetilde{R}_B}|}{|[x_i]_{\widetilde{R}_B}|} + \sum_{i=p+1}^{|U|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}\right)\\ &= -\frac{1}{|U|}\sum_{i=p+1}^{|U|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}\\ &= -\frac{|U'|}{|U|}\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log\frac{|[x_i]_{\widetilde{R}_B} \cap [x_i]_{\widetilde{R}_D}|}{|[x_i]_{\widetilde{R}_B}|}\\ &= \frac{|U'|}{|U|}H^{U'}(D|B).\end{aligned}$$

Hence, $\frac{Sig_2(a,B,D,U)}{Sig_2(a,B,D,U')} = \frac{|U'|}{|U|}$. Therefore, one has that $\forall a, b \in C - B$, if $Sig_2(a, B, D, U) \geq Sig_2(b, B, D, U)$, then $Sig_2(a, B, D, U') \geq Sig_2(b, B, D, U')$. This completes the proof. □

Theorem 6. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - \{x | \mu_{POS_B^U(D)}(x) = 1, x \in U\}$. For $\forall a, b \in C - B$, if $Sig_3(a, B, D, U) \geq Sig_3(b, B, D, U)$, then $Sig_3(a, B, D, U') \geq Sig_3(b, B, D, U')$.*

Proof. From the definition of $\mu_{POS_B^U(D)}(x)$, we have that

$$
\begin{aligned}
&\mu_{POS^U_{B\cup\{a\}}(D)}(x) = 1, \quad \forall x \in U \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \mu^U_{\underline{B\cup\{a\}}(X)}(x) = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \inf_{x_j \in U} \max\left\{1 - \widetilde{R}_{B\cup\{a\}}(x, x_j),\ \mu_X(x_j)\right\} = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \inf_{x_j \in U} \max\left\{1 - \max\left\{\widetilde{R}_B(x, x_j),\ \widetilde{R}_{\{a\}}(x, x_j)\right\},\ \mu_X(x_j)\right\} = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \forall x_j \in U, \max\left\{1 - \max\left\{\widetilde{R}_B(x, x_j),\ \widetilde{R}_{\{a\}}(x, x_j)\right\},\ \mu_X(x_j)\right\} = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \forall x_j \in U, \max\left\{1 - \widetilde{R}_B(x, x_j),\ \mu_X(x_j)\right\} = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \inf_{x_j \in U} \max\left\{1 - \widetilde{R}_B(x, x_j),\ \mu_X(x_j)\right\} = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \mu^U_{\underline{B}(X)}(x) = 1 \\
&\Rightarrow \quad \exists X \in U/D, \text{ s.t., } \mu_{POS^U_B(D)}(x) = 1, \quad \forall x \in U.
\end{aligned}
$$

From the definition of $Sig_3(a, B, D, U) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$, we know that

$$
\begin{aligned}
&\frac{Sig_3(a, B, D, U)}{Sig_3(a, B, D, U')} \\
&= \frac{\gamma^U_{B\cup\{a\}}(D) - \gamma^U_B(D)}{\gamma^{U'}_{B\cup\{a\}}(D) - \gamma^{U'}_B(D)} \\
&= \frac{|U'|}{|U|} \frac{\sum_{x\in U} \mu_{POS^U_{B\cup\{a\}}(D)}(x) - \sum_{x\in U} \mu_{POS^U_B(D)}(x)}{\sum_{x\in U} \mu_{POS^{U'}_{B\cup\{a\}}(D)}(x) - \sum_{x\in U} \mu_{POS^{U'}_B(D)}(x)} \\
&= \frac{|U'|}{|U|} \frac{\sum_{x\in U'} \mu_{POS^U_{B\cup\{a\}}(D)}(x) + \sum_{x\in U-U'} \mu_{POS^U_{B\cup\{a\}}(D)}(x) - \sum_{x\in U'} \mu_{POS^U_B(D)}(x) - \sum_{x\in U-U'} \mu_{POS^U_B(D)}(x)}{\sum_{x\in U} \mu_{POS^{U'}_{B\cup\{a\}}(D)}(x) - \sum_{x\in U} \mu_{POS^{U'}_B(D)}(x)} \\
&= \frac{|U'|}{|U|} \frac{[\sum_{x\in U'} \mu_{POS^U_{B\cup\{a\}}(D)}(x) - \sum_{x\in U'} \mu_{POS^U_B(D)}(x)] + [\sum_{x\in U-U'} \mu_{POS^U_{B\cup\{a\}}(D)}(x) - \sum_{x\in U-U'} \mu_{POS^U_B(D)}(x)]}{\sum_{x\in U} \mu_{POS^{U'}_{B\cup\{a\}}(D)}(x) - \sum_{x\in U} \mu_{POS^{U'}_B(D)}(x)} \\
&= \frac{|U'|}{|U|} \frac{[\sum_{x\in U'} \mu_{POS^U_{B\cup\{a\}}(D)}(x) - \sum_{x\in U'} \mu_{POS^U_B(D)}(x)] + [|U - U'| - |U - U'|]}{\sum_{x\in U} \mu_{POS^{U'}_{B\cup\{a\}}(D)}(x) - \sum_{x\in U} \mu_{POS^{U'}_B(D)}(x)} \\
&= \frac{|U'|}{|U|}.
\end{aligned}
$$

Because $\frac{|U'|}{|U|} \geq 0$ and if $Sig_3(a, B, D, U) \geq Sig_3(b, B, D, U)$, then $Sig_3(a, B, D, U') \geq Sig_3(b, B, D, U')$. This completes the proof. □

References

[1] T. Beaubouef, F.E. Perty, G. Arora, Information-theoretic measures of uncertainty for rough sets and rough relational databases, Inf. Sci. 109 (1998) 185–195.

[2] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognit. Lett. 26 (2005) 1632–1640.

[3] O. Boehm, D.R. Hardoon, L.M. Manevitz, Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms, Int. J. Mach. Learn. Cybern. 2 (3) (2011) 125–134.
[4] D.G. Chen, L. Zhang, S.Y. Zhao, Q.H. Hu, P.F. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (2012) 385–389.
[5] D. Dübois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.
[6] I. Düntsch, G. Gediga, Uncertainty measures of rough set prediction, Artif. Intell. 106 (1998) 109–137.
[7] G. Gediga, I. Düntsch, Rough approximation quality revisited, Artif. Intell. 132 (2001) 219–234.
[8] J.W. Guan, D.A. Bell, Rough computational methods for information systems, Artif. Intell. 105 (1998) 77–103.
[9] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, Int. J. Comput. Intell. 11 (2) (1995) 323–338.
[10] Q. Hu, W. Pedrycz, D. Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 40 (1) (2010) 137–150.
[11] Q. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognit. 40 (2007) 3509–3521.
[12] Q. Hu, D. Yu, W. Pedrycz, D. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (11) (2011) 1649–1667.
[13] Q. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (5) (2006) 414–423.
[14] Q. Hu, L. Zhang, S. An, D. Zhang, D. Yu, On robust fuzzy rough set models, IEEE Trans. Fuzzy Syst. 20 (4) (2012) 636–651.
[15] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. Data Eng. 16 (12) (2004) 1457–1471.
[16] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute reduction, IEEE Trans. Fuzzy Syst. 15 (1) (2007) 73–89.
[17] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (4) (2009) 824–838.
[18] J.Y. Liang, C.Y. Dang, K.S. Chin, C.M. Yam Richard, A new method for measuring uncertainty and fuzziness in rough set theory, Int. J. Gen. Syst. 31 (4) (2002) 331–342.
[19] J.Y. Liang, Y.H. Qian, Information granules and entropy theory in information systems, Sci. China, Ser. F 51 (10) (2008) 1427–1444.
[20] X.D. Liu, W. Pedrycz, T.Y. Chai, M.L. Song, The development of fuzzy rough sets with the use of structures and algebras of axiomatic fuzzy sets, IEEE Trans. Knowl. Data Eng. 21 (3) (2009) 443–462.
[21] S. Mitra, H. Banka, W. Pedrycz, Rough-fuzzy collaborative clustering, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 36 (4) (2006) 795–805.
[22] Z. Pawlak, Rough sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, London, 1991.
[23] Z. Pawlak, J.W. Grzymala-Busse, R. Slowiski, W. Ziako, Rough sets, Commun. ACM 38 (11) (1995) 89–95.
[24] W. Pedrycz, A. Bargiela, Granular clustering: a granular signature of data, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 32 (2) (2002) 212–224.
[25] W. Pedrycz, Z.A. Sosnowski, Designing decision trees with the use of fuzzy granulation, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 3 (2) (2000) 151–159.
[26] Y.H. Qian, J.Y. Liang, C.Y. Dang, Incomplete multigranulation rough set, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 40 (2) (2010) 420–431.
[27] 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.
[28] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recognit. 44 (2011) 1658–1670.
[29] Y.H. Qian, J.Y. Liang, Y.Y. Yao, C.Y. Dang, MGRS: a multi-granulation rough set, Inf. Sci. 180 (2010) 949–970.
[30] T. Ruckstieb, C. Osendorfer, P. Smagt, Minimizing data consumption with sequential online feature selection, Int. J. Mach. Learn. Cybern. 4 (3) (2013) 235–243.
[31] D. Sen, S.K. Pal, Generalized rough sets, entropy and image ambiguity measures, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 39 (1) (2009) 117–128.
[32] C.E. Shannon, The mathematical theory of communication, Bell Syst. Tech. J. 27 (3–4) (1948) 373–423, 623–656.
[33] A. Sharma, S. Imoto, S. Miyano, V. Sharma, Null space based feature selection method for gene expression data, Int. J. Mach. Learn. Cybern. 3 (4) (2012) 269–276.
[34] N. Subrahmanya, Y.C. Shin, A variational Bayesian framework for group feature selection, Int. J. Mach. Learn. Cybern. 4 (6) (2013) 609–619.
[35] E.C.C. Tang, D.G. Chen, D.S. Yeung, X.Z. Wang, J.W.T. Lee, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (5) (2008) 1130–1141.
[36] X.Z. Wang, E.C.C. Tang, S.Y. Zhao, D.G. Chen, D.S. Yeung, Learning fuzzy rules from fuzzy samples based on rough set technique, Inf. Sci. 177 (2007) 4493–4514.
[37] G.Y. Wang, H. Yu, D.C. Yang, Decision table reduction based on conditional information entropy, Chin. J. Comput. 25 (11) (2002) 1–8.
[38] M.J. Wierman, Measuring uncertainty in rough set theory, Int. J. Gen. Syst. 28 (4) (1999) 283–297.
[39] W.Z. Wu, J.S. Mi, W.X. Zhang, Generalized fuzzy rough sets, Inf. Sci. 151 (2003) 263–282.
[40] W.Z. Wu, W.X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Inf. Sci. 159 (2004) 233–254.
[41] Z.X. Xie, Y. Xu, Sparse group LASSO based uncertain feature selection, Int. J. Mach. Learn. Cybern. 5 (2) (2014) 201–210.
[42] W.H. Xu, Y.F. Liu, T.J. Li, Intuitionistic fuzzy ordered information system, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 21 (3) (2013) 367–390.
[43] Y.Y. Yao, Probabilistic rough set approximations, Int. J. Approx. Reason. 49 (2) (2008) 255–271.

[44] Y.Y. Yao, Interpreting concept learning in cognitive informatics and granular computing, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 39 (4) (2009) 855–865.
[45] Y.Y. Yao, Y. Zhao, J. Wang, On reduct construction algorithms, Lect. Notes Comput. Sci. 5150 (2008) 100–117.
[46] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee, X.Z. Wang, On the generalization of fuzzy rough sets, IEEE Trans. Fuzzy Syst. 13 (3) (2005) 343–361.
[47] S.Y. Zhao, E.C.C. Tsang, D.G. Chen, The model of fuzzy variable precision rough sets, IEEE Trans. Fuzzy Syst. 17 (2) (2009) 451–467.
[48] S.Y. Zhao, E.C.C. Tsang, X.Z. Wang, Building a rule-based classifier—a fuzzy rough set approach, IEEE Trans. Knowl. Data Eng. 22 (5) (2009) 624–638.