Artificial Intelligence 174 (2010) 597–618

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Positive approximation: An accelerator for attribute reduction in rough set theory

Yuhua Qian [a,c], Jiye Liang [a,*], Witold Pedrycz [b], Chuangyin Dang [c]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006, Shanxi, China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
[c] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 15 July 2009
Received in revised form 6 April 2010
Accepted 7 April 2010
Available online 9 April 2010

Keywords:
Rough set theory
Attribute reduction
Decision table
Positive approximation
Granular computing

ABSTRACT

Feature selection is a challenging problem in areas such as pattern recognition, machine learning and data mining. Considering a consistency measure introduced in rough set theory, the problem of feature selection, also called attribute reduction, aims to retain the discriminatory power of original features. Many heuristic attribute reduction algorithms have been proposed however, quite often, these methods are computationally time-consuming. To overcome this shortcoming, we introduce a theoretic framework based on rough set theory, called positive approximation, which can be used to accelerate a heuristic process of attribute reduction. Based on the proposed accelerator, a general attribute reduction algorithm is designed. Through the use of the accelerator, several representative heuristic attribute reduction algorithms in rough set theory have been enhanced. Note that each of the modified algorithms can choose the same attribute reduct as its original version, and hence possesses the same classification accuracy. Experiments show that these modified algorithms outperform their original counterparts. It is worth noting that the performance of the modified algorithms becomes more visible when dealing with larger data sets.

1. Introduction

Feature selection, also called attribute reduction, is a common problem in pattern recognition, data mining and machine learning. In recent years, we encounter databases in which both the number of objects becomes larger and their dimensionality (number of attributes) gets larger as well. Tens, hundreds, and even thousands of attributes are stored in many real-world application databases [6,12,37]. Attributes that are irrelevant to recognition tasks may deteriorate the performance of learning algorithms [44,45]. In other words, storing and processing all attributes (both relevant and irrelevant) could be computationally very expensive and impractical. To deal with this issue, as was pointed out in [20], some attributes can be omitted, which will not seriously impact the resulting classification (recognition) error, cf. [20]. Therefore, the omission of some attributes could not only be tolerable but even desirable relatively to the costs involved in such cases [32].

In feature selection, we encounter two general strategies, namely wrappers [16] and filters. The former employs a learning algorithm to evaluate the selected attribute subsets, and the latter selects attributes by being guided by some significance measures such as information gain [23,46], consistency [6,41], distance [15], dependency [30], and others. These

* Corresponding author. Tel.: +86 0351 7018176; fax: +86 0351 7018176.
E-mail addresses: jinchengqyh@sxu.edu.cn (Y.H. Qian), ljy@sxu.edu.cn (J.Y. Liang), pedrycz@ee.ualberta.ca (W. Pedrycz), mecdang@cityu.edu.hk (C.Y. Dang).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2010.04.018

measures can be divided into two main categories: distance-based measures and consistency-based measures [20]. Rough set theory by Pawlak [33–36] is a relatively new soft computing tool for the analysis of a vague description of an object, and has become a popular mathematical framework for pattern recognition, image processing, feature selection, neurocomputing, data mining and knowledge discovery from large data sets [7,11,31]. Attribute reduction in rough set theory offers a systematic theoretic framework for consistency-based feature selection, which does not attempt to maximize the class separability but rather attempts to retain the discernible ability of original features for the objects from the universe [13,14,53].

Generally speaking, one always needs to handle two types of data, viz. those that assume numerical values and symbolic values. For numerical values, there are two types of approaches. One relies on fuzzy rough set theory, and the other is concerned with the discretization of numerical attributes. In order to deal with numerical attributes or hybrid attributes, several approaches have been developed in the literature. Pedrycz and Vukovich regarded features as granular rather than numerical [37]. Shen and Jenshen generalized the dependency function in classical rough set framework to the fuzzy case and proposed a fuzzy-rough QUICKREDUCT algorithm [13,14,48]. Bhatt and Gopal provided a concept of fuzzy-rough sets formed on compact computational domain, which is utilized to improve the computational efficiency [3,4]. Hu et al. presented a new entropy to measure of the information quantity in fuzzy sets [21] and applied this particular measure to reduce hybrid data [22]. Data discretization is another important approach to deal with numerical values, in which we usually discretize numerical values into several intervals and associate the intervals with a set of symbolic values, see [5,28]. In the "classical" rough set theory, the attribute reduction method takes all attributes as those which assume symbolic values. Through preprocessing of original data, one can use the classical rough set theory to select a subset of features that is the most suitable for a given recognition problem.

In the last twenty years, many techniques of attribute reduction have been developed in rough set theory. The concept of the β-reduct proposed by Ziarko provides a suite of reduction methods in the variable precision rough set model [60]. An attribute reduction method was proposed for knowledge reduction in random information systems [57]. Five kinds of attribute reducts and their relationships in inconsistent systems were investigated by Kryszkiewicz [18], Li et al. [24] and Mi et al. [29], respectively. By eliminating some rigorous conditions required by the distribution reduct, a maximum distribution reduct was introduced by Mi et al. in [29]. In order to obtain all attribute reducts of a given data set, Skowron [49] proposed a discernibility matrix method, in which any two objects determine one feature subset that can distinguish them. According to the discernibility matrix viewpoint, Qian et al. [42,43] and Shao et al. [47] provided a technique of attribute reduction for interval ordered information systems, set-valued ordered information systems and incomplete ordered information systems, respectively. Kryszkiewicz and Lasek [17] proposed an approach to discovery of minimal sets of attributes functionally determining a decision attribute. The above attribute reduction methods are usually computationally very expensive, which are intolerable for dealing with large-scale data sets with high dimensions. To support efficient attribute reduction, many heuristic attribute reduction methods have been developed in rough set theory, cf. [19,20,22,25,26,39,52, 54–56]. Each of these attribute reduction methods can extract a single reduct from a given decision table.[1] For convenience, from the viewpoint of heuristic functions, we classify these attribute reduction methods into four categories: positive-region reduction, Shannon's entropy reduction, Liang's entropy reduction and combination entropy reduction. Hence, we review only four representative heuristic attribute reduction methods.

(1) Positive-region reduction

The concept of positive region was proposed by Pawlak in [33], which is used to measure the significance of a condition attribute in a decision table. While the idea of attribute reduction using positive region was originated by J.W. Grzymala-Busse in [9] and [10], and the corresponding algorithm ignores the additional computation required for selecting significant attributes. Then, Hu and Cercone [19] proposed a heuristic attribute reduction method, called positive-region reduction, which remains the positive region of target decision unchanged. The literature [20] gave an extension of this positive-region reduction for hybrid attribute reduction in the framework of fuzzy rough set. Owing to the consistency of ideas and strategies of these methods, we regard the method from [19] as their representative. These reduction methods are the first attempt to heuristic attribute reduction algorithms in rough set theory.

(2) Shannon's entropy reduction

The entropy reducts have first been introduced in 1993/1994 by Skowron in his lectures at Warsaw University. Based on the idea, Slezak introduced Shannon's information entropy to search reducts in the classical rough set model [50–52]. Wang et al. [54] used conditional entropy of Shannon's entropy to calculate the relative attribute reduction of a decision information system. In fact, several authors also have used variants of Shannon's entropy or mutual information to measure uncertainty in rough set theory and construct heuristic algorithm of attribute reduction in rough set theory [22,55,56]. Here

[1] The attribute reduct obtained preserves a particular property of a given decision table. However, as Prof. Bazan said, from the viewpoint of stability of attribute reduct, the selected reduct may be of bad quality [1,2]. To overcome this problem, Bazan developed a method for dynamic reducts to get a stable attribute reduct from a decision table. How to accelerate the method for dynamic reducts is an interesting topic in further work.

we only select the attribute reduction algorithm in the literature [54] as their representative. This reduction method remains the conditional entropy of target decision unchanged.

(3) Liang's entropy reduction

Liang et al. [25] defined a new information entropy to measure the uncertainty of an information system and applied the entropy to reduce redundant features [26]. Unlike Shannon's entropy, this information entropy can measure both the uncertainty of an information system and the fuzziness of a rough decision in rough set theory. This reduction method can preserve the conditional entropy of a given decision table. In fact, the mutual information form of Liang's entropy also can be used to construct a heuristic function of an attribute reduction algorithm. For simplicity, we here ignore its discussion.

(4) Combination entropy reduction

In general, the objects in an equivalence class cannot be distinguished each other, but the objects in different equivalence classes can be distinguished each other in rough set theory. Therefore, in a broad sense, the knowledge content of a given attribute set can be characterized by the entire number of pairs of the objects which can be distinguished each other on the universe. Based on this consideration, Qian and Liang [39] presented the concept of combination entropy for measuring the uncertainty of information systems and used its conditional entropy to select a feature subset. This reduction method can obtain an attribute subset that possesses the same number of pairs of the elements which can be distinguished each other as the original decision table. This measure focuses on a completely different point of view, which is mainly based on the intuitionistic knowledge content nature of information gain.

Each of these above methods preserves a particular property of a given information system or 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. In this paper, we are not concerned with how to discretize numerical attributes and construct a heuristic function for attribute reduction. Our objective is to focus on how to improve the time efficiency of a heuristic attribute reduction algorithm. We propose a new rough set framework, which is called positive approximation. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure of a rough set using a granulation order. Based on the positive approximation, we develop a common accelerator for improving the time efficiency of a heuristic attribute reduction, which provides a vehicle of making algorithms of rough set based feature selection techniques faster. By incorporating the accelerator into each of the above four representative heuristic attribute reduction methods, we construct their modified versions. Numerical experiments show that each of the modified methods can choose the same attribute subset as that of the corresponding original method while greatly reducing computing time. 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. Some basic concepts in rough set theory are briefly reviewed in Section 2. In Section 3, we establish the positive approximation framework and investigate some of its main properties. In Section 4, through analyzing the rank preservation of four representative significance measures of attributes, we develop a general modified attribute reduction algorithm based on the positive approximation. Experiments on nine public data sets show that these modified algorithms outperform their original counterparts in terms of computational time. Finally, Section 5 concludes this paper by bringing some remarks and discussions.

2. Preliminaries

In this section, we will review several basic concepts in rough set theory. Throughout this paper, we suppose that the universe U is a finite nonempty set.

Let U be a finite and nonempty set called the universe and $R \subseteq U \times U$ an equivalence relation on U. Then $K = \langle U, R\rangle$ is called an approximation space [33–36]. The equivalence relation R partitions the set U into disjoint subsets. This partition of the universe is called a quotient set induced by R, denoted by U/R. It represents a very special type of similarity between elements of the universe. If two elements $x, y \in U$ $(x \neq y)$ belong to the same equivalence class, we say that x and y are indistinguishable under the equivalence relation R, i.e., they are equal in R. We denote the equivalence class including x by $[x]_R$. Each equivalence class $[x]_R$ may be viewed as an information granule consisting of indistinguishable elements [59]. The granulation structure induced by an equivalence relation is a partition of the universe.

Given an approximation space $K = \langle U, R\rangle$ and an arbitrary subset $X \subseteq U$, one can construct a rough set of the set on the universe by elemental information granules in the following definition:

$$\begin{cases} \underline{R}X = \bigcup\{[x]_R \mid [x]_R \subseteq X\}, \\ \overline{R}X = \bigcup\{[x]_R \mid [x]_R \cap X \neq \emptyset\}, \end{cases}$$

where $\underline{R}X$ and $\overline{R}X$ are called R-lower approximation and R-upper approximation with respect to R, respectively. The order pair $\langle \underline{R}X, \overline{R}X\rangle$ is called a rough set of X with respect to the equivalence relation R. Equivalently, they also can be written as

$$\begin{cases} \underline{R}X = \{x \mid [x]_R \subseteq X\}, \\ \overline{R}X = \{x \mid [x]_R \cap X \neq \emptyset\}. \end{cases}$$

There are two kinds of attributes for a classification problem, which can be characterized by a decision table $S = (U, C \cup D)$ with $C \cap D = \emptyset$, where an element of C is called a condition attribute, C is called a condition attribute set, an element of D is called a decision attribute, and D is called a decision attribute set. Each nonempty subset $B \subseteq C$ determines an equivalence relation in the following way:

$$R_B = \{(x, y) \in U \times U \mid a(x) = a(y),\ \forall a \in B\},$$

where $a(x)$ and $a(y)$ denote the values of objects x and y under a condition attribute a, respectively. This equivalence relation R_B partitions U into some equivalence classes given by

$$U/R_B = \{[x]_B \mid x \in U\}, \quad \text{for simplicity, } U/R_B \text{ will be replaced by } U/B,$$

where $[x]_B$ denotes the equivalence class determined by x with respect to B, i.e., $[x]_B = \{y \in U \mid (x, y) \in R_B\}$.

Assume the objects are partitioned into r mutually exclusive crisp subsets $\{Y_1, Y_2, \ldots, Y_r\}$ by the decision attributes D. Given any subset $B \subseteq C$ and R_B is the equivalence relation induced by B, then one can define the lower and upper approximations of the decision attributes D as

$$\begin{cases} \underline{R_B} D = \{\underline{R_B} Y_1, \underline{R_B} Y_2, \ldots, \underline{R_B} Y_r\}, \\ \overline{R_B} D = \{\overline{R_B} Y_1, \overline{R_B} Y_2, \ldots, \overline{R_B} Y_r\}. \end{cases}$$

Denoted by $POS_B(D) = \bigcup_{i=1}^{r} \underline{R_B} Y_i$, it is called the positive region of D with respect to the condition attribute set B.

We define a partial relation $\preccurlyeq$ on the family $\{B \mid B \subseteq C\}$ as follows: $P \preccurlyeq Q$ (or $Q \succcurlyeq P$) if and only if, for every $P_i \in U/P$, there exists $Q_j \in U/Q$ such that $P_i \subseteq Q_j$, where $U/P = \{P_1, P_2, \ldots, P_m\}$ and $U/Q = \{Q_1, Q_2, \ldots, Q_n\}$ are partitions induced by $P, Q \subseteq C$, respectively [40]. In this case, we say that Q is coarser than P, or P is finer than Q. If $P \preccurlyeq Q$ and $U/P \neq U/Q$, we say Q is strictly coarser than P (or P is strictly finer than Q), denoted by $P \prec Q$ (or $Q \succ P$).

It becomes clear that $P \prec Q$ if and only if, for every $X \in U/P$, there exists $Y \in U/Q$ such that $X \subseteq Y$, and there exists $X_0 \in U/P$, $Y_0 \in U/Q$ such that $X_0 \subset Y_0$.

3. Positive approximation and its properties

A partition induced by an equivalence relation provides a granulation world for describing a target concept [38]. Thus, a sequence of granulation worlds stretching from coarse to fine granulation can be determined by a sequence of attribute sets with granulations from coarse to fine in the power set of attributes, which is called a positive granulation world. If the granulation worlds are arranged from the fine to the coarse one, then the sequence is called converse granulation worlds [27,40].

In this section, we introduce a new set-approximation approach called positive 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 positive granulation world. Given a decision table $S = (U, C \cup D)$, $U/D = \{Y_1, Y_2, \ldots, Y_r\}$ is called a target decision, in which each equivalence class Y_i $(i \leqslant r)$ 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 $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \ldots, R_n\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots \succcurlyeq R_n$ $(R_i \in 2^C)$. Given $P_i = \{R_1, R_2, \ldots, R_i\}$, 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

$$\begin{cases} \underline{P_i}(X) = \bigcup_{k=1}^{i} \underline{R_k} X_k, \\ \overline{P_i}(X) = \overline{R_i} X, \end{cases}$$

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

Correspondingly, the boundary of X is given as

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

Theorem 1. *Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \ldots, R_n\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots \succcurlyeq R_n$ $(R_i \in 2^C)$. Given $P_i = \{R_1, R_2, \ldots, R_i\}$, then $\forall P_i$ $(i = 1, 2, \ldots, n)$, we have*

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

Fig. 1 visualizes the mechanism of the positive approximation.

In Fig. 1, let $P_1 = \{R_1\}$ and $P_2 = \{R_1, R_2\}$ with $R_1 \succcurlyeq R_2$ be two granulation orders. $\underline{R_1} X_1$ is the lower approximation of X_1 obtained by the equivalence relation R_1, and $\underline{R_2} X_2$ is the lower approximation of X_2 obtained by the equivalence

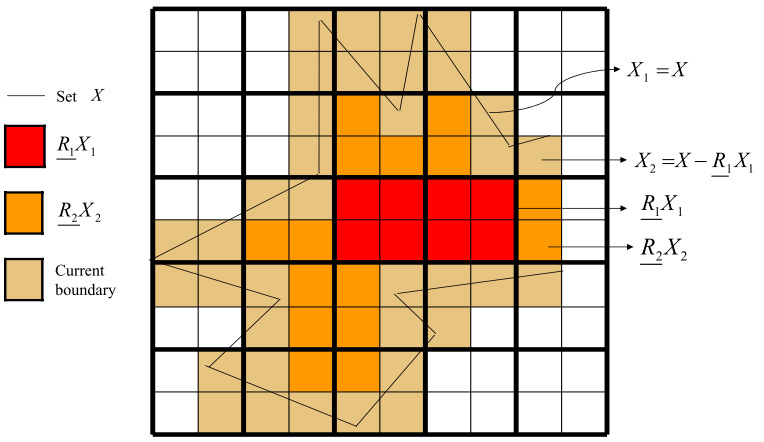


Fig. 1. Sketch map of the positive approximation.

relation R_2. Hence, $\underline{P_2}(X) = \underline{R_1}X_1 \cup \underline{R_2}X_2 = \underline{R_2}X$. The mechanism characterizes the structure of a rough set approximation, which can be used to gradually compute the lower approximation of a target concept/decision.

Theorem 2. *Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \ldots, R_n\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots \succcurlyeq R_n$ ($R_i \in 2^C$). Given $P_i = \{R_1, R_2, \ldots, R_i\}$, then $\forall P_i$ $(i = 1, 2, \ldots, n)$, we have*

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

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

In order to illustrate that the essence of the positive approximation is concentrated on the changes in the construction of the target concept X (equivalence classes in the lower approximation of X with respect to P_i), one can redefine P_i-positive approximation of X by using some equivalence classes on U. The structures of P_i-lower approximation $\underline{P_i}(X)$ and P_i-upper approximation $\overline{P_i}(X)$ of P_i-positive approximation of X can be represented as follows

$$\begin{cases} S(\underline{P_i}(X)) = \{[x]_{R_k} \mid [x]_{R_k} \subseteq \underline{R_k}X_k,\ k \leqslant i,\ X_1 = X,\ X_{k+1} = \underline{R_k}X_k\}, \\ S(\overline{P_i}(X)) = \{[x]_{R_i} \mid [x]_{R_i} \cap X \neq \emptyset\}, \end{cases}$$

where $[x]_{R_i}$ represents the equivalence class including x in the partition U/R_i.

Example 1. Let $U = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$, $X = \{e_1, e_2, e_3, e_4, e_7, e_8\}$ and $U/R_1 = \{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_5, e_6, e_7, e_8\}\}$, $U/R_2 = \{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_5, e_6\}, \{e_7, e_8\}\}$ be two partitions on U.

Obviously, $R_1 \succcurlyeq R_2$ holds. Thus, one can construct two granulation orders (a family of equivalence relations) $P_1 = \{R_1\}$ and $P_2 = \{R_1, R_2\}$.

By computing the positive approximation of X, one can easily obtain that

$$\begin{aligned} &S\big(\underline{P_1}(X)\big) = \big\{\{e_1\}, \{e_2\}, \{e_3, e_4\}\big\}, \\ &S\big(\overline{P_1}(X)\big) = \big\{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_5, e_6, e_7, e_8\}\big\}, \\ &S\big(\underline{P_2}(X)\big) = \big\{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_7, e_8\}\big\} \quad \text{and} \\ &S\big(\overline{P_2}(X)\big) = \big\{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_7, e_8\}\big\}. \end{aligned}$$

That is to say, the target concept X can be described by using granulation orders $P_1 = \{R_1\}$ and $P_2 = \{R_1, R_2\}$, respectively.

Definition 2. Let $S = (U, C \cup D)$ be a decision table, $P_i = \{R_1, R_2, \ldots, R_i\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots R_i$ ($R_i \in 2^C$) and $U/D = \{Y_1, Y_2, \ldots, Y_r\}$. Lower approximation and upper approximation of D with respect to P_i are defined as

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

$\underline{P_i}D$ 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 $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \ldots, R_n\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots \succcurlyeq R_n$ $(R_i \in 2^C)$. Given $P_i = \{R_1, R_2, \ldots, R_i\}$, we have*

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

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

Example 2. Let $S = (U, C \cup D)$ be a decision table, where $U = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$, $C = \{a_1, a_2\}$, $U/D = \{\{e_1, e_2, e_3, e_4, e_7, e_8\}, \{e_5, e_6\}\}$ and $U/\{a_1\} = \{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_5, e_6, e_7, e_8\}\}$, $U/C = \{\{e_1\}, \{e_2\}, \{e_3, e_4\}, \{e_5, e_6\}, \{e_7, e_8\}\}$ be two partitions on U.

Obviously, $\{a_1\} \succcurlyeq C$ holds. Thus, one can construct two granulation orders (a family of equivalence relations) $P_1 = \{a_1\}$ and $P_2 = \{a_1, a_2\}$.

By computing the positive approximation of D, one can easily obtain that

$$POS^U_{P_1}(D) = POS^{U_1}_{P_1}(D) = \{e_1, e_2, e_3, e_4\}, \quad \text{where } U_1 = U,$$
$$U_2 = U - POS^U_{P_1}(D) = \{e_5, e_6, e_7, e_8\},$$
$$POS^{U_2}_{P_2}(D) = \{e_5, e_6\}.$$

Hence,

$$POS^U_{P_2}(D) = \{e_1, e_2, e_3, e_4, e_5, e_6\} = POS^U_{P_1}(D) \cup POS^{U_2}_{P_2}(D).$$

That is to say, the target decision D can be positively approximated by using granulation orders P_1 and P_2 on the gradually reduced universe, respectively. This mechanism implies the idea of the accelerator proposed in this paper for improving the computing performance of a heuristic attribute reduction algorithm.

The dependency function (or level of consistency [5]) is used to characterize the dependency degree of an attribute subset with respect to a given decision [8,33]. Given a decision table $S = (U, C \cup D)$, the dependency function of condition attributes C with respect to the decision attribute 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 $0 \leqslant \gamma_P(D) \leqslant 1$.

The dependency function reflects the granulation order P's power to dynamically approximate D. When $\gamma = 1$, one says D completely depends on the granulation order P. It means that the decision can be precisely described by the information granules generated by the granulation order P. This dependency function can be used to measure the significance of categorical attributes relative to the decision and construct a heuristic function for designing an attribute reduction algorithm.

4. FSPA: feature selection based on the positive approximation

Each feature selection method preserves a particular property of a given information system, which is based on a certain predetermined heuristic function. In rough set theory, attribute reduction is about finding some attribute subsets that have the minimal attributes and retain some particular properties. For example, the dependency function keeps the approximation power of a set of condition attributes. To design a heuristic attribute reduction algorithm, three key problems should be considered, which are significance measures of attributes, search strategy and stopping (termination) criterion. As there are symbolic attributes and numerical attributes in real-world data, one needs to proceed with some preprocessing. Through attribute discretization, it is easy to induce an equivalence partition. However, the existing heuristic attribute reduction methods are computationally intensive which become infeasible in case of large-scale data. As already noted, we do not reconstruct significance measures of attributes and design new stopping criteria, but improve the search strategies of the existing algorithms by exploiting the proposed concept of positive approximation.

4.1. Forward attribute reduction algorithms

In rough set theory, to support efficient attribute reduction, many heuristic attribute reduction methods have been developed, in which forward greedy search strategy is usually employed, cf. [19,20,22,25,26,39,52]. In this kind of attribute reduction approaches, two important measures of attributes are used for heuristic functions, which are inner importance

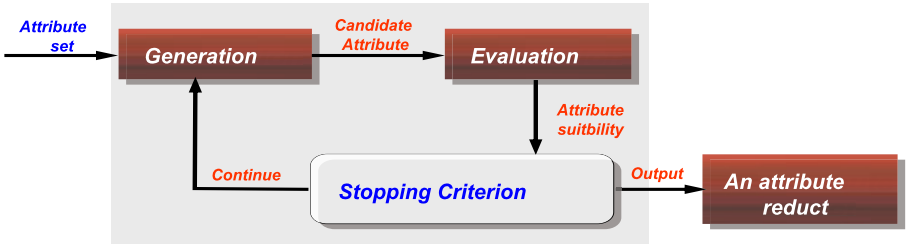


Fig. 2. The process of forward greedy attribute reduction algorithm.

measure and outer importance measure. The inner importance measure is applicable to determine the significance of every attribute, while the outer importance measure 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, starting with the attribute with the maximal inner importance, we take the attribute with the maximal outer significance into the attribute subset in each loop until this feature subset satisfies the stopping criterion, and then we can get 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: Compute $Sig^{inner}(a_k, C, D)$, $k \leqslant |C|$; $//Sig^{inner}(a_k, C, D)$ is the inner importance measure of the attribute a_k;
Step 3: Put a_k into *red*, where $Sig^{inner}(a_k, C, D, U) > 0$;
Step 4: While $EF(red, D) \neq EF(C, D)$ Do //This provides a stopping criterion.
$\{red \leftarrow red \cup \{a_0\}$, where $Sig^{outer}(a_0, red, D) = \max\{Sig^{outer}(a_k, red, D), a_k \in C - red\}\}$; $//Sig^{outer}(a_k, C, D)$ is the outer importance measure of the attribute a_k;
Step 5: Return *red* and end.

This algorithm can obtain an attribute reduct from a given decision table. Fig. 2 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 the mechanism of the algorithm.

Remark. Given any definition of attribute reduct and heuristic function, using the above attribute reduction framework, one can heuristically find an attribute reduct (a feature subset) that preserves a particular property of a decision table. If we survey the attribute reduct from the viewpoint of rough classifiers, these attribute reduction algorithms may lead to overfitting in the approximation of concepts, which will weaken the generalization ability of rough classifiers induced by the attribute reducts obtained. This problem could be caused by two cases. One is that the attribute subset induced by an attribute reduction algorithm with forward greedy searching strategy may be redundant. That is to say, there are some redundant attributes in the attribute subset obtained from the definition of a given attribute reduct. The other is that the definition of each of attribute reductions does not take into account the generalization ability of the rough classifier induced by the attribute reduct obtained. These two situations yield the same overfitting problem as a decision tree does when the tree has too long paths. Hence, it is very desirable to solve the overfitting problem of feature selection for learning a rough classifier in the framework of rough set theory. This issue will be addressed in future work.

4.2. Four representative significance measures of attributes

For efficient attribute reduction, many heuristic attribute reduction methods have been developed in rough set theory, see [19,20,22,25,26,39,52,54–56]. For convenience, as was pointed out in the introduction part of this paper, we only focus on the four representative attribute reduction methods here.

Given a decision table $S = (U, C \cup D)$, one can obtain the condition partition $U/C = \{X_1, X_2, \ldots, X_m\}$ and the decision partition $U/D = \{Y_1, Y_2, \ldots, Y_n\}$. Through these notations, in what follows we review four types of significance measures of attributes.

The idea of attribute reduction using positive region was first originated by Grzymala-Busse in Refs. [9] and [10], and the corresponding algorithm ignores the additional computation of choice of significant attributes. Hu and Cercone proposed a heuristic attribute reduction method, called positive-region reduction (PR), which remains the positive region of target decision unchanged [19]. 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 $\forall a \in B$. The significance measure of a in B is defined as

$$Sig_1^{inner}(a, B, D) = \gamma_B(D) - \gamma_{B-\{a\}}(D),$$

where $\gamma_B(D) = \frac{|POS_B(D)|}{|U|}$.

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

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

As Shannon's information entropy was introduced to search reducts in classical rough set model [52], Wang et al. used its conditional entropy to calculate the relative attribute reduction of a decision information system [54]. In fact, several authors also have used variants of Shannon's entropy to measure uncertainty in rough set theory and construct heuristic algorithm of attribute reduction. Here we only select the attribute reduction algorithm in the literature [54] as their representative. This reduction method remains the conditional entropy of target decision unchanged, denoted by SCE, in which the conditional entropy reads as

$$H(D|B) = -\sum_{i=1}^{m} p(X_i) \sum_{j=1}^{n} p(Y_j|X_i) \log\big(p(Y_j|X_i)\big),$$

where $p(X_i) = \frac{|X_i|}{|U|}$ and $p(Y_j|X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$. Using the conditional entropy, the definitions of the significance measures are expressed in the following way.

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

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

Definition 7. 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

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

As was pointed out in [25], Shannon's entropy is not a fuzzy entropy, and cannot measure the fuzziness of a rough decision in rough set theory. Hence, Liang et al. defined another information entropy and its conditional entropy to measure the uncertainty of an information system and applied the proposed entropy to reduce redundant features [25,26]. This reduction method can preserve the conditional entropy of a given decision table, denoted here by LCE. The conditional entropy used in the study is defined as

$$E(D|C) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Y_j \cap X_i|}{|U|} \frac{|Y_j^c \cap X_i^c|}{|U|}.$$

The corresponding significance measures are listed as follows.

Definition 8. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in B$. The significance measure of a in B is defined as

$$Sig_3^{inner}(a, B, D) = E\big(D|B - \{a\}\big) - E(D|B).$$

Definition 9. 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

$$Sig_3^{outer}(a, B, D) = E(D|B) - E\big(D|B \cup \{a\}\big).$$

Based on the intuitionistic knowledge content nature of information gain, Qian and Liang in [39] presented the concept of combination entropy for measuring the uncertainty of information systems and used its conditional entropy to obtain a feature subset. This reduction method can obtain an attribute subset that possesses the same number of pairs of the elements which can be distinguished each other as the original decision table, denoted here by CCE. The following definition of the conditional entropy is considered

$$CE(D|C) = \sum_{i=1}^{m} \left(\frac{|X_i|}{|U|} \frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n} \frac{|X_i \cap Y_j|}{|U|} \frac{C_{|X_i \cap Y_j|}^2}{C_{|U|}^2} \right),$$

where $C_{|X_i|}^2 = \frac{|X_i| \times (|X_i|-1)}{2}$ denotes the number of the pairs of the objects which are not distinguishable each other in the equivalence class X_i.

The conditional entropy also can be used to construct the corresponding significance measures of attributes in decision tables.

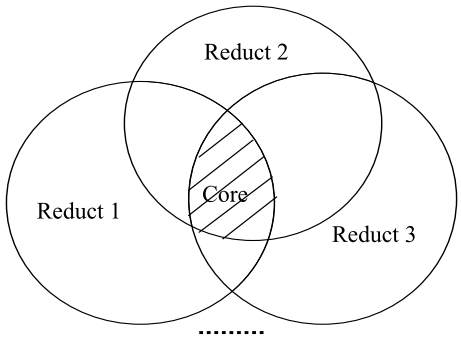


Fig. 3. The relationship between the core and all attribute reducts.

Definition 10. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in B$. The significance measure of a in B is defined as

$$Sig_4^{inner}(a, B, D) = CE\big(D|B - \{a\}\big) - CE(D|B).$$

Definition 11. 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

$$Sig_4^{outer}(a, B, D) = CE(D|B) - CE\big(D|B \cup \{a\}\big).$$

All the definitions used above are used to select an attribute in a heuristic attribute reduction algorithm. For a given decision table, the intersection of all attribute reducts is said to be indispensable and is called the core. Each attribute in the core must be in every attribute reduct of the decision table. The core may be an empty set. The relationship between the core and all attribute reducts can be displayed by Fig. 3.

The above four kinds of significance measures can be used to find the core attributes. The following theorem is of interest with this regard.

Theorem 4. *(See [26,33,39,54].) Let $S = (U, C \cup D)$ be a decision table and $a \in C$. If $Sig_\Delta^{inner}(a, C, D) > 0$ ($\Delta = \{1, 2, 3, 4\}$), then a is a core attribute of S in the context of type Δ.*

From the definition of the core, one can see that each attribute in the core must be in every attribute reduct of the decision table. It is well known that, if $Sig_\Delta^{inner}(a, C, D) = 0$ ($\Delta = \{1, 2, 3, 4\}$), then one still can find at least one attribute reduct when a is deleted. If $Sig_\Delta^{inner}(a, C, D) > 0$ ($\Delta = \{1, 2, 3, 4\}$), then the attribute a is indispensable in all attribute reducts. Therefore, the attribute a must be a core attribute of S in the context of type Δ.

In a heuristic attribute reduction algorithm, based on the above theorem, one can find an attribute reduct by gradually adding selected attributes to the core attributes.

4.3. Rank preservation of significance measures of attributes

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 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 $Sig_1^{outer}(a, B, D, U) \geqslant Sig_1^{outer}(b, B, D, U)$, then $Sig_1^{outer}(a, B, D, U') \geqslant Sig_1^{outer}(b, B, D, U')$.*

Proof. From the definition of $Sig_1^{outer}(a, B, D) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$, 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

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

$$= \frac{|U'|}{|U|}\frac{|POS^U_{B\cup\{a\}}(D)| - |POS^U_B(D)|}{|POS^{U'}_{B\cup\{a\}}(D)| - |POS^{U'}_B(D)|}$$
$$= \frac{|U'|}{|U|}\frac{|POS^U_{B\cup\{a\}}(D)| - |POS^U_B(D)|}{|POS^U_{B\cup\{a\}}(D)| - |POS^U_B(D)|}$$
$$= \frac{|U'|}{|U|}.$$

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

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 6. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS^U_B(D)$. For $\forall a, b \in C - B$, if $Sig_2^{outer}(a, B, D, U) \geqslant Sig_2^{outer}(b, B, D, U)$, then $Sig_2^{outer}(a, B, D, U') \geqslant Sig_2^{outer}(b, B, D, U')$.*

Proof. Let $U/B = \{X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_m\}$, $U/D = \{Y_1, Y_2, \ldots, Y_n\}$, where $X_{p+1}, X_{p+2}, \ldots, X_m \subseteq POS^U_B(D)$. Since for each equivalence class $X \subseteq POS_B(D)$, there exists a decision class Y such that $X \cap Y = X$. Let us denote Shannon's conditional entropy in the universe U by $H^U(D|B)$. Then it follows that

$$\begin{aligned}
H^U(D|B) &= -\sum_{i=1}^{m} p(X_i)\sum_{j=1}^{n} p(Y_j|X_i)\log\big(p(Y_j|X_i)\big)\\
&= -\left(\sum_{i=1}^{p} p(X_i)\sum_{j=1}^{n} p(Y_j|X_i)\log\big(p(Y_j|X_i)\big) + \sum_{i=p+1}^{m} p(X_i)\sum_{j=1}^{n} p(Y_j|X_i)\log\big(p(Y_j|X_i)\big)\right)\\
&= -\left(\sum_{i=1}^{p}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|X_i|}\log\frac{|X_i\cap Y_j|}{|X_i|} + \sum_{i=p+1}^{m}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|X_i|}\log\frac{|X_i\cap Y_j|}{|X_i|}\right)\\
&= -\left(\sum_{i=1}^{p}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|X_i|}\log\frac{|X_i\cap Y_j|}{|X_i|} + \sum_{i=p+1}^{m}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i|}{|X_i|}\log\frac{|X_i|}{|X_i|}\right)\\
&= -\sum_{i=1}^{p}\frac{|X_i|}{|U|}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|X_i|}\log\frac{|X_i\cap Y_j|}{|X_i|}\\
&= -\frac{|U'|}{|U|}\sum_{i=1}^{p}\frac{|X_i|}{|U'|}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|X_i|}\log\frac{|X_i\cap Y_j|}{|X_i|}\\
&= \frac{|U'|}{|U|}H^{U'}(D|B).
\end{aligned}$$

Therefore, we have that $\frac{Sig_2^{outer}(a,B,D,U)}{Sig_2^{outer}(a,B,D,U')} = \frac{|U'|}{|U|}$. Thus, if $Sig_2^{outer}(a, B, D, U) \geqslant Sig_2^{outer}(b, B, D, U)$, $\forall a, b \in C - B$, then $Sig_2^{outer}(a, B, D, U') \geqslant Sig_2^{outer}(b, B, D, U')$. This completes the proof. □

Then, we obtain the rank preservation of significance measures of attributes based on Liang's conditional entropy, which is given by the following theorem.

Theorem 7. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS^U_B(D)$. For $\forall a, b \in C - B$, if $Sig_3^{outer}(a, B, D, U) \geqslant Sig_3^{outer}(b, B, D, U)$, then $Sig_3^{outer}(a, B, D, U') \geqslant Sig_3^{outer}(b, B, D, U')$.*

Proof. Suppose $U/B = \{X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_m\}$, $U/D = \{Y_1, Y_2, \ldots, Y_n\}$, where $X_{p+1}, X_{p+2}, \ldots, X_m \subseteq POS^U_B(D)$. For each equivalence class $X \subseteq POS_B(D)$, there exists a decision class Y such that $X \cap Y = X$, i.e., $X \subseteq Y$. We denote the Liang's conditional entropy in the universe U by $E^U(D|B)$. Then, one has

$$
\begin{aligned}
E^U(D|B) &= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|Y_j^c - X_i^c|}{|U|}\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|X_i - Y_j|}{|U|}\\
&= \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|X_i - Y_j|}{|U|} + \sum_{i=p+1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|X_i - Y_j|}{|U|}\\
&= \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|X_i - Y_j|}{|U|} + \sum_{i=p+1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|\emptyset|}{|U|}\\
&= \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|X_i - Y_j|}{|U|}\\
&= \frac{|U'|^2}{|U|^2}\sum_{i=1}^{p}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U'|}\frac{|X_i - Y_j|}{|U'|}\\
&= \frac{|U'|^2}{|U|^2}E^{U'}(D|B).
\end{aligned}
$$

Hence, we have $\frac{Sig_3^{outer}(a,B,D,U)}{Sig_3^{outer}(a,B,D,U')} = \frac{|U'|^2}{|U|^2}$. Therefore, for any $a, b \in C - B$, if $Sig_3^{outer}(a, B, D, U) \geqslant Sig_3^{outer}(b, B, D, U)$, then $Sig_3^{outer}(a, B, D, U') \geqslant Sig_3^{outer}(b, B, D, U')$. This completes the proof. □

Finally, similar to the above three theorems, one can show that significance measures of attributes based on the conditional combination entropy also possesses the property of rank preservation, which is shown as follows.

Theorem 8. *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 $Sig_4^{outer}(a, B, D, U) \geqslant Sig_4^{outer}(b, B, D, U)$, then $Sig_4^{outer}(a, B, D, U') \geqslant Sig_4^{outer}(b, B, D, U')$.*

Proof. Suppose $U/B = \{X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_m\}$, $U/D = \{Y_1, Y_2, \ldots, Y_n\}$, where $X_{p+1}, X_{p+2}, \ldots, X_m \subseteq POS_B^U(D)$. Hence, for any equivalence class $X \subseteq POS_B(D)$, there exists a decision class Y such that $X \cap Y = X$, i.e., $X \subseteq Y$. Denote the conditional combination entropy in the universe U by $CE^U(D|B)$. Then, one has

$$
\begin{aligned}
CE^U(D|B) &= \sum_{i=1}^{m}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U|}^2}\right)\\
&= \sum_{i=1}^{p}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U|}^2}\right) + \sum_{i=p+1}^{m}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U|}^2}\right)\\
&= \sum_{i=1}^{p}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U|}^2}\right) + \sum_{i=p+1}^{m}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2}\right)\\
&= \sum_{i=1}^{p}\left(\frac{|X_i|}{|U|}\frac{C_{|X_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U|}^2}\right)\\
&= \frac{|U'|C_{|U'|}^2}{|U|C_{|U|}^2}\sum_{i=1}^{p}\left(\frac{|X_i|}{|U'|}\frac{C_{|X_i|}^2}{C_{|U'|}^2} - \sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U'|}\frac{C_{|X_i\cap Y_j|}^2}{C_{|U'|}^2}\right)\\
&= \frac{|U'|C_{|U'|}^2}{|U|C_{|U|}^2}CE^{U'}(D|B).
\end{aligned}
$$

In the sequel we obtain $\frac{Sig_4^{outer}(a,B,D,U)}{Sig_4^{outer}(a,B,D,U')} = \frac{|U'|C_{|U'|}^2}{|U|C_{|U|}^2}$. Therefore, for any $a, b \in C-B$, if $Sig_4^{outer}(a, B, D, U) \geqslant Sig_4^{outer}(b, B, D, U)$, then $Sig_4^{outer}(a, B, D, U') \geqslant Sig_4^{outer}(b, B, D, U')$. This completes the proof. □

From Theorems 5–8, one can see that they can be uniformly represented by the following theorem.

Theorem 9 *(Rank preservation). 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* $Sig_\Delta^{outer}(a, B, D, U) \geqslant Sig_\Delta^{outer}(b, B, D, U)$ $(\Delta = \{1, 2, 3, 4\})$, *then* $Sig_\Delta^{outer}(a, B, D, U') \geqslant Sig_\Delta^{outer}(b, B, D, U')$.

From this theorem, 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.

Besides the above four kind attribute reducts, as we know, there are many other different kind of attribute reducts. We do not discuss the rank preservation of each existing reducts taking into account the compact of the paper. In fact, rank preservation of attributes selected is based on the monotonicity of positive region of target decision in a heuristic feature selection process. In other words, the rank of attributes selected will be unchanged when the scale of positive region of target decision increases as the number of selected attributes becomes bigger.

4.4. Attribute reduction algorithm based on the positive approximation

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 a 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, one knows that the rank preservation of attributes in the context of the positive approximation. Hence, we can construct an improved forward search algorithm based on the positive approximation, which is formulated as follows. In this general 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.

Algorithm Q1. A general improved feature selection algorithm based on the positive approximation (FSPA).

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: Compute $Sig^{inner}(a_k, C, D, U)$, $k \leqslant |C|$;
Step 3: Put a_k into *red*, where $Sig^{inner}(a_k, C, D, U) > 0$;// These attributes form the core of the given decision table;
Step 4: $i \leftarrow 1$, $R_1 = red$, $P_1 = \{R_1\}$ and $U_1 \leftarrow U$;
Step 5: While $EF^{U_i}(red, D) \neq EF^{U_i}(C, D)$ Do
{Compute the positive region of positive approximation $POS_{P_i}^U(D)$,
$U_i = U - POS_{P_i}^U(D)$,
$i \leftarrow i + 1$,
$red \leftarrow red \cup \{a_0\}$, where $Sig^{outer}(a_0, red, D, U_i) = \max\{Sig^{outer}(a_k, red, D, U_i), a_k \in C - red\}$,
$R_i \leftarrow R_i \cup \{a_0\}$,
$P_i \leftarrow \{R_1, R_2, \ldots, R_i\}$};
Step 6: Return *red* and end.

Computing the significance measure of an attribute $Sig^{inner}(a_k, C, D, U)$ is one of the key steps in FSPA. Xu et al. in [57] gave a quick algorithm with time complexity $O(|U|)$. Hence, the time complexity of computing the core in Step 2 is $O(|C||U|)$. In Step 5, we begin with the core 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 $O(\sum_{i=1}^{|C|} |U_i|(|C| - i + 1))$. Thus the time complexity of FSPA is $O(|U||C| + \sum_{i=1}^{|C|} |U_i|(|C| - i + 1))$. However, the time complexity of a classical heuristic algorithm is $O(|U||C| + \sum_{i=1}^{|C|} |U|(|C| - i + 1))$. Obviously, the time complexity of FSPA is much lower than that of each of classical heuristic attribute reduction algorithms. Hence, one can draw a conclusion that the general feature selection algorithm based on the positive approximation (FSPA) may significantly reduce the computational time for attribute reduction from decision tables. To stress these findings, the time complexity of each step in original algorithms and FSPA is shown as Table 1.

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

(1) One can select the same attribute in each loop of the improved algorithm and that of the original one. This provides a restriction of keeping the result of an attribute reduction algorithm.

Table 1
The complexities description.

Algorithms	Step 2	Step 3	Step 5	Other steps
Each of original algorithms	$O(\|C\|\|U\|)$	$O(\|C\|)$	$O(\sum_{i=1}^{\|C\|} \|U\|(\|C\|-i+1))$	Constant
FSPA	$O(\|C\|\|U\|)$	$O(\|C\|)$	$O(\sum_{i=1}^{\|C\|} \|U_i\|(\|C\|-i+1))$	Constant

Table 2
Data sets description.

	Data sets	Cases	Features	Classes
1	Mushroom	5644	22	2
2	Tic-tac-toe	958	9	2
3	Dermatology	358	34	6
4	Kr-vs-kp	3196	36	2
5	Breast-cancer-wisconsin	683	9	2
6	Backup-large.test	376	35	19
7	Shuttle	58000	9	7
8	Letter-recognition	20000	16	26
9	Ticdata2000	5822	85	2

(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 one key factor of the accelerated algorithm.
(3) Time consumption of computing the stopping criterion is also significantly reduced via gradually decreasing the size of data set. This is the other important factor of the improved algorithm.

Based on the above three speedup factors, we draw such a conclusion that: the general modified algorithm can significantly reduce the computational time of each existing attribute reduction algorithm while producing the same attribute reducts and classification accuracies as those coming from the original ones.

It is deserved to point out that each of these modified algorithms can not solve the overfitting problem in the approximation of concepts and improve the generalization ability of the rough classifier induced by the obtained attribute reduct. The general improved algorithm only devotes to largely reducing the computational time of original attribute reduction algorithms (see Algorithm 1).

4.5. Time efficiency analysis of algorithms

Many heuristic attribute reduction methods have been developed for symbolic data [19,20,22,25,26,39,52,54–56]. The four heuristic algorithms mentioned in Section 4.2 are very representative. The objective of the following experiments is to show the time efficiencies of the proposed general framework for selecting a feature subset. The data used in the experiments are outlined in Table 2, which were all downloaded from UCI Repository of machine learning databases.

In this subsection, in order to compare the above four representative attribute reduction algorithms (PR, SCE, LCE and CCE) with the modified ones, we employ nine UCI data sets from Table 2 to verify the performance of time reduction of the modified algorithms, which are all symbolic data (Shuttle and Ticdata2000 are preprocessed by discretization with entropy). In these nine data sets, Mushroom and Breast-cancer-wisconsin are two data sets with missing values. For uniform treatment of all data sets, we remove the objects with missing values.

From the rank preservation of significance measures of attributes, we know that each modified attribute reduction algorithm can obtain the same attribute reduct as its original version. Therefore, in the following experiments, we only consider attribute reducts obtained and computational time, and do not compare their classification accuracies.

For any heuristic attribute reduction algorithm in rough set theory, the computation of classification is the first key step. To date, many quick classification algorithms have been proposed for improving the efficiency of attribute reduction. However, the computation of classification is only the pretreatment of data in any heuristic attribute reduction algorithm. Hence, for convenient comparison, we will adopt the classification algorithm with the time complexity $O(|C||U|)$ [58]. 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, ..., 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 one 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#*.

Table 3
The time and attribute reduction of the algorithms PR and FSPA-PR.

Data sets	Original features	PR algorithm		FSPA-PR algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Mushroom	22	3	24.8750	3	20.4531
Tic-tac-toe	9	8	0.3594	8	0.3125
Dermatology	34	10	0.8438	10	0.4375
Kr-vs-kp	36	29	28.0313	29	21.5781
Breast-cancer-wisconsin	9	4	0.1250	4	0.0938
Backup-large.test	35	10	0.6563	10	0.4219
Shuttle	9	4	906.0625	4	712.2500
Letter-recognition	16	11	282.6406	11	112.6250
Ticdata2000	85	24	886.4531	24	296.3750

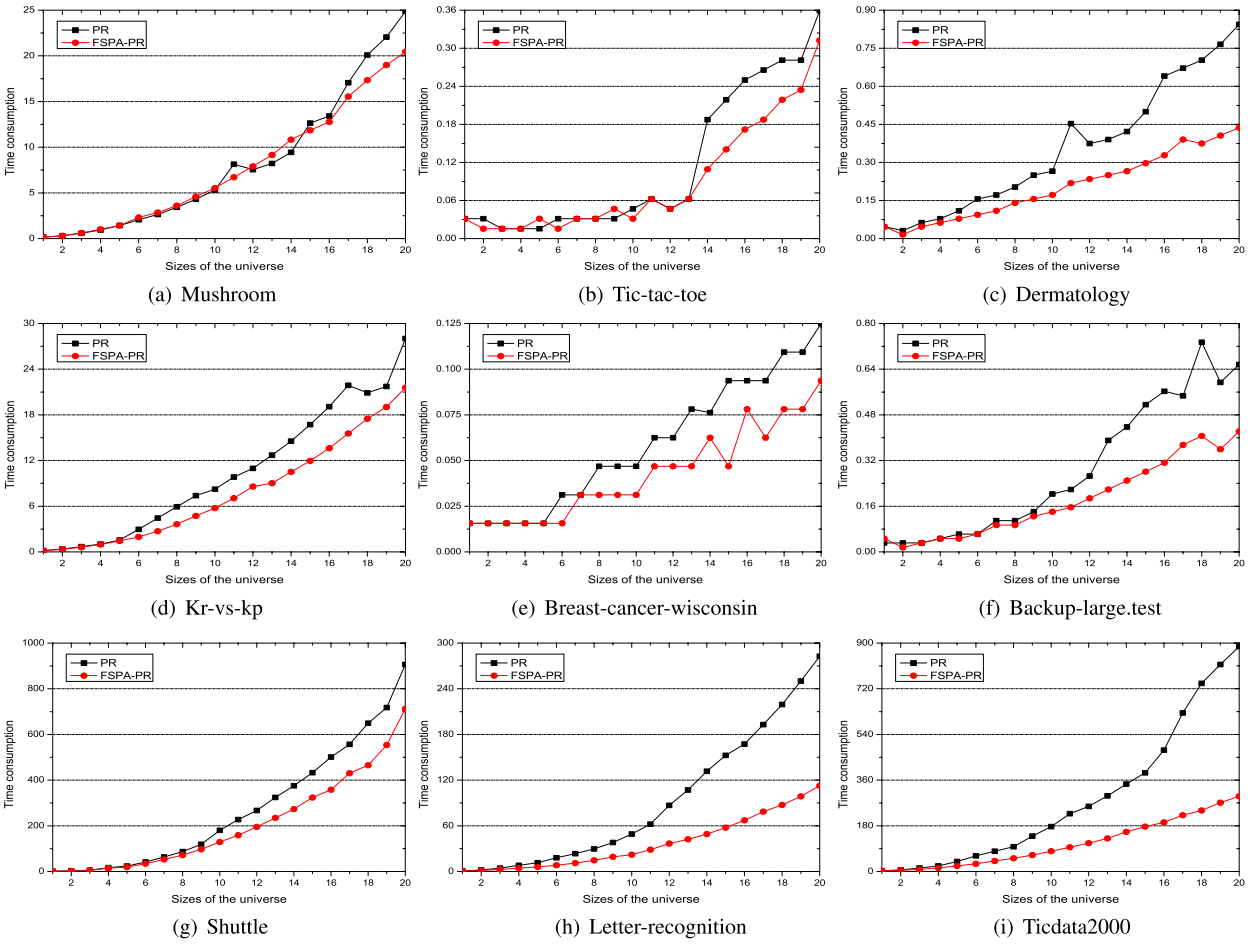


Fig. 4. Times of PR and FSPA-PR versus the size of data.

4.5.1. PR and FSPA-PR

In the sequence of experiments, we compare PR with FSPA-PR on the nine real world data sets shown in Table 2. The experimental results of these nine data sets are shown in Table 3 and Fig. 4. 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 PR and the accelerated algorithm FSPA-PR on nine data sets. While Fig. 4 displays more detailed change trend of each of two algorithms with size of data set becoming increasing.

It is easy to note from Table 3 and Fig. 4 that the computing time of each of these two algorithms increases with the increase of the size of data. Nevertheless this relationship is not strictly monotonic. For example, as the size of data set varies from the 18th to the 19th in sub-figure (f), the computing time dropped. We observe the same effect in sub-figures (c) and (e). One could envision that this situation must have occurred because different numbers of features selected.

As one of the important advantages of the FSPA, as shown in Table 3 and Fig. 4, we see that the modified algorithms are much more faster than their original counterparts on the basis of selecting the same feature subset. Sometimes, the effect of this reduction can reduce over half the computational time. For example, the reduced time achieves 170.0156 seconds on the data set Letter-recognition, while the reduced time is 590.0781 seconds on the data set Ticdata200. Furthermore the differences are profoundly larger when the size of the data set increases. Owing to the rank preservation of significance measures of attributes, the feature subset obtained by each of the modified algorithm is the same as the one produced by the original algorithm.

4.5.2. SCE and FSPA-SCE

It is well known that, the attribute reduct induced by Shannon's information entropy keeps the probabilistic distribution 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 one induced by the positive-region reduction.

In what follows, we compare SCE with FSPA-SCE on those nine real world data sets shown in Table 2 from computational time and selected feature subsets. Table 4 presents the comparisons of selected features and computational time with original algorithm SCE and the accelerated algorithm FSPA-SCE on nine data sets. While Fig. 5 gives more detailed change trendline of each of two algorithms with size of data set becoming increasing.

From Table 4 and Fig. 5, it is easy to see that the modified algorithms is consistently faster than their original counterparts. Sometimes, the reduced time can almost achieves seven-eighths of the original computational time. For example, the reduced time achieves 4275.4531 seconds on the data set Letter-recognition, and the reduced time achieves 7109.7657 seconds on the data set Ticdata2000. In particular, the feature subset obtained by each of the modified algorithm is the same as the one produced by the original algorithm, which benefits from the rank preservation of significance measures of attributes based on the positive approximation. 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.

4.5.3. LCE and FSPA-LCE

The attribute reduct induced by Liang's information entropy also keeps the probabilistic distribution of original data set, which is based on a more strict definition of attribute reduct. The attribute reduct obtained by this approach is often much longer than one induced by the positive-region reduction.

In the following experiments, we will compare LCE with FSPA-LCE on the nine real world data sets shown in Table 2. The comparisons of selected features and computational time with original algorithm SCE and the accelerated algorithm FSPA-SCE on nine data sets are shown in Table 5, and more detailed change trendline of each of two algorithms with size of data set becoming increasing are given in Fig. 6.

As one of the important advantages of the FSPA, as shown in Table 5 and Fig. 6, we see that the modified algorithms are much more faster than their original counterparts. Furthermore the differences are profoundly larger when the size of the data set increases. Sometimes, the computational time of the modified algorithm only is almost one-fifteenths of the computational time of algorithm LCE. On the data set Ticdata2000, for example, FSPA-LCE only needs 1805.5625 seconds, while LCE uses 27962.6250 seconds. Like FSPA-PR and FSPA-SCE, the attribute reduct obtained by the modified algorithm FSPA-LCE is the same as the one proceed by the original algorithm LCE owing to the rank preservation of significance measures of attributes.

4.5.4. CCE and FSPA-CCE

Finally, we compare CCE with FSPA-CCE on the nine real world data sets shown in Table 2. In Table 6, it is shown that the comparisons of selected features and computational time with original algorithm SCE and the accelerated algorithm FSPA-SCE on nine data sets. In Fig. 7, we display more detailed change trendline of each of two algorithms with size of data set becoming increasing. Similarly, in each of these sub-figures (a)–(i), 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.

From Table 6 and Fig. 7, it is easy to see that the modified algorithm is consistently faster than its original counterpart. Sometimes, the reduced time can be over seven-eighths of the original computational time. For example, the reduced time achieves 7213.4688 seconds on the data set Ticdata2000 and the reduced time achieves 4507.9062 seconds on the data set Letter-recognition. In particular, the feature subset obtained by the modified algorithm is the same as the one produced by the original algorithm, which is guaranteed by the rank preservation of significance measures of attributes based on the positive approximation. Furthermore the differences are profoundly larger when the size of the data set increases. One can say that attribute reduction based on the accelerator should be a good solution.

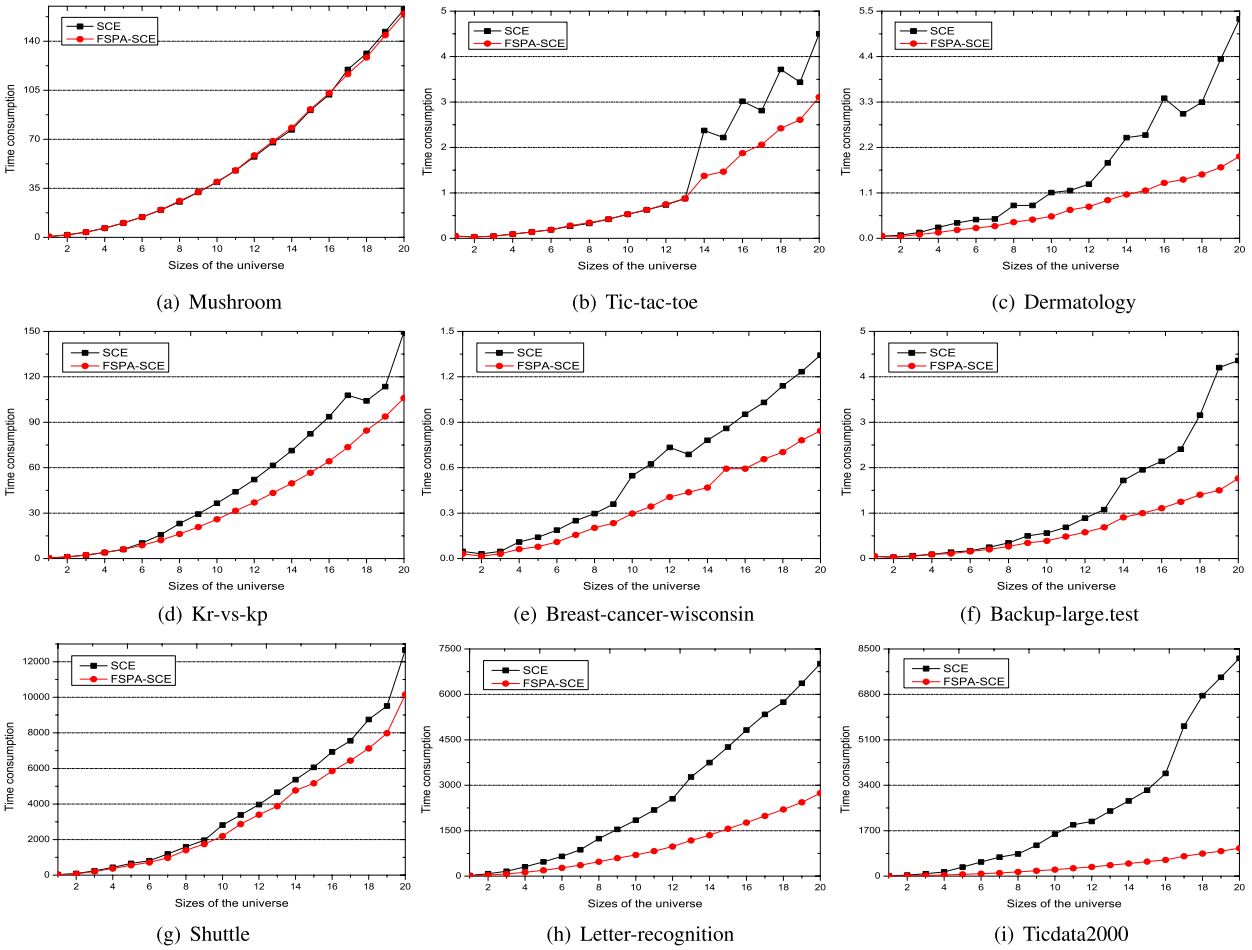
As one of the important advantages of the FSPA, as shown in Table 6 and Fig. 7, we see that the modified algorithm is much faster than its original counterpart. Furthermore the differences are profoundly larger when the size of the data set increases. Owing to the rank preservation of significance measures of attributes, the feature subset obtained by each of the modified algorithm is the same as the one produced by the original algorithm.

4.6. Stability analysis of algorithms

The stability of a heuristic attribute reduction algorithm determines the stability of its classification accuracy. The objective of this suite of experiments is to compare the stability of the computing time and attribute reduction of each of

Table 4
The time and attribute reduction of the algorithms SCE and FSPA-SCE.

Data sets	Original features	SCE algorithm		FSPA-SCE algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Mushroom	22	4	162.6406	4	159.5938
Tic-tac-toe	9	8	4.5000	8	3.1094
Dermatology	34	11	5.3125	11	1.9844
Kr-vs-kp	36	29	149.6250	29	105.9844
Breast-cancer-wisconsin	9	4	1.3438	4	0.8438
Backup-large.test	35	10	4.3594	10	1.7656
Shuttle	9	4	12665.3906	4	10153.1719
Letter-recognition	16	11	7015.7031	11	2740.2500
Ticdata2000	85	24	8153.6563	24	1043.8906



(a) Mushroom (b) Tic-tac-toe (c) Dermatology

(d) Kr-vs-kp (e) Breast-cancer-wisconsin (f) Backup-large.test

(g) Shuttle (h) Letter-recognition (i) Ticdata2000

Fig. 5. Times of SCE and FSPA-SCE versus the size of data.

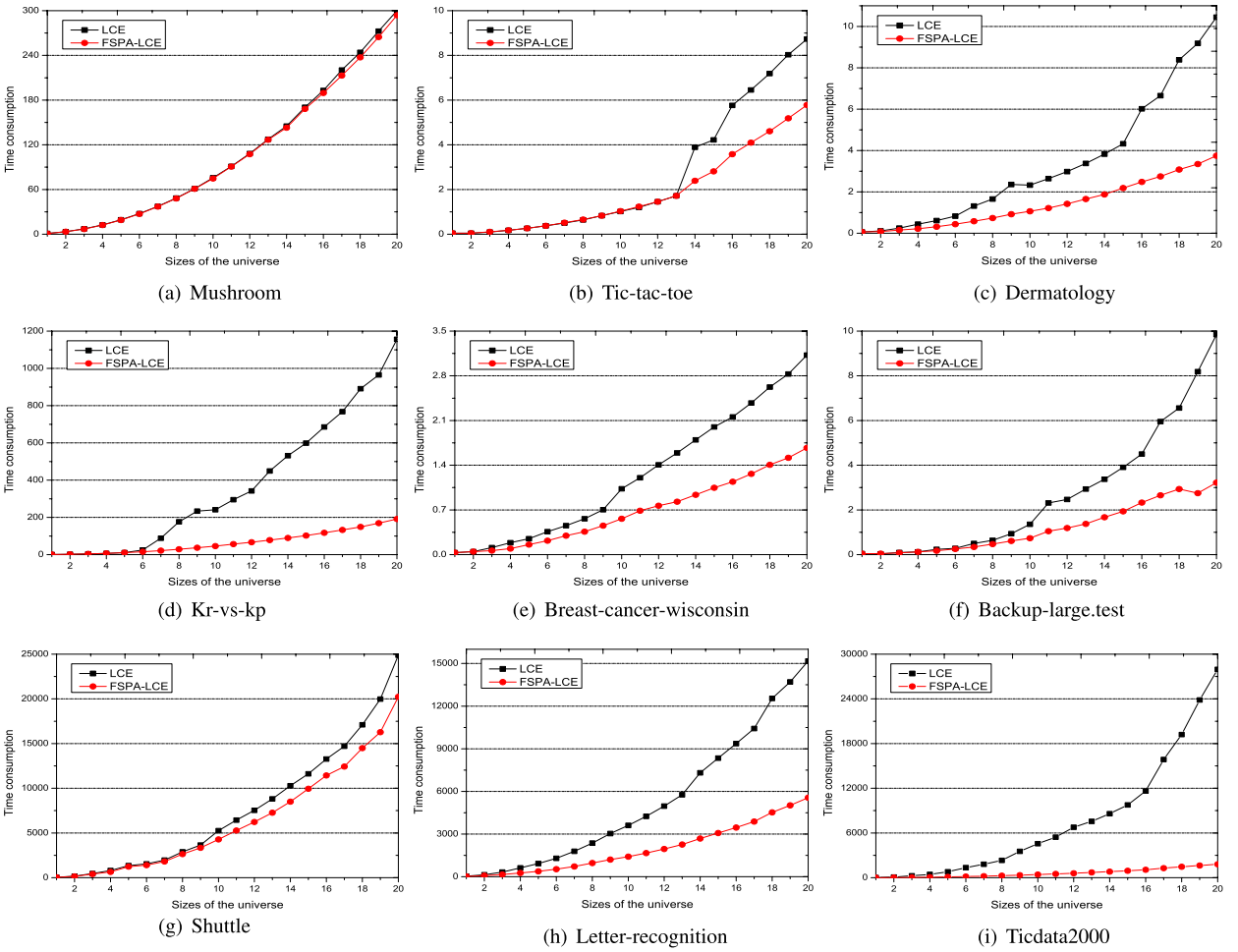
the modified algorithms with those obtained when running the original methods, We use the nine real-world data sets as shown in Table 2.

In the experiments, in order to evaluate the stability of feature subset selected with 10-fold cross validation, we introduce several definitions and necessary notations. Let $X_1, X_2, \ldots, X_{10}$ be the 10 data sets coming from a given universe U. We denote the reduct induced by the universe U by C_0. The reducts induced by the data set X_i will be denoted by C_i ($i \leqslant 10$), respectively. To measure the difference of two reducts C_i and C_j, we use the following distance:

$$D(C_i, C_j) = 1 - \frac{|C_i \cap C_j|}{|C_i \cup C_j|}.$$

Table 5
The time and attribute reduction of the algorithms LCE and FSPA-LCE.

Data sets	Original features	LCE algorithm		FSPA-LCE algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Mushroom	22	4	300.2188	4	294.0000
Tic-tac-toe	9	8	8.7344	8	5.7813
Dermatology	34	10	10.4531	10	3.7500
Kr-vs-kp	36	29	1156.1250	29	191.1250
Breast-cancer-wisconsin	9	5	3.1250	5	1.6719
Backup-large.test	35	10	9.8438	10	3.2188
Shuttle	9	4	24883.6250	4	20228.3906
Letter-recognition	16	12	15176.7656	12	5558.7813
Ticdata2000	85	24	27962.6250	24	1805.5625



(a) Mushroom (b) Tic-tac-toe (c) Dermatology

(d) Kr-vs-kp (e) Breast-cancer-wisconsin (f) Backup-large.test

(g) Shuttle (h) Letter-recognition (i) Ticdata2000

Fig. 6. Times of LCE and FSPA-LCE versus the size of data.

Next we calculate the mean value of the 10 distances:

$$\mu = \frac{1}{10}\sum_{i=1}^{10}\left(1 - \frac{|C_i \cap C_0|}{|C_i \cup C_0|}\right),$$

where C_0 is the reduct induced by the universe U.

This standard deviation

$$\sigma = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(D(C_i, C_0) - \mu\right)^2}$$

Table 6
The time and attribute reduction of the algorithms CCE and FSPA-CCE.

Data sets	Original features	CCE algorithm		FSPA-CCE algorithm	
		Selected features	Time (s)	Selected features	Time (s)
Mushroom	22	4	166.9219	4	159.6406
Tic-tac-toe	9	8	6.7656	8	3.1406
Dermatology	34	10	5.8281	10	2.2656
Kr-vs-kp	36	29	149.7500	29	105.7500
Breast-cancer-wisconsin	9	4	1.3594	4	0.8906
Backup-large.test	35	9	4.5781	9	1.9844
Shuttle	9	4	13718.8750	4	10948.9219
Letter-recognition	16	11	7118.2656	11	2610.3594
Ticdata2000	85	24	8262.0469	24	1048.5781

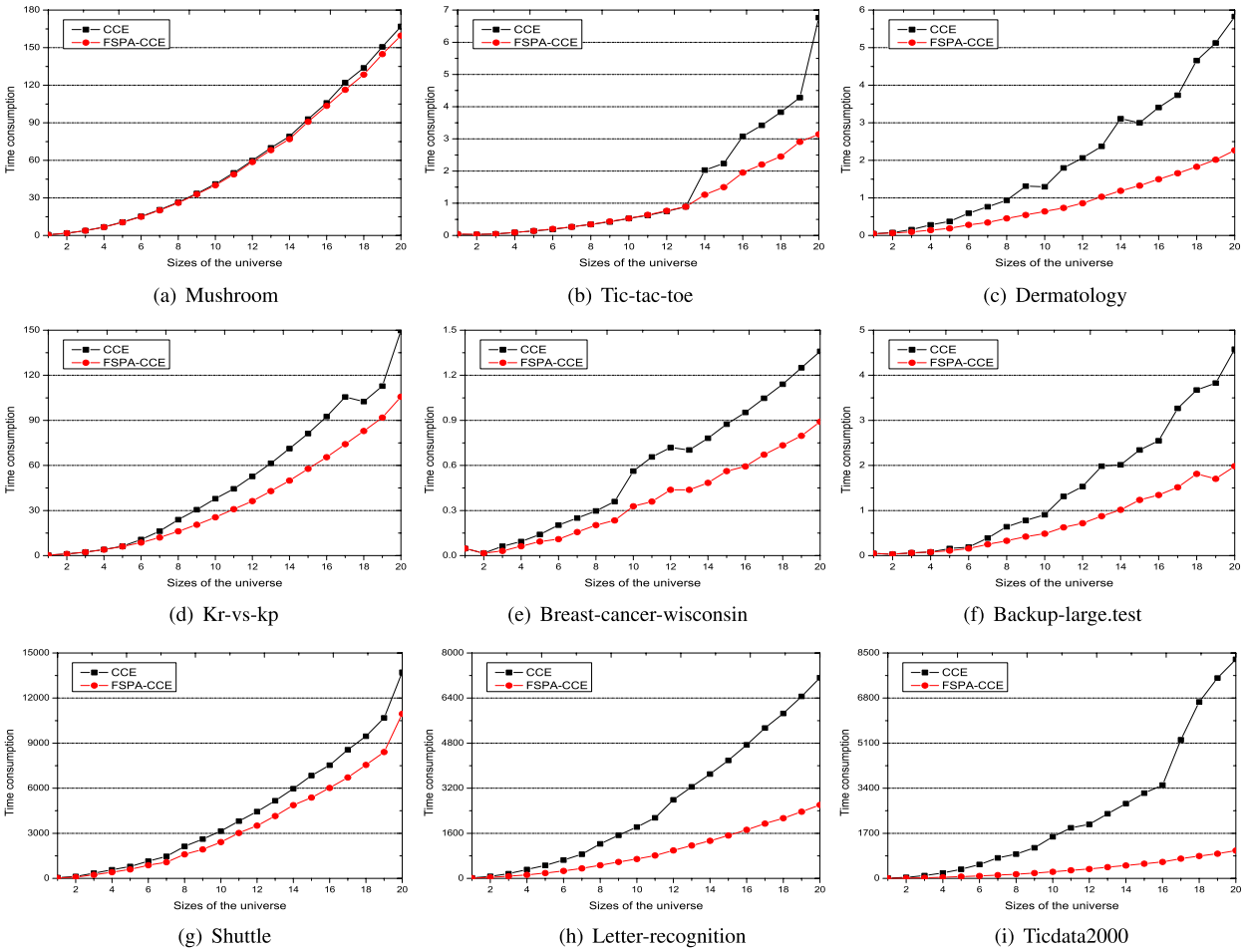


Fig. 7. Times of CCE and FSPA-CCE versus the size of data.

is used to characterize the stability of the reduct result induced by a heuristic attribute reduction algorithm. The lower the value of the standard deviation, the higher the stability of the algorithm. Similarly, we also can use the standard deviation to evaluate the stability of the computing time.

As before we used the same heuristic attribute reduction algorithms along with their modifications. The results reported in Tables 7–10 are obtained using the 10-fold cross validation.

Table 7 reveals that FSPA-PR comes with a far lower mean time and the standard deviation than the ones produced by the original PR. The FSPA-PR's stability is the same as the one reported for the PR. In other words, as an accelerator for attribute reduction, the positive approximation can be used to significantly reduce the time consumption of the algorithm PR. The much smaller standard deviation implies that the modified algorithm FSPA-PR exhibits a far better robustness than the original PR. We also note that the modified algorithm has not affected the stability of reducts induced by the original method (we obtained the same attribute reduct on the same data set). The mechanism can be well interpreted by the

Table 7
The stabilities of the time and attribute reduction of algorithms PR and FSPA-PR.

Data sets	PR's time	FSPA-PR's time	PR's stability	FSPA-PR's stability
Mushroom	16.8359 ± 0.2246	14.8438 ± 0.2130	0.0000 ± 0.0000	0.0000 ± 0.0000
Tic-tac-toe	0.3234 ± 0.0222	0.2391 ± 0.0262	0.0000 ± 0.0000	0.0000 ± 0.0000
Dermatology	0.8234 ± 0.0494	0.3922 ± 0.0109	0.2142 ± 0.1692	0.2142 ± 0.1692
Kr-vs-kp	25.0781 ± 4.3400	16.2438 ± 0.2232	0.0675 ± 0.0652	0.0675 ± 0.0652
Breast-cancer-wisconsin	0.1156 ± 0.0104	0.0813 ± 0.0094	0.1733 ± 0.2736	0.1733 ± 0.2736
Backup-large.test	0.6344 ± 0.0788	0.3891 ± 0.0331	0.4187 ± 0.1830	0.4187 ± 0.1830
Shuttle	778.6959 ± 29.4587	551.6750 ± 10.6770	0.0250 ± 0.0750	0.0250 ± 0.0750
Letter-recognition	224.1219 ± 7.3887	90.5797 ± 1.5252	0.2222 ± 0.2020	0.2222 ± 0.2020
Ticdata2000	698.1016 ± 54.8386	248.8391 ± 6.5261	0.2058 ± 0.0862	0.2058 ± 0.0862

Table 8
The stabilities of the time and attribute reduction of algorithms SCE and FSPA-SCE.

Data sets	SCE's time	FSPA-SCE's time	SCE's stability	FSPA-SCE's stability
Mushroom	130.6234 ± 0.9870	126.1625 ± 0.8873	0.0000 ± 0.0000	0.0000 ± 0.0000
Tic-tac-toe	3.8359 ± 0.0614	2.5045 ± 0.0617	0.1111 ± 0.1111	0.1111 ± 0.1111
Dermatology	4.0500 ± 0.3197	1.6266 ± 0.0422	0.5312 ± 0.1000	0.5312 ± 0.1000
Kr-vs-kp	126.7734 ± 15.7752	83.2891 ± 0.9501	0.0675 ± 0.0652	0.0675 ± 0.0652
Breast-cancer-wisconsin	1.2156 ± 0.0894	0.7500 ± 0.0677	0.3562 ± 0.3099	0.3562 ± 0.3099
Backup-large.test	3.7234 ± 0.3919	1.4188 ± 0.0655	0.3599 ± 0.2521	0.3599 ± 0.2521
Shuttle	9749.1705 ± 308.8128	8158.8490 ± 209.5685	0.0250 ± 0.0750	0.0250 ± 0.0750
Letter-recognition	5891.5906 ± 181.0442	2282.8141 ± 73.0362	0.1689 ± 0.1823	0.1689 ± 0.1823
Ticdata2000	7107.3904 ± 105.7970	861.2000 ± 9.7081	0.2485 ± 0.0830	0.2485 ± 0.0830

Table 9
The stabilities of the time and attribute reduction of algorithms LCE and FSPA-LCE.

Data sets	LCE's time	FSPA-LCE's time	LCE's stability	FSPA-LCE's stability
Mushroom	241.9891 ± 1.3425	236.0313 ± 1.6868	0.0000 ± 0.0000	0.0000 ± 0.0000
Tic-tac-toe	7.3328 ± 0.0601	4.7531 ± 0.1007	0.1778 ± 0.0889	0.1778 ± 0.0889
Dermatology	8.2875 ± 0.6289	3.0938 ± 0.0617	0.1852 ± 0.1783	0.1852 ± 0.1783
Kr-vs-kp	228.9547 ± 27.4934	154.4984 ± 2.0417	0.0675 ± 0.0652	0.0675 ± 0.0652
Breast-cancer-wisconsin	2.5969 ± 0.0493	1.4031 ± 0.0554	0.2333 ± 0.1528	0.2333 ± 0.1528
Backup-large.test	7.9094 ± 0.4949	2.7109 ± 0.1746	0.1617 ± 0.1630	0.1617 ± 0.1630
Shuttle	17717.9594 ± 391.4628	14392.4496 ± 99.2163	0.0250 ± 0.0750	0.0250 ± 0.0750
Letter-recognition	12334.2729 ± 80.6504	4252.5578 ± 71.4054	0.1914 ± 0.1436	0.1914 ± 0.1436
Ticdata2000	19582.6515 ± 385.2873	1463.7391 ± 14.5646	0.1744 ± 0.1192	0.1744 ± 0.1192

Table 10
The stabilities of the time and attribute reduction of algorithms CCE and FSPA-CCE.

Data sets	CCE's time	FSPA-CCE's time	CCE's stability	FSPA-CCE's stability
Mushroom	133.9672 ± 0.9331	129.3531 ± 1.2343	0.0000 ± 0.0000	0.0000 ± 0.0000
Tic-tac-toe	3.8391 ± 0.0297	2.5172 ± 0.0439	0.1778 ± 0.0889	0.1778 ± 0.0889
Dermatology	4.6469 ± 0.3029	1.8016 ± 0.0335	0.2735 ± 0.1698	0.2735 ± 0.1698
Kr-vs-kp	130.0047 ± 17.5668	86.3641 ± 0.9297	0.0733 ± 0.0780	0.0733 ± 0.0780
Breast-cancer-wisconsin	1.1969 ± 0.0865	0.7406 ± 0.0298	0.1200 ± 0.1600	0.1200 ± 0.1600
Backup-large.test	3.8016 ± 0.3155	1.5875 ± 0.1018	0.3426 ± 0.1780	0.3426 ± 0.1780
Shuttle	9564.8752 ± 68.5368	7440.3281 ± 25.0001	0.0250 ± 0.0750	0.0250 ± 0.0750
Letter-recognition	5956.0833 ± 43.7866	2171.0000 ± 36.5273	0.1370 ± 0.1450	0.1370 ± 0.1450
Ticdata2000	6726.4778 ± 42.1287	859.4672 ± 10.7790	0.1742 ± 0.0894	0.1742 ± 0.0894

rank preservation of the significance measures of attributes used in the algorithms PR and FSPA-PR (see Theorem 5 and Theorem 6). From Tables 8–10, one draw the same conclusions.

4.7. Related discussion

In this subsection, we summarize the advantages of the accelerator-positive approximation for attribute reduction and offer some explanatory comments. Based on the experimental evidence, we can affirm that:

- Each of the accelerated algorithms preserves the attribute reduct induced by the corresponding original one.

In Section 4.3, we have proved the rank preservation of the four significance measures of attributes, which implies that one selects the same attribute in each loop of each of modified algorithms and that of the corresponding original one. Naturally, one can obtain the same attribute reduct on the same data set. Hence, each of the accelerated algorithms does not affect 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-positive approximation, the size of data set could be reduced in each loop of each of modified algorithms. Therefore, the computational time for determining partitions, significance measures of attributes and judging stopping criterion in the reduced data set would be much smaller than that encountered for the entire data set. Evidently, these modified algorithms outperform the previous methods.

- 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.

5. Conclusions

To overcome the limitations of the existing heuristic attribute reduction schemes, in this study, a theoretic framework based on rough set theory have been proposed, called the positive approximation, which can be used to accelerate algorithms of heuristic attribute reduction. Based on this framework, a general heuristic feature selection algorithm (FSPA) has been presented. Several representative heuristic attribute reduction algorithms encountered in rough set theory have been revised and modified. Note that each of the modified algorithms can choose the same feature subset as the original attribute reduction algorithm. Experimental studies pertaining to nine UCI data sets show that the modified algorithms can significantly reduce computing time of attribute reduction while producing the same attribute reducts and classification accuracies as those coming from the original methods. The results show that the attribute reduction based on the positive approximation is an effective accelerator and can efficiently obtain an attribute reduct.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. The authors also wish to thank PhD candidate Feng Wang for numeric experiments and data statistics for several months and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education for the usage of about 1000 CPU hours on two computers (Inter(R) Core(TM)2 Quad CPU Q9400, 2.66 GHz and 3.37 GB memory) for the empirical study.

This work was supported by the National Natural Science Foundation of China (Nos. 60773133, 60903110, 70971080), National Key Basic Research and Development Program of China (973) (No. 2007CB311002), GRF: CityU 113308 of the Government of Hong Kong SAR, the National High Technology Research and Development Program of China (No. 2007AA01Z165), and the Natural Science Foundation of Shanxi Province, China (Nos. 2008011038, 2009021017-1).

References

[1] J.G. Bazan, A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery 1: Methodology and Applications, Studies in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg, Germany, 1998, pp. 321–365.
[2] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problems, in: L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Springer-Verlag, Heidelberg, Germany, 2000, pp. 49–88.
[3] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Letters 26 (2005) 965–975.
[4] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters 26 (2005) 1632–1640.
[5] M.R. Chmielewski, J.W. Grzymala Busse, Global discretization of continuous attributes as preprocessing for machine learning, International Journal of Approximate Reasoning 15 (4) (1996) 319–331.
[6] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[7] I. Düntsch, G. Gediga, Uncertainty measures of rough set prediction, Artificial Intelligence 106 (1998) 109–137.
[8] G. Gediga, I. Düntsch, Rough approximation quality revisited, Artificial Intelligence 132 (2001) 219–234.
[9] J.W. Grzymala-Busse, An algorithm for computing a single covering, in: J.W. Grzymala-Busse (Ed.), Managing Uncertainty in Expert Systems, Kluwer Academic Publishers, 1991, p. 66.

[10] J.W. Grzymala-Busse, LERS—a system for learning from examples based on rough sets, in: R. Slowinski (Ed.), Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, 1992, pp. 3–18.
[11] J.W. Guan, D.A. Bell, Rough computational methods for information systems, Artificial Intelligence 105 (1998) 77–103.
[12] I. Guyon, A. Elisseeff, An introduction to variable feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[13] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions on Knowledge and Data Engineering 16 (12) (2004) 1457–1471.
[14] R. Jensen, Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE Press/Wiley & Sons, 2008.
[15] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, Proc. AAAI 92 (1992) 129–134.
[16] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[17] M. Kryszkiewicz, P. Lasek, FUN: fast discovery of minimal sets of attributes functionally determining a decision attribute, Transactions on Rough Sets 9 (2008) 76–95.
[18] M. Kryszkiewicz, Comparative study of alternative type of knowledge reduction in inconsistent systems, International Journal of Intelligent Systems 16 (2001) 105–120.
[19] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, International Journal of Computational Intelligence 11 (2) (1995) 323–338.
[20] Q.H. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3509–3521.
[21] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2) (2006) 191–201.
[22] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (5) (2006) 414–423.
[23] C.K. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Information Processing and Management 42 (2006) 155–165.
[24] D.Y. Li, B. Zhang, Y. Leung, On knowledge reduction in inconsistent decision information systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12 (5) (2004) 651–672.
[25] J.Y. Liang, K.S. Chin, C.Y. Dang, C.M. Yam Richid, A new method for measuring uncertainty and fuzziness in rough set theory, International Journal of General Systems 31 (4) (2002) 331–342.
[26] J.Y. Liang, Z.B. Xu, The algorithm on knowledge reduction in incomplete information systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (1) (2002) 95–103.
[27] J.Y. Liang, Y.H. Qian, C.Y. Chu, D.Y. Li, J.H. Wang, Rough set approximation based on dynamic granulation, Lecture Notes in Computer Science 3641 (2005) 701–708.
[28] H. Liu, R. Setiono, Feature selection via discretization, IEEE Transactions on Knowledge and Data Engineering 9 (4) (1997) 642–645.
[29] J.S. Mi, W.Z. Wu, W.X. Zhang, Comparative studies of knowledge reductions in inconsistent systems, Fuzzy Systems and Mathematics 17 (3) (2003) 54–60.
[30] M. Modrzejewski, Feature selection using rough set theory, in: Proceedings of European Conference on Machine Learning, 1993, pp. 213–226.
[31] H.S. Nguyen, Approximate Boolean reasoning: Foundations and applications in data mining, Lecture Notes in Computer Science 3100 (2006) 334–506.
[32] T. Pavlenko, On feature selection, curse-of-dimensionality and error probability in discriminant analysis, Journal of Statistical Planning and Inference 115 (2003) 565–584.
[33] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, 1991.
[34] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (1) (2007) 3–27.
[35] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (2007) 28–40.
[36] Z. Pawlak, A. Skowron, Rough sets and boolean reasoning, Information Sciences 177 (1) (2007) 41–73.
[37] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognition 35 (2002) 825–834.
[38] L. Polkowski, On convergence of rough sets, in: R. Slowinski (Ed.), Intelligent Decision Support: Handbook of Applications and Advances of Rough Set Theory, vol. 11, Kluwer, Dordrecht, 1992, pp. 305–311.
[39] Y.H. Qian, J.Y. Liang, Combination entropy and combination granulation in rough set theory, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16 (2) (2008) 179–193.
[40] Y.H. Qian, J.Y. Liang, C.Y. Dang, Converse approximation and rule extraction from decision tables in rough set theory, Computer and Mathematics with Applications 55 (2008) 1754–1765.
[41] Y.H. Qian, J.Y. Liang, C.Y. Dang, Consistency measure, inclusion degree and fuzzy measure in decision tables, Fuzzy Sets and Systems 159 (2008) 2353–2377.
[42] Y.H. Qian, J.Y. Liang, C.Y. Dang, Interval ordered information systems, Computer and Mathematics with Applications 56 (2008) 1994–2009.
[43] Y.H. Qian, J.Y. Liang, C.Y. Dang, D.W. Tang, Set-valued ordered information systems, Information Sciences 179 (2009) 2809–2832.
[44] Y.H. Qian, J.Y. Liang, D.Y. Li, H.Y. Zhang, C.Y. Dang, Measures for evaluating the decision performance of a decision table in rough set theory, Information Sciences 178 (2008) 181–202.
[45] Y.H. Qian, J.Y. Liang, C.Y. Dang, Incomplete multi-granulations rough set, IEEE Transactions on Systems, Man and Cybernetics: Part A 40 (2) (2010) 420–431.
[46] R. Quinlan, Induction of decision rules, Machine Learning 1 (1) (1986) 81–106.
[47] M.W. Shao, W.X. Zhang, Dominance relation and rules in an incomplete ordered information system, International Journal of Intelligent Systems 20 (2005) 13–27.
[48] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (2004) 1351–1363.
[49] A. Skowron, Extracting laws from decision tables: a rough set approach, Computational Intelligence 11 (1995) 371–388.
[50] D. Slezak, Approximate reducts in decision tables, Research report, Institute of Computer Science, Warsaw University of Technology, 1995.
[51] D. Slezak, Foundations of entropy-based Bayesian networks: theoretical results & rough set based extraction from data, in: IPMU'00, Proceedings of the 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, vol. 1, Madrid, Spain, 2000, pp. 248–255.
[52] D. Slezak, Approximate entropy reducts, Fundamenta Informaticae 53 (3–4) (2002) 365–390.
[53] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.
[54] G.Y. Wang, H. Yu, D.C. Yang, Decision table reduction based on conditional information entropy, Chinese Journal of Computer 25 (7) (2002) 759–766.
[55] G.Y. Wang, J. Zhao, J.J. An, A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundamenta Informaticae 68 (3) (2005) 289–301.
[56] S.X. Wu, M.Q. Li, W.T. Huang, S.F. Liu, An improved heuristic algorithm of attribute reduction in rough set, Journal of System Sciences and Information 2 (3) (2004) 557–562.

[57] W.Z. Wu, M. Zhang, H.Z. Li, J.S. Mi, Knowledge reduction in random information systems via Dempster–Shafer theory of evidence, Information Sciences 174 (2005) 143–164.
[58] Z.Y. Xu, Z.P. Liu, B.R. Yang, W. Song, A quick attribute reduction algorithm with complexity of $\max(O(|C||U|), O(|C|^2|U/C|))$, Chinese Journal of Computer 29 (3) (2006) 391–398.
[59] Y.Y. Yao, Information granulation and rough set approximation, International Journal of Intelligent Systems 16 (1) (2001) 87–104.
[60] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.