Pattern Recognition 44 (2011) 1658–1670

Contents lists available at ScienceDirect

Pattern Recognition

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

An efficient accelerator for attribute reduction from incomplete data in rough set framework ☆

Yuhua Qian [a,b], Jiye Liang [a,*], Witold Pedrycz [c], 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
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

ARTICLE INFO

Article history:
Received 7 December 2009
Received in revised form
16 February 2011
Accepted 18 February 2011
Available online 24 February 2011

Keywords:
Feature selection
Rough set theory
Incomplete information systems
Positive approximation
Granular computing

ABSTRACT

Feature selection (attribute reduction) from large-scale incomplete data is a challenging problem in areas such as pattern recognition, machine learning and data mining. In rough set theory, feature selection from incomplete data aims to retain the discriminatory power of original features. To address this issue, many feature selection algorithms have been proposed, however, these algorithms are often computationally time-consuming. To overcome this shortcoming, we introduce in this paper a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. As an application of the proposed accelerator, a general feature selection algorithm is designed. By integrating the accelerator into a heuristic algorithm, we obtain several modified representative heuristic feature selection algorithms in rough set theory. 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, both the number and dimensionality of items in data sets have grown dramatically. For examples, tens, hundreds, and even thousands of attributes are stored in many real-world application databases [1–3]. It is well known that attributes irrelevant to recognition tasks may deteriorate the performance of learning algorithms [4]. In other words, storing and processing all attributes (both relevant and irrelevant) could be computationally very expensive and impractical. To address this issue, as pointed out in [5], some attributes can be omitted, which will not seriously affect the resulting classification (recognition) accuracy. Therefore, the omission of some attributes could be not only tolerable but also even desirable relative to the computational costs involved [6].

In feature selection, there are two general strategies, namely wrappers [7] and filters. The former employs a learning algorithm to evaluate the selected attribute subsets, and the latter selects attributes according to some significance measure such as information gain [8], consistency [9], distance [10], dependency [11], and others. These measures can be divided into two main categories: distance-based measures and consistency-based measures [5]. 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 discerning ability of original features for the objects from the universe [12,13].

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 hybrid attributes, several approaches have been developed in the literature [12,14–21]. In classical rough set theory, the attribute reduction algorithm takes all attributes to be symbolic values. After preprocessing original data, one can use classical rough set theory to select a subset of features that is most suitable for a given recognition problem.

Feature selection based on rough set theory starts from a data table, which is also called an information system and contains data about objects of interest that are characterized by a finite set of attributes. According to whether or not there are missing data (null values), information systems are classified into two

☆This is an extended version of the paper presented at the Eighth IEEE International Conference on Machine Learning and Cybernetics, Baoding, 2009, China.

* Corresponding author. Tel./fax: +86 0351 7018176.
E-mail addresses: jinchengqyh@126.com (Y. Qian), ljy@sxu.edu.cn (J. Liang), pedrycz@ee.ualberta.ca (W. Pedrycz), mecdang@cityu.edu.hk (C. Dang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.020

categories: complete and incomplete. In general, by an incomplete information system, we mean a system with missing data (null values) [22,23]. For an incomplete information system, if condition attributes and decision attributes are distinguished from each other, then it is called an incomplete decision table. Feature selection from incomplete data usually starts from incomplete decision tables.

In the last two decades, many techniques for attribute reduction have been developed in rough set theory [25–29]. Especially £ in order to obtain all attribute reducts of a given data set, Skowron proposed a discernibility matrix method [30]. However, these feature selection algorithms are usually time-consuming to process large-scale data. Aiming at efficient feature selection, many heuristic attribute reduction algorithms have been developed in rough set theory, cf. [5,19,31–37]. Each of these algorithms preserves a particular property of a given information system. To accomplish attribute reduction from incomplete decision tables, similar to the discernibility matrix proposed by Skowron, Kryszkiewicz gave a generalized discernibility matrix to obtain all attribute reducts of an incomplete decision table [24]. To efficiently obtain an attribute reduct, several heuristic attribution reduction approaches have been presented [38–41]. For convenience of our further development, we review several representative heuristic attribute reduction algorithms in the context of incomplete data here. Applying the idea of positive-region reduction, Yang and Shu proposed a heuristic feature selection algorithm in incomplete decision tables, which keeps the positive region of target decision unchanged [39]. Liang et al. defined new information entropy to measure the uncertainty of an incomplete information system [23] and applied the corresponding conditional entropy to reduce redundant features [38]. Qian and Liang [40] presented the combination entropy for measuring the uncertainty of an incomplete information system and used its conditional entropy to obtain a feature subset. As Shannon's information entropy was introduced to search reducts in classical rough set model [34], an extension of its conditional entropy also can be used to calculate a relative attribute reduct of an incomplete decision table. However, the above algorithms are still computationally time-consuming to deal with large-scale data sets.

In this study, we will not consider how to discretize numerical attributes and construct a heuristic function for feature selection. Our objective is how to improve computational efficiency of a heuristic attribute reduction algorithm in the context of incomplete data. A brief version of this work has been published in the literature [42]. In this extended version, we propose a new rough set framework, which is called positive approximation in incomplete information systems. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure of an incomplete rough set using a granulation order. Based on the positive approximation, we develop a common accelerator for improving computational efficiency of a heuristic feature selection, which provides a vehicle of making rough set-based feature selection algorithms faster. By incorporating the accelerator into each of the above representative heuristic attribute reduction algorithms, we obtain their modified versions. Numerical experiments show that each of the modified algorithms can choose the same feature subset as that of the corresponding original algorithm while greatly reducing computational time. Furthermore, we would like to stress that the improvement becomes more visible when the data sets become larger.

The rest of the paper is organized as follows. Some basic concepts are briefly reviewed in Section 2, which include incomplete information systems, incomplete rough set model, incomplete variable precision rough set model and partial relations. In Section 3, we establish a positive approximation framework in incomplete information systems and investigate some of its main properties. In Section 4, by analyzing the rank preservation of several representative significance measures of attributes, a general algorithm based on the positive approximation is first introduced, and a series of experimental studies are then conducted, which focus on comparison of computational efficiency and stability of the selected attributes. Finally, Section 5 concludes the paper with some remarks and discussions.

2. Preliminaries

In this section, we will review several basic concepts such as incomplete information systems, tolerance relation and partial relation. Throughout this paper, we suppose that the universe U is a finite non-empty set.

An information system is a pair $S=(U,A)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, and for every $a\in A$, there is a mapping a, $a: U\to V_a$, where V_a is called the value set of a.

It may happen that some of attribute values for an object are missing. To distinguish such a situation from the other, a so-called null value, denoted by $*$, is usually assigned to those attributes. If V_a contains a null value for at least one attribute $a\in A$, then S is called an incomplete information system; otherwise it is a complete one [24].

Let $S=(U,A)$ be an information system and $P\subseteq A$ an attribute set. We define a binary relation on U as follows:

$$SIM(P)=\{(u,v)\in U\times U|\ \forall a\in P,\ a(u)=a(v)\quad \text{or } a(u)=*\quad \text{or } a(v)=*\}.$$

In fact, $SIM(P)$ is a tolerance relation on U. The concept of a tolerance relation has a wide variety of applications in classification. It can be easily shown that $SIM(P)=\bigcap_{a\in P}SIM(\{a\})$. Let $S_P(u)$ denote the set $\{v\in U|(u,v)\in SIM(P)\}$. $S_P(u)$ is the maximal set of objects which are possibly indistinguishable by P with u. Let $U/SIM(P)$ denote the family sets $\{S_P(u)|u\in U\}$, which is the classification or the knowledge induced by P. A member $S_P(u)$ from $U/SIM(P)$ will be called a tolerance class or an information granule. It should be noticed that the tolerance classes in $U/SIM(P)$ do not yield a partition of U in general. They form a cover of U, i.e., $S_P(u)\neq\emptyset$ for every $u\in U$, and $\bigcup_{u\in U}S_P(u)=U$.

Let $S=(U,A)$ be an incomplete information system, X a subset of U, and $P\subseteq A$ an attribute set. In the rough set model, based on the tolerance relation [24,43], X can be characterized by $\overline{SIM(P)}X$ and $\underline{SIM(P)}X$, where

$$\begin{cases}\underline{SIM(P)}X=\cup\{Y\in U/SIM(P)|Y\subseteq X\},\\ \overline{SIM(P)}X=\cup\{Y\in U/SIM(P)|Y\cap X\neq\emptyset\}.\end{cases}$$

There are two kinds of attributes for a classification problem. Each of them 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. Assume the objects are partitioned into r mutually exclusive crisp subsets $\{X_1,X_2,\ldots,X_r\}$ by the decision attribute set D. Given any subset $P\subseteq C$ and the tolerance relation $SIM(P)$ induced by P, one can then define the lower and upper approximations of the decision attribute set D as

$$\begin{cases}\underline{SIM(P)}D=\{\underline{SIM(P)}X_1,\underline{SIM(P)}X_2,\ldots,\underline{SIM(P)}X_r\},\\ \overline{SIM(P)}D=\{\overline{SIM(P)}X_1,\overline{SIM(P)}X_2,\ldots,\overline{SIM(P)}X_r\}.\end{cases}$$

Let $POS_P(D)=\bigcup_{i=1}^{r}\underline{SIM(P)}X_i$, which is called the positive region of D with respect to the condition attribute set P.

To formulate the variable precision rough set model, Ziako [44] used the relative degree of misclassification function c and the granule-based definition of approximation. Using the approximating method of the variable precision rough set model, we will introduce an incomplete variable precision rough set framework for more flexible feature selection. Let $S=(U,A)$ be an incomplete information system, X a subset of U, and $P\subseteq A$ an attribute set. Given a threshold $\beta\in[0,0.5]$, X is approximated by $\underline{SIM(P)}^{\beta}X$ and $\overline{SIM(P)}^{\beta}X$, where

$$\begin{cases}\underline{SIM(P)}^{\beta}X=\{u|D(S_P(u),S_P(u)\cap X)\le\beta,u\in X\},\\ \overline{SIM(P)}^{\beta}X=\{u|D(S_P(u),S_P(u)\cap X)\le 1-\beta,u\in U\}..\end{cases}$$

Unlike the classical variable rough set model, the proposed incomplete variable rough set model has the property that $\underline{SIM(P)}^{\beta}X\subseteq X\subseteq\overline{SIM(P)}^{\beta}X$.

Assume the objects are partitioned into r mutually exclusive crisp subsets $\{X_1,X_2,\ldots,X_r\}$ by the decision attribute set D. Given any subset $P\subseteq C$ and the tolerance relation $SIM(P)$ induced by P, one can then define the lower and upper approximations of the decision attribute set D as

$$\begin{cases}\underline{SIM(P)}^{\beta}D=\{\underline{SIM(P)}^{\beta}X_1,\underline{SIM(P)}^{\beta}X_2,\ldots,\underline{SIM(P)}^{\beta}X_r\},\\ \overline{SIM(P)}^{\beta}D=\{\overline{SIM(P)}^{\beta}X_1,\overline{SIM(P)}^{\beta}X_2,\ldots,\overline{SIM(P)}^{\beta}X_r\}..\end{cases}$$

Similar to the definition of positive region in the incomplete rough set model, we come to the definition of positive region of a variable rough decision as $POS_P^{\beta}(D)=\bigcup_{i=1}^{r}\underline{SIM(P)}^{\beta}X_i$, which is called the β-positive region of D with respect to the condition attribute set P. Using the β-positive region, one can construct a new heuristic function for feature selection from incomplete data in the context of variable rough set framework.

Let $S=(U,A)$ be an incomplete information system. We define a partial relation $\preceq$ (or $\succcurlyeq$) on 2^A as follows [23,45]: we say that Q is coarser than P (or P is finer than Q), denoted by $P\preceq Q$ (or $Q\succcurlyeq P$), if and only if $S_P(u_i)\subseteq S_Q(u_i)$ for $i\in\{1,2,\ldots,|U|\}$. If $P\preceq Q$ and $P\neq Q$, we say that Q is strictly coarser than P (or P is strictly finer than Q) and denoted by $P\prec Q$ (or $Q\succ P$). In fact, $P\prec Q\Leftrightarrow$ for $i\in\{1,2,\ldots,|U|\}$, it follows that $S_P(u_i)\subseteq S_Q(u_i)$, and there exists $j\in\{1,2,\ldots,|U|\}$ such that $S_P(u_j)\subset S_Q(u_j)$.

3. Positive approximation in incomplete information systems

For a given incomplete data set, a cover induced by a tolerance relation provides a granulation world for describing a target concept. Therefore, 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 fine to coarse, then the sequence is called converse granulation worlds.

In this section, we introduce a new set-approximation method called positive approximation to incomplete information systems and investigate some of its important properties, in which a target concept is approximated by a positive granulation world. These concepts and properties will be helpful for understanding the notion of a granulation order and set approximation under a granulation order in the context of incomplete data.

Definition 1. Let $S=(U,A)$ be an incomplete information system, $X\subseteq U$, and $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1\succcurlyeq P_2\succcurlyeq\cdots\succcurlyeq P_n$ ($P_i\in 2^A$). Given $\mathbf{P}_i=\{P_1,P_2,\ldots,P_i\}$, we define $\mathbf{P}_i$-lower approximation $\underline{\mathbf{P}_i}(X)$ and $\mathbf{P}_i$-upper approximation $\overline{\mathbf{P}_i}(X)$ of $\mathbf{P}_i$-positive approximation of X as

$$\begin{cases}\underline{\mathbf{P}_i}(X)=\bigcup_{k=1}^{i}\underline{SIM(P_k)}X_k,\\ \overline{\mathbf{P}_i}(X)=\overline{SIM(P_i)}X,\end{cases}$$

where $X_1=X$ and $X_k=X-\bigcup_{j=1}^{k-1}\underline{\mathbf{P}_j}(X_j)$ for $k=2,3,\ldots,i$, $i=1,2,\ldots,n$.

Definition 1 shows that a target concept can be approached by the change of the lower approximation $\underline{\mathbf{P}_i}(X)$ and the upper approximation $\overline{\mathbf{P}_i}(X)$.

In order to illustrate the essence that positive approximation is mainly concerned with the change of the construction of the target concept X (tolerance classes in lower approximation of X with respect to $\mathbf{P}$) in incomplete information systems, we can redefine $\mathbf{P}$-positive approximation of X by using tolerance classes on U. Therefore, the structures of $\mathbf{P}$-lower approximation $\underline{\mathbf{P}}(X)$ and $\mathbf{P}$-upper approximation $\overline{\mathbf{P}}(X)$ of $\mathbf{P}$-positive approximation of X can be represented as follows:

$$\{\langle\underline{\mathbf{P}}(X)\rangle=\{S_{P_i}(u)\mid S_{P_i}(u)\subseteq X_i,i\le n,u\in U\},\langle\overline{\mathbf{P}}(X)\rangle=\{S_{P_n}(u)\mid S_{P_n}(u)\cap X\neq\emptyset,u\in U\},$$

where $X_1=X$, $X_i=X-\bigcup_{k=1}^{i-1}\underline{SIM(P_k)X_k}$ for $i=2,\ldots,n$, and $\langle\cdot\rangle$ denotes the structure of a rough approximation.

In the following, we show how positive approximation in an incomplete information system works through an illustrative example.

Example 1. Suppose that $S=(U,A)$ is an incomplete information system with $U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$, $P,Q\subseteq A$ are two attribute sets, $X=\{u_1,u_2,u_3,u_5,u_6\}$, $SIM(P)=\{\{u_1,u_2\},\{u_1,u_2\},\{u_2,u_3\},\{u_3,u_4,u_5\},\{u_4,u_5,u_6\},\{u_4,u_5,u_6\}\}$, $SIM(Q)=\{\{u_1\},\{u_2\},\{u_3\},\{u_4,u_5\},\{u_4,u_5\},\{u_5,u_6\}\}$. Obviously, $P\succcurlyeq Q$ holds. Hence, we can construct a granulation order (a family of tolerance relations) $\mathbf{P}=\{P,Q\}$, where $\mathbf{P}_1=\{P\}$ and $\mathbf{P}_2=\{P,Q\}$. Computing the positive approximation of X with respect to $\mathbf{P}$, we obtain that

$$\langle\underline{\mathbf{P}_1}(X)\rangle=\{\{u_1,u_2\},\{u_1,u_2\},\{u_2,u_3\}\},$$

$$\langle\overline{\mathbf{P}_1}(X)\rangle=\{\{u_1,u_2\},\{u_1,u_2\},\{u_2,u_3\},\{u_3,u_4,u_5\},\{u_4,u_5,u_6\},\{u_4,u_5,u_6\}\},$$

$$\langle\underline{\mathbf{P}_2}(X)\rangle=\{\{u_1,u_2\},\{u_1,u_2\},\{u_2,u_3\},\{u_5,u_6\}\},$$

$$\langle\overline{\mathbf{P}_2}(X)\rangle=\{\{u_1\},\{u_2\},\{u_3\},\{u_4,u_5\},\{u_4,u_5\},\{u_5,u_6\}\}.$$

The target concept X is described by using the granulation order $\mathbf{P}=\{P,Q\}$.

In practices, a granulation order on an attribute set can be appointed by users or experts or built according to the significance of each attribute. In particular, in an incomplete decision table, some certain/uncertain decision rules can be extracted through constructing the positive approximation of a target decision.

Definition 2. Let $S=(U,C\cup D)$ be an incomplete decision table, $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1\succcurlyeq P_2\succcurlyeq\cdots\succcurlyeq P_n$, and $U/D=\{X_1,X_2,\ldots,X_r\}$ a decision (partition) on U. A lower approximation and a upper approximation of D related to $\mathbf{P}$ are defined by

$$\begin{cases}\underline{\mathbf{P}}D=\{\underline{\mathbf{P}}(X_1),\underline{\mathbf{P}}(X_2),\ldots,\underline{\mathbf{P}}(X_r)\},\\ \overline{\mathbf{P}}D=\{\overline{\mathbf{P}}(X_1),\overline{\mathbf{P}}(X_2),\ldots,\overline{\mathbf{P}}(X_r)\}.\end{cases}$$

In this paper, $\underline{\mathbf{P}}D$ is also called the positive region of D with respect to the granulation order $\mathbf{P}$, denoted by $POS_{\mathbf{P}}^{U}(D)=\bigcup_{k=1}^{r}\underline{\mathbf{P}}(X_k)$.

Theorem 1. *Let $S=(U,A)$ be an incomplete information system, X a subset of U and $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1 \succcurlyeq P_2 \succcurlyeq \cdots \succcurlyeq P_n$ $(P_i \in 2^A)$. Then, for $i=1,2,\ldots,n$, given that $\mathbf{P}_i=\{P_1,P_2,\ldots,P_i\}$, we have*

$$POS^U_{\mathbf{P}_{i+1}}(D)=POS^U_{\mathbf{P}_i}(D)\cup POS^{U_{i+1}}_{P_{i+1}}(D),$$

where $U_1=U$ and $U_{i+1}=U-POS^U_{\mathbf{P}_i}(D)$.

This theorem shows that a given target decision can be positively approximated by using granulation orders on the gradually reduced universe, which leads to the idea of the accelerator proposed in this paper for improving the computational performance of a heuristic attribute reduction algorithm.

In what follows, we investigate the form of positive approximation in the incomplete variable rough set framework. According to the definition of positive approximation in Definition 1, we define β-positive approximation in an incomplete information system as follows.

Definition 3. Let $S=(U,A)$ be an incomplete information system, $X\subseteq U$ and $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1 \succcurlyeq P_2 \succcurlyeq \cdots \succcurlyeq P_n$ $(P_i \in 2^A)$. Given that $\mathbf{P}_i=\{P_1,P_2,\ldots,P_i\}$, we define $\mathbf{P}_i^\beta$-lower approximation $\underline{\mathbf{P}_i}^\beta(X)$ and $\mathbf{P}_i^\beta$-upper approximation $\overline{\mathbf{P}_i}^\beta(X)$ of $\mathbf{P}_i^\beta$-positive approximation of X as

$$\begin{cases}\underline{\mathbf{P}_i}^\beta(X)=\bigcup\limits_{k=1}^{i}\underline{SIM(P_k)}^\beta X_k,\\ \overline{\mathbf{P}_i}^\beta(X)=\overline{SIM(P_i)}^\beta X,.\end{cases}$$

where $X_1=X$ and $X_k=X-\bigcup_{j=1}^{k-1}\underline{\mathbf{P}_j}^\beta(X_j)$ for $k=2,3,\ldots,i$, $i=1,2,\ldots,n$.

Definition 4. Let $S=(U,C\cup D)$ be an incomplete decision table, $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1 \succcurlyeq P_2 \succcurlyeq \cdots \succcurlyeq P_n$, and $U/D=\{X_1,X_2,\ldots,X_r\}$ a decision (partition) on U. A lower approximation and a upper approximation of D related to $\mathbf{P}$ are defined as

$$\begin{cases}\underline{\mathbf{P}}^\beta D=\{\underline{\mathbf{P}}^\beta(X_1),\underline{\mathbf{P}}^\beta(X_2),\ldots,\underline{\mathbf{P}}^\beta(X_r)\},\\ \overline{\mathbf{P}}^\beta D=\{\overline{\mathbf{P}}^\beta(X_1),\overline{\mathbf{P}}^\beta(X_2),\ldots,\overline{\mathbf{P}}^\beta(X_r)\}.\end{cases}$$

In this paper, $\underline{\mathbf{P}}^\beta D$ is also called the positive region of D with respect to the granulation order $\mathbf{P}$, denoted by $POS^{\beta U}_{\mathbf{P}}(D)=\bigcup_{k=1}^{r}\underline{\mathbf{P}}^\beta X_k$.

Theorem 2. *Let $S=(U,A)$ be an incomplete information system, X a subset of U and $\mathbf{P}=\{P_1,P_2,\ldots,P_n\}$ a family of attribute sets with $P_1 \succcurlyeq P_2 \succcurlyeq \cdots \succcurlyeq P_n$ $(P_i \in 2^A)$. Then, for $i=1,2,\ldots,n$, given that $\mathbf{P}_i=\{P_1,P_2,\ldots,P_i\}$, we have*

$$POS^{\beta U}_{\mathbf{P}_{i+1}}(D)=POS^{\beta U}_{\mathbf{P}_i}(D)\cup POS^{\beta U_{i+1}}_{P_{i+1}}(D),$$

where $U_1=U$ and $U_{i+1}=U-POS^{\beta U}_{\mathbf{P}_i}(D)$.

This theorem indicates that through the proposed incomplete variable rough set model, a given target decision also can be positively approximated by using granulation orders on the gradually reduced universe, which shows that idea of the accelerator proposed in this paper can be used to improve the computational performance of a heuristic attribute reduction algorithm in the context of incomplete variable rough set framework.

4. Feature selection based on the positive approximation

Feature selection based on rough set theory is about finding some attribute subsets that have the minimal number of attributes and retain some special properties. To construct a heuristic feature selection algorithm, three key issues should be taken into consideration, which are significance measures of attributes, search strategy and stopping (termination) criterion. However, the existing heuristic attribute reduction algorithms 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 in incomplete data.

4.1. Several representative significance measures of attributes

For efficient feature selection, many heuristic attribute reduction algorithms have been developed in the context of incomplete data, see [38–41]. For convenience, we only focus on two representative attribute reduction algorithms from incomplete data.

Given an incomplete decision table $S=(U,C\cup D)$, one can obtain the condition classification $U/SIM(C)=\{S_C(u_1),S_C(u_2),\ldots,S_C(u_{|U|})\}$ and the decision partition $U/D=\{X_1,X_2,\ldots,X_r\}$. In fact, we can denote the decision partition by the tolerance class of each object on the universe, that is $U/SIM(D)=\{S_D(u_1),S_D(u_2),\ldots,S_D(u_{|U|})\}$. Without loss of generality, let $X_j=\{u_{j1},u_{j2},\ldots,u_{js_j}\}$, where $|X_j|=s_j$ and $\sum_{j=1}^{r}s_j=|U|$. Then, the relationship between U/D and $U/SIM(D)$ is as follows:

$$X_j=S_D(u_{j1})=S_D(u_{j2})=\cdots=S_D(u_{js_j}),$$

$$|X_j|=|S_D(u_{j1})|=|S_D(u_{j2})|=\cdots=|S_D(u_{js_j})|.$$

Using this relationship, one can equivalently redefine the positive region of an incomplete decision table by

$$POS_C(D)=\{u \mid S_P(u)\subseteq S_D(u),u\in U\}.$$

Given the above denotations, in what follows we review two types of significance measures of attributes.

Hu and Cercone proposed a heuristic feature selection algorithm, called positive-region reduction (PR), which keeps the positive region of target decision unchanged [31]. Applying the idea of positive-region reduction, Yang and Shu gave a heuristic feature selection algorithm in incomplete decision tables (IPR), which also keeps the positive region of target decision unchanged [39]. In this algorithm, the significance measures of attributes are defined as follows.

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

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

where $\gamma_B(D)=|POS_B(D)|/|U|$.

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

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

In [23], Liang et al. defined information entropy to measure the uncertainty of an incomplete information system and applied the entropy to reduce redundant features. This reduction algorithm is denoted here by ILCE. The conditional entropy used in the study was defined as

$$E(D|C)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}(|S_C(u_i)|-|S_C(u_i)\cap S_D(u_i)|), \tag{1}$$

where $S_C(u_i)\in U/SIM(C)$ and $S_D(u_i)\in U/SIM(D)$. The corresponding significance measures are listed as follows.

Definition 7. Let $S=(U,C\cup D)$ be an incomplete decision table and $B\subseteq C$. $\forall a\in B$, the significance measure of a in B is defined as

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

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

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

In the incomplete variable rough set model, according to the ideas of Definitions 5 and 6, we can design an algorithm using the corresponding significance measures for searching an attribute reduct, which keeps the β-positive region of target decision unchanged. In this algorithm, the significance measures of attributes are defined as follows.

Definition 9. Let $S=(U,C\cup D)$ be an incomplete decision table and $B\subseteq C$. $\forall a\in B$, the significance measure of a in B is defined as

$$Sig_3^{inner}(a,B,D)=\gamma_B^{\beta}(D)-\gamma_{B-\{a\}}^{\beta}(D),$$

where $\gamma_B^{\beta}(D)=|POS_B^{\beta}/(D)|/|U|$.

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

$$Sig_3^{outer}(a,B,D)=\gamma_{B\cup\{a\}}^{\beta}(D)-\gamma_B^{\beta}(D).$$

All the definitions above are used in a heuristic feature selection algorithm to select an attribute from incomplete data. For a given incomplete 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 incomplete decision table. The core may be an empty set. The two kinds of significance measures can be used to find the core attributes. The following theorem is of interest with this regard.

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

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

4.2. Rank preservation of significance measures of attributes

As mentioned above, each of significance measures of attributes provides some heuristics to guide the mechanism for forward searching a feature subset. Unlike the discernibility matrix, the time consumption of the heuristic algorithms has been largely reduced. Nevertheless, these algorithms still could be very time-consuming. To introduce an improved strategy for heuristic feature selections, we concentrate on the rank preservation of the three significance measures of attributes based on the positive approximation introduced in an incomplete decision table.

For a clearer representation, we denote the significance measure of an attribute by $Sig_{\Delta}^{outer}(a,B,D,U)$ ($\Delta\in\{1,2,3\}$), which denotes the value of the significance measure on the universe U; and we write the tolerance class induced by u with respect to B on the universe U as $S_B^U(u)$.

Firstly, we investigate the rank preservation of the significance measure of attributes based on the dependency measure in incomplete decision tables. To do it in a much clearer way, we introduce the following two lemmas.

Lemma 1. *Let A,B,C,A',B',C' be six finite sets, where $A'=A\cup C$ and $B'=B\cup C'$. If $A'\subseteq B'$ and $C'\cap(A\cup B)=\emptyset$, then $A\subseteq B$.*

Lemma 2. *Let $S=(U,C\cup D)$ be an incomplete decision table, $B\subseteq C$ and $U'=U-POS_B^U(D)$. If $S_{B\cup\{a\}}^U(u')\subseteq S_D^U(u')$ and $u'\in U'$, then $S_{B\cup\{a\}}^{U'}(u')\subseteq S_D^{U'}(u')$.*

By using these two lemmas, one can prove the following theorem of rank preservation with respect to the significance measure of attributes based on the dependency measure in incomplete decision tables.

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

Secondly, we study the rank preservation of the significance measure of attributes based on Liang's conditional entropy. To do it, we need the following lemma.

Lemma 3. *Let $S=(U,C\cup D)$ be an incomplete decision table, $B\subseteq C$ and $U'=U-POS_B^U(D)$. Then,*

$$|S_B^U(u')|-|S_B^U(u')\cap S_D^U(u')|=|S_B^{U'}(u')|-|S_B^{U'}(u')\cap S_D^{U'}(u')|,u'\in U'.$$

From Lemma 3, one can get the rank preservation of the significance measure of attributes based on Liang's condition entropy in incomplete decision tables, which is as follows.

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

For feature selection in the incomplete variable rough set framework, if $\beta=0$, we also can obtain the following rank preservation result.

Theorem 6. *Let $S=(U,C\cup D)$ be an incomplete decision table, $B\subseteq C$, $U'=U-POS_B^U(D)$, and $\beta=0$. $\forall a,b\in C-B$, if $Sig_3^{outer}(a,B,D,U)\geq Sig_3^{outer}(b,B,D,U)$, then $Sig_3^{outer}(a,B,D,U')\geq Sig_3^{outer}(b,B,D,U')$.*

From Theorems 4,5,6, it is easy to observe that the sequence of attributes selected in the process of feature selection will be kept unchanged when reducing the lower approximation of positive approximation in an incomplete decision table. This property can be used to improve computational efficiency of a heuristic feature selection algorithm, while retaining the same selected feature subset from a given incomplete data set as its original version. If we use the support vector machine (SVM) or the decision tree method to construct a classifier, then the same feature subset selected must possess the same classification accuracy. From the viewpoint of classifiers, these attribute reduction algorithms may lead to the overfitting problem as a decision tree does when the tree has too long paths, which will weaken the generalization ability of classifiers induced by the attribute reducts obtained. Hence, it is very desirable to solve the overfitting problem of feature selection for learning a classifier in the framework of rough set theory. This issue will be addressed in future work. As pointed out in the introduction part, this study does not aim to improve the classification accuracy of a classifier induced by an attribute reduct, but only focuses on largely reducing computational time of original attribute reduction algorithms. In fact, a heuristic algorithm with the proposed accelerator will have the same classification accuracy as before.

Note that in the context of incomplete variable rough set model, when $0<\beta\leq 0.5$, the sequence of attributes selected in the process of feature selection may be changed in an incomplete decision table. It can be understood by the definition of incomplete variable rough set, in which the used inclusion degree

function is not monotonic. However, it is not disappointing from the following two reasons:

(1) When the stop criterion of an accelerated feature selection algorithm is satisfied, if the reduced universe is not an empty set, then the attribute reduct obtained by the algorithm must include the same features as those obtained by the original algorithm and have the same approximating ability as that obtained by the original algorithm. This is because the stop criterion requires $\gamma_{red}^{\beta U_i}(D)=\gamma_C^{\beta U_i}(D)$ and $\gamma_{red}^{\beta U}(D)=\gamma_C^{\beta U}(D)$ (see Algorithm 2).
(2) If the reduced universe is equal to an empty set, then we must have extracted a feature subset with the dependency degree $\gamma_B^{\beta}(D)=(|POS_B^{\beta}(D)|/|U|)=1$. In this situation, all objects in the universe have been put in the lower approximation of the target decision, hence the obtained feature subset has a much better approximation ability than the original feature subset. This does provide a satisfying and interesting feature subset for a much better approximation.

4.3. Feature selection algorithms based on the positive approximation

The objective of rough set-based feature selection is to find a subset of attributes which retains some special properties as the original data 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. In most applications, it is enough to find a single reduct. Based on the significance measures of attributes, some heuristic algorithms have been proposed in the literature, most of which are greedy forward search algorithms. These search algorithms start with a non-empty set, and keep adding one or several attributes of high significance into a pool each time until the dependence no longer increases.

In a feature selection algorithm based on rough set theory, we need to compute tolerance classes induced by the condition attributes in an incomplete decision table. This process largely affects computational time of an algorithm for feature selection. In order to design an efficient feature selection algorithm, we first give a fast algorithm for acquiring tolerance classes (QAATC) from a given incomplete decision table, which is mainly based on the idea of radix sorting algorithm. Its time complexity is $O(|A||U|+\sum_{j=1}^{|A|}\sum_{k=1}^{j-1}|*_{a_k}||V_{a_k}|)$, where $|*_{a_k}|$ is the number of objects with missing value $*$ under attribute a_k, and $|V_{a_k}|$ represents the number of values (is not equal to $*$) under attribute a_k, respectively. It is well known that rough set theory is mainly used for knowledge discovery from symbolic data, in which the number of values under each attribute is so small that it often can be seen as a constant. Hence, the time complexity of the algorithm is almost not affected by $|V_{a_k}|$. In addition, the number of objects with "$*$" under each attribute often is also much smaller, and its maximum value is $|U|$ in one worst attribute that can not provide any classification information. Therefore, we can further reduce the time complexity to

$$O\left(|A||U|+\sum_{j=1}^{|A|}\sum_{k=1}^{j-1}|*_{a_k}||V_{a_k}|\right)\approx O(|A||U|+|A|^2|U|)=O(|A|^2|U|).$$

Hence, this algorithm will show its advantage for calculating tolerance classes from large-scale incomplete data in which the dimensionality has much smaller effect than the size of objects on computational time. Taking into account the compactness of the article, we omit the description of the algorithm here.

From the discussion in the previous subsection, we obtain an improved forward search algorithm based on the positive approximation and the fast algorithm for acquiring tolerance classes, which is formulated as follows. In this general algorithmic framework, we denote the evaluation function (stop criterion) by $EF^U(B,D)=EF^U(C,D)$. For example, if one adopts Liang's conditional entropy, then the evaluation function is $E^U(B,D)=E^U(C,D)$. That is to say, if $EF^U(B,D)=EF^U(C,D)$, then B is said to be an attribute reduct.

Algorithm 1. A general accelerated incomplete feature selection algorithm based on the positive approximation (IFSPA)

Input: An incomplete 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 $Sig^{inner}(a_k,C,D,U)$, $k\le|C|$;
*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: $i\leftarrow1$, $R_1=red$, $\mathbf{P}_1=\{R_1\}$ and $U_1\leftarrow U$;
*Step*5: While $U_i\neq\emptyset$ and $EF^{U_i}(red,D)\neq EF^{U_i}(C,D)$ Do
{Compute the positive region of positive approximation $POS_{\mathbf{P}_i}^U(D)$,
$U_i=U-POS_{\mathbf{P}_i}^U(D)$,
$i\leftarrow i+1$,
$red\leftarrow red\cup\{a_0\}$, where $Sig^{outer}(a_0,red,D,U_i)=\max\{Sig^{outer}(a_k,red,D,U_i),\ a_k\in C-red\}$,
$R_i\leftarrow R_i\cup\{a_0\}$,
$\mathbf{P}_i\leftarrow\{R_1,R_2,\ldots,R_i\}$ };
*Step*6: return *red* and end.

For feature selection from incomplete data in the incomplete variable rough set framework, we can also modify a feature selection algorithm using the β-positive approximation as follows.

Algorithm 2. An accelerated incomplete feature selection algorithm based on the β-positive approximation (IFSPA-IVPR)

Input: An incomplete decision table $S=(U,C\cup D)$ and the threshold $\beta\le0.5$;
Output: One reduct *red*.
*Step*1: $red\leftarrow\emptyset$;//*red* is the pool to conserve the selected attributes
*Step*2: Compute $Sig_3^{inner}(a_k,C,D,U)$, $k\le|C|$;
*Step*3: Put a_k into *red*, where $Sig_3^{inner}(a_k,C,D,U)>0$;
*Step*4: $i\leftarrow1$, $R_1=red$, $\mathbf{P}_1=\{R_1\}$ and $U_1\leftarrow U$;
*Step*5: While $U_i\neq\emptyset$ and $\gamma_{red}^{\beta U_i}(D)\neq\gamma_C^{\beta U_i}(D)$ Do
{Compute the positive region of positive approximation $POS_{\mathbf{P}_i}^{\beta U}(D)$,
$U_i=U-POS_{\mathbf{P}_i}^{\beta U}(D)$,
$i\leftarrow i+1$,
$red\leftarrow red\cup\{a_0\}$, where $Sig_3^{outer}(a_0,red,D,U_i)=\max\{Sig_3^{outer}(a_k,red,D,U_i),\ a_k\in C-red\}$,
$R_i\leftarrow R_i\cup\{a_0\}$,
$\mathbf{P}_i\leftarrow\{R_1,R_2,\ldots,R_i\}$ };
*Step*6: return *red* and end.

To determine the time complexity for Algorithms 1 and 2 in the following, we use the same framework. Computing the

Table 1
The complexity description.

Algorithms	Step 2	Step 3	Step 5	Other steps
Original one	$O(\|C\|^2\|U\|^2)$	$O(\|C\|)$	$O\left(\sum_{i=1}^{\|C\|}(\|C\|-i+1)^2\|U\|^2\right)$	Constant
IFSPA	$O\left(\|C\|^2\|U\|+\|C\|\sum_{j=1}^{\|C\|}\sum_{k=1}^{j-1}\|*_{a_k}\|\|V_{a_k}\|\right)$	$O(\|C\|)$	$O\left(\sum_{i=1}^{\|C\|}((\|C\|-i+1)^2\|U_i\|+(\|C\|-i+1)\sum_{j=1}^{\|C\|-i+1}\sum_{k=1}^{j-1}\|*_{a_k}^{U_i}\|\|V_{a_k}^{U_i}\|)\right)$	Constant

significance measure of an attribute $Sig^{inner}(a_k,C,D,U)$ is one of key steps in IFSPA. The fast algorithm for computing tolerance classes has the time complexity $O(|C||U|+\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_{a_k}||V_{a_k}|)$, $a_k \in C$. Hence, the time complexity of computing the core in Step 2 is $O(|C|^2|U|+|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_{a_k}||V_{a_k}|)$. In Step 5, we begin with the core and add an attribute with the maximal significance into the set in each stage until a reduct is obtained, and this process is called a forward reduction algorithm. To estimate the time complexity of Step 5, we denote the number of objects with missing value $*$ under attribute a_k and the number of values (is not equal to $*$) under attribute a_k on the universe U by $|*_{a_k}^U|$ and $|V_{a_k}^U|$, respectively. Hence, one can obtain that the time complexity of Step 5 is given by

$$O\left(\sum_{i=1}^{|C|}((|C|-i+1)^2|U_i|+(|C|-i+1)\sum_{j=1}^{|C|-i+1}\sum_{k=1}^{j-1}|*_{a_k}^{U_i}||V_{a_k}^{U_i}|)\right).$$

These results together show that the time complexity of IFSPA is as follows:

$$O\left(|C|^2|U|+|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_{a_k}||V_{a_k}|+\sum_{i=1}^{|C|}((|C|-i+1)^2|U_i|+(|C|-i+1)\sum_{j=1}^{|C|-i+1}\sum_{k=1}^{j-1}|*_{a_k}^{U_i}||V_{a_k}^{U_i}|)\right).$$

According to the analysis on time complexity of Algorithm QAATC, the time complexity of IFSPA can be approximately estimated as $O(|C|^3|U|+\sum_{i=1}^{|C|}((|C|-i+1)^2|U_i|+(|C|-i+1)^3|U_i|))$. However, the time complexity of a classical heuristic algorithm is $O(|C|^2|U|^2+\sum_{i=1}^{|C|}(|C|-i+1)^2|U|^2)$. Obviously, the time complexity of IFSPA is lower than that of each of classical heuristic attribute reduction algorithms for incomplete data. The proposed accelerator may be a more reasonable choice for improving the time complexity of feature selection from large-scale incomplete data in which the dimensionality has much smaller effect than the size of objects on computational time. Hence, one can draw the conclusion that the general incomplete feature selection algorithm based on the positive approximation (IFSPA) may significantly reduce the computational time for feature selection from decision tables. To stress these findings, the time complexity of each step in original algorithms and IFSPA is given in Table 1, respectively.

4.4. Computational efficiency of algorithms

Many heuristic feature selection algorithms have been developed for incomplete data. The three heuristic algorithms mentioned in Subsection 4.1 are quite representative. The objective of the following experiments is to show computational efficiency of the proposed general framework for selecting a feature subset from incomplete data. In what follows, we perform experimental analysis on the classical incomplete rough set model and the incomplete variable rough set model, respectively. Data used in the experiments are outlined in Table 2, which were all downloaded from UCI Repository of machine learning databases.

Table 2
Data sets description.

	Data sets	Samples	Features	Classes
1	Audiology.standardized	200	69	24
2	Soybean-large	307	35	19
3	Dermatology	366	34	6
4	Breast-cancer-wisconsin	699	9	2

4.4.1. Computational efficiency of algorithms based on the incomplete rough set framework

In order to compare the two representative feature selection algorithms (IPR and ILCE) with the modified ones, we employ in this subsection four UCI data sets from Table 2 to verify the performance of the modified algorithms in time reduction, which are all symbolic data with missing values.

With regard to any heuristic feature selection algorithm for incomplete data in rough set theory, the computation of tolerance classes is the first key step. For convenience of comparison, we will use algorithm QAATC in this paper.

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 the six data sets into 20 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, and so on. These data sets are used to calculate the computational time used by each of the original feature selection algorithms and the corresponding modifications and to show it vis-a-vis the size of a universe. These algorithms are run on a personal computer with Windows XP, Pentium(R) D 3.4 GHz processor and 1.00 GB memory.

In the sequence of experiments, we compare IPR (and ILCE) with IFSPA-IPR (and IFSPA-ILCE) on the four real-world data sets in Table 2. We show the experimental results in Figs. 1–6. In each of these 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 gives the computational time.

It is easy to see from Figs. 1 and 2 that the computational time of each of these two algorithms usually 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 (b) in Fig. 1, the computational time decreases. One can also observe the same effect in sub-figures (a) and (c) of Fig. 1 and sub-figure (b) of Fig. 2. Therefore, one could envision that this situation must have occurred because different numbers of features are selected.

From Figs. 1 and 2, we find that the modified algorithms are much faster than their original counterparts. Furthermore, the differences become larger and 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 algorithms is the same as the one produced by the original algorithm.

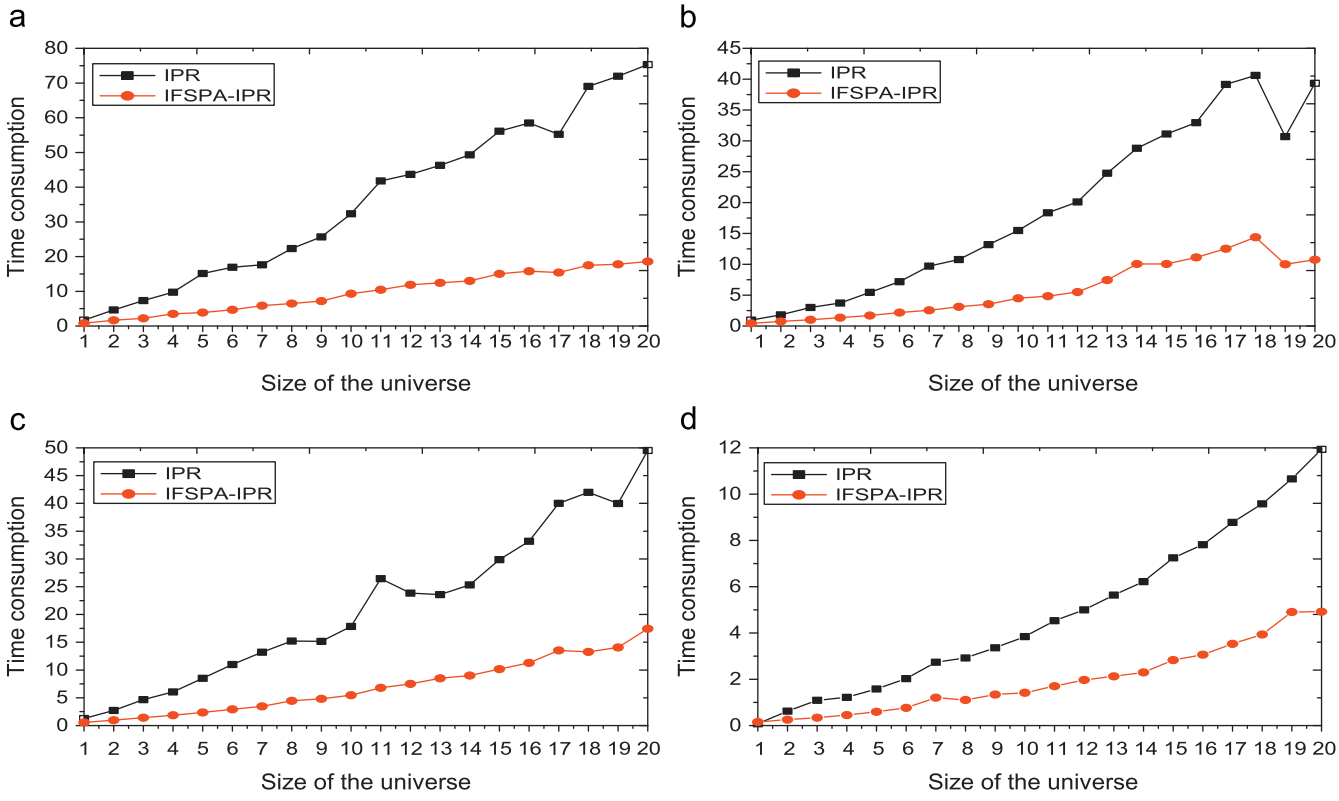

Fig. 1. Times of IPR and IFSPA-IPR versus size of data. (a) Audiology.standardized, (b) soybean-large, (c) dermatology and (d) breast-cancer-wisconsin.

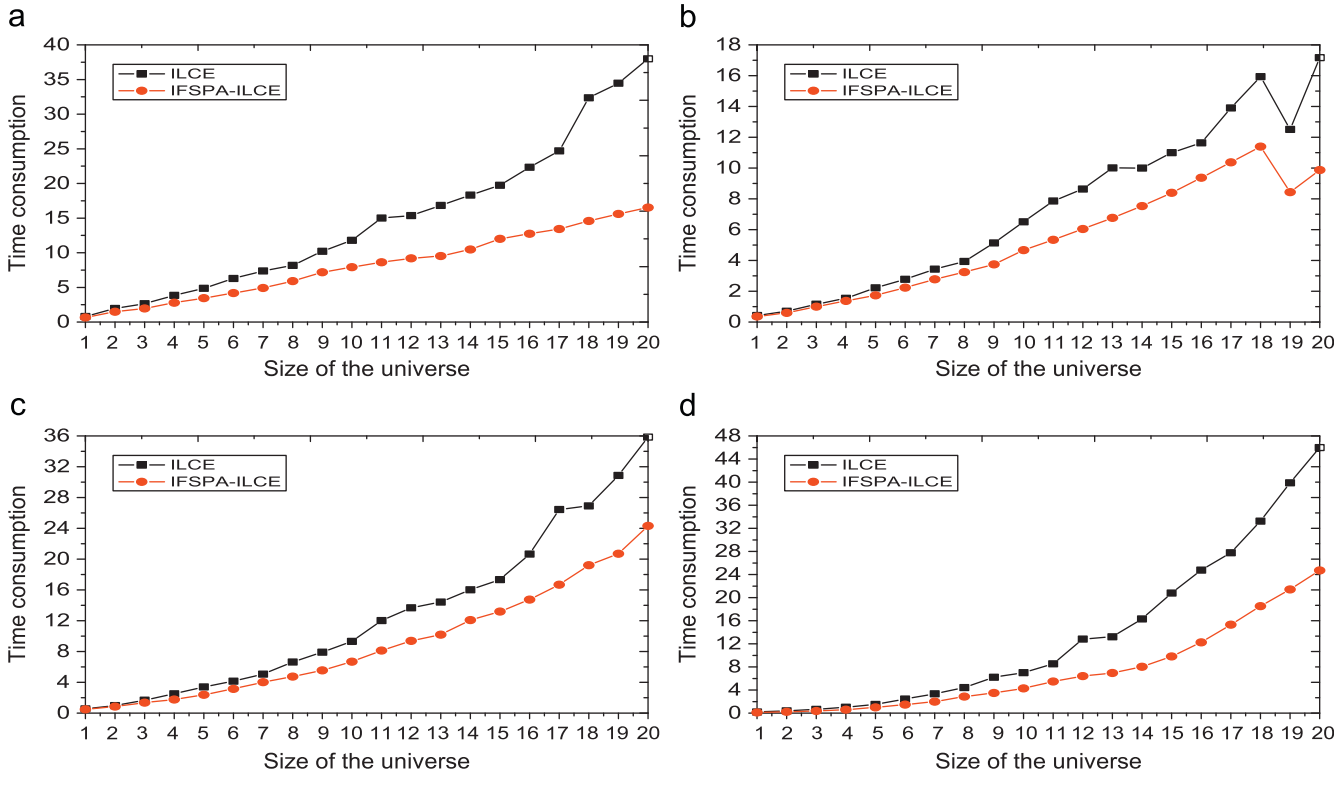

Fig. 2. Times of ILCE and IFSPA-ILCE versus size of data. (a) Audiology.standardized, (b) soybean-large, (c) dermatology and (d) breast-cancer-wisconsin.

4.4.2. Computational efficiency of algorithms based on the incomplete variable rough set framework

In this subsection, we compare the feature selection algorithms without the accelerator (IVPR) with the modified one (IFSPA-IVPR) in the incomplete variable rough set framework. We will still employ those four UCI data sets from Table 2 to verify the performance of the modified algorithm in time reduction.

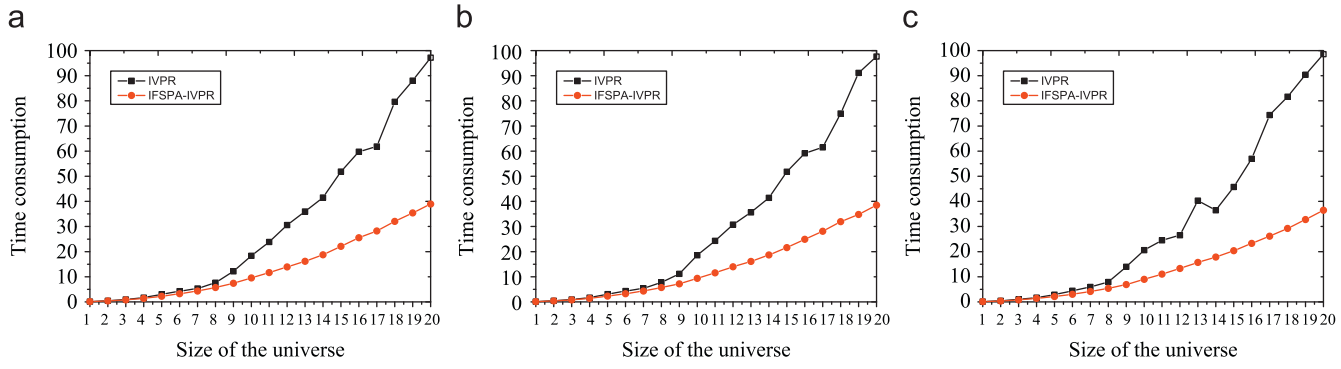

Fig. 3. Times of IVPR and IFSPA-IVPR versus size of data (Audiology.standardized). (a) $\beta=0$, (b) $\beta=0.1$ and (c) $\beta=0.2$.

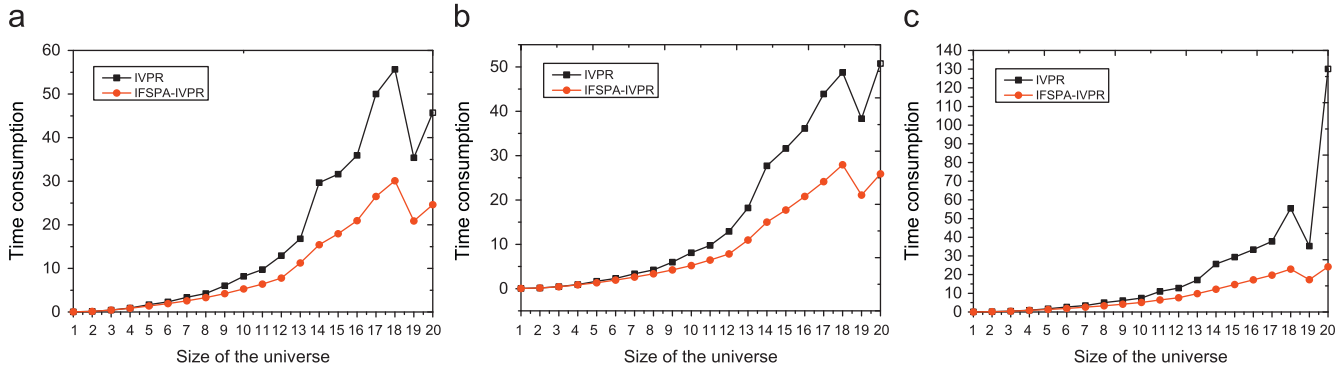

Fig. 4. Times of IVPR and IFSPA-IVPR versus size of data (Soybean-large). (a) $\beta=0$, (b) $\beta=0.1$ and (c) $\beta=0.2$.

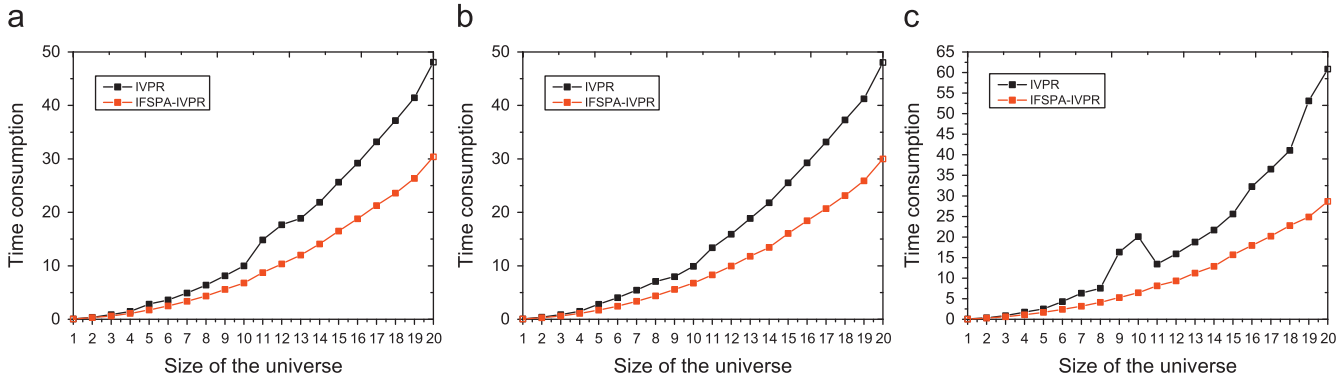

Fig. 5. Times of IVPR and IFSPA-IVPR versus size of data (Dermatology). (a) $\beta=0$, (b) $\beta=0.1$ and (c) $\beta=0.2$.

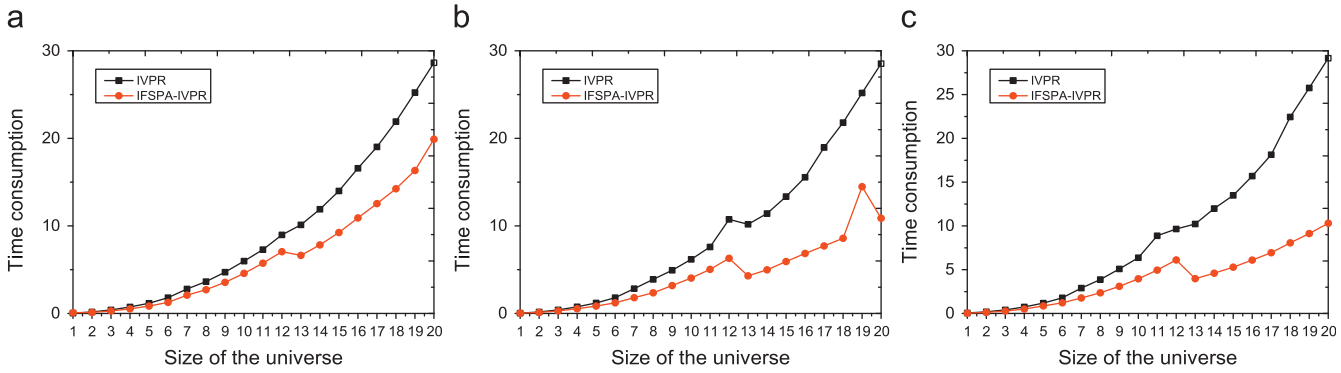

Fig. 6. Times of IVPR and IFSPA-IVPR versus size of data (Breast-cancer-wisconsin). (a) $\beta=0$, (b) $\beta=0.1$ and (c) $\beta=0.2$.

In the experimental analysis, to demonstrate the performance improvement of the modified algorithm, we set the value of β to vary from 0 to 0.2, that is $\beta=0$, 0.1, and 0.2, respectively. We report the experimental results in Figs. 3–6.

It can be seen from Figs. 3–6 that the computational time of each of the two algorithms usually increases with the increase of the size of data. Nevertheless this relationship is also not strictly monotonic. For example, one can see this phenomenon from the sub-figures (a), (b) and (c) in Fig. 4. We also observe the same effect from Figs. 3, 5 and 6. Therefore, one could envision that this situation must have occurred because of different numbers of features selected.

As shown in Figs. 3–6, the modified algorithms are consistently much faster than their original counterparts. Furthermore, the differences become larger and larger when the size of the data set increases. Although we cannot ensure the rank of attributes in the process of feature selection remains unchanged, the feature subset obtained by the modified algorithm has the following properties: either the obtained feature subset has the same approximation ability as that of the original one, which can be ensured by the stop criterion in the modified algorithm; or the obtained feature subset has a much better approximation ability than the original one, in which all objects in the universe have been put in the lower approximation of the target decision.

From the four figures, one can find the phenomenon that the proposed algorithm is able to maintain a steady increase in processing time, whereas the original algorithm incurs an unexpected high surge in processing time. This phenomenon may result from three possible reasons: (1) the accelerated algorithm does take much smaller processing time when the universe is gradually reduced; (2) since the original algorithm selects more attributes than the modified algorithm on the same data set, their processing time has a remarkable difference; and (3) the accelerated algorithm has put all objects in the lower approximation of the target decision in the process of attribute reduction, so it can save more processing time than the original one.

4.5. Stability analysis of algorithms

The stability of a heuristic feature selection algorithm determines the stability of its classification accuracy. The objective of experiments in this subsection is to compare the stabilities of computational time and feature selection of each of the modified algorithms with those obtained when running the original algorithms. In the experiments, we still use the four real-world data sets in Table 2.

In order to evaluate the stability of the feature subset selected with 10-fold cross validation, we introduce several definitions and necessary notations as follows. Let $X_1,X_2,\ldots,X_{10}$ be the 10 data sets coming from a given universe U. We use C_0 to denote the reduct induced by the universe U. The reducts induced by the data set X_i will be denoted by C_i ($1\leq i\leq 10$), respectively. To measure the difference between two reducts C_i and C_j, we use the distance

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

Next we calculate the mean value of the 10 distances:

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

The standard deviation

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

is used to characterize the stability of the reduct result induced by a heuristic feature selection algorithm for incomplete data. 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 computational time.

As before, we use the same heuristic feature selection algorithms along with their modifications. The results reported in Tables 3–5 are obtained from applications of the 10-fold cross validation.

Table 3 reveals that IFSPA-IPR has a far lower mean computational time and standard deviation than the ones produced by the original IPR. The IFSPA-IPR's stability is the same as that reported for the IPR. In other words, as an accelerator for feature selection, the positive approximation can be used to significantly reduce the computational time of the algorithm IPR. The much smaller standard deviation implies that the modified algorithm IFSPA-IPR exhibits a far better robustness than the original IPR. We also note that the modified algorithm has no effect on the stability of reducts induced by the original algorithm (we obtained the same attribute reduct on the same incomplete data set). The mechanism can be well interpreted by the rank preservation of the significance measures of attributes used in the algorithms IPR and IFSPA-IPR (see Theorem 4). From Table 4, one can draw the same conclusions.

From Table 5, it can be noted that IFSPA-IVPR has a far lower mean computational time and standard deviation than the ones produced by the original IVPR. In other words, as an accelerator for feature selection, the positive approximation can be used to significantly reduce the computational time of the algorithm IVPR. The much smaller standard deviation implies that the modified algorithm IFSPA-IVPR also exhibits a far better robustness than the original IVPR. For the stability of attribute reducts, it can be seen that when the threshold $\beta=0$, the IFSPA-IVPR's stability is the same as that reported for the IVPR. This conclusion can be ensured by Theorem 6 (rank preservation theorem). From the theorem it is easy to see that the sequence of attributes in the process of feature selection will be kept unchanged when reducing the lower approximation of positive approximation in an incomplete decision table. This property can be used to improve computational efficiency of a heuristic feature selection algorithm, while retaining the same selected feature subset from a given incomplete data set. For other cases, the property may not hold in the incomplete variable rough set framework, which is because one may have extracted a much better feature subset with the dependency degree of one in the process of feature selection.

From the definition of positive approximation and the corresponding accelerated algorithm, it follows that the stopping criterion of each of accelerated feature selection algorithms ensures that the feature subset selected has the same classification ability as the original feature set, while each modified algorithm has much smaller computational time. Hence, we think that the idea of the accelerator (positive approximation) is promising to be applied to other types of forward search feature selection approaches based on measures like information gain, distance, dependency and information entropy.

5. Conclusions

To overcome the limitations of the existing feature selection schemes, a theoretic framework based on tolerance relations have been proposed in this study, which is called the positive approximation and can be used to accelerate heuristic algorithms for feature selection from incomplete data. Based on this framework, a general heuristic incomplete feature selection algorithm (IFSPA) has been presented. For feature selection from incomplete data, several representative heuristic algorithms in rough set theory have been modified. Each of the modified algorithms can choose the same feature subset as the original one. Experimental studies pertaining to four UCI data sets show that the modified algorithms can largely reduce

Table 3
The stabilities of the time and feature selection of the algorithms IPR and IFSPA-IPR.

Data sets	IPR's time	IFSPA-IPR's time	IPR's stability	IFSPA-IPR's stability
Audiology.standardized	72.2859 ± 5.2078	17.2050 ± 1.1261	0.2650 ± 0.1393	0.2650 ± 0.1393
Soybean-large	34.2984 ± 3.9632	10.8922 ± 1.4842	0.2312 ± 0.2071	0.2312 ± 0.2071
Dermatology	43.0281 ± 1.9503	15.6281 ± 0.5176	0.2921 ± 0.2293	0.2921 ± 0.2293
Breast-cancer-wisconsin	12.5109 ± 1.7053	4.4656 ± 0.8354	0.0700 ± 0.1552	0.0700 ± 0.1552

Table 4
The stabilities of the time and feature selection of the algorithms ILCE and IFSPA-ILCE.

Data sets	ILCE's time	IFSPA-ILCE's time	ILCE's stability	IFSPA-ILCE's stability
Audiology.standardized	37.9297 ± 2.7544	21.0484 ± 1.2110	0.1987 ± 0.1071	0.1987 ± 0.1071
Soybean-large	18.9297 ± 1.9388	12.2275 ± 0.7403	0.1773 ± 0.1364	0.1773 ± 0.1364
Dermatology	34.4672 ± 0.6046	24.4734 ± 0.6038	0.2564 ± 0.1803	0.2564 ± 0.1803
Breast-cancer-wisconsin	39.0875 ± 3.1274	25.9265 ± 2.4856	0.0733 ± 0.1171	0.0733 ± 0.1171

Table 5
The stabilities of the time and feature selection of the algorithms IVPR and IFSPA-IVPR.

Data sets	β	IVPR's time	IFSPA-IVPR's time	IVPR's stability	IFSPA-IVPR's stability
Audiology.standardized	0.0	77.7891 ± 3.1100	31.9922 ± 1.7034	0.2650 ± 0.1393	0.2650 ± 0.1393
	0.1	77.3484 ± 3.6537	31.2969 ± 1.5882	0.2430 ± 0.1407	0.1837 ± 0.0923
	0.2	77.2594 ± 3.7293	30.1219 ± 1.1077	0.2402 ± 0.1759	0.1963 ± 0.1526
Soybean-large	0.0	42.3766 ± 5.2753	21.8516 ± 3.0888	0.3630 ± 0.1940	0.3630 ± 0.1940
	0.1	48.9141 ± 19.4519	21.6031 ± 2.4870	0.3869 ± 0.2251	0.3692 ± 0.2079
	0.2	94.9578 ± 24.4370	20.1313 ± 2.1082	0.1247 ± 0.2422	0.4840 ± 0.1371
Dermatology	0.0	39.6219 ± 0.3885	24.2484 ± 0.4813	0.2278 ± 0.1991	0.2278 ± 0.1991
	0.1	40.5563 ± 0.9885	23.5922 ± 0.4280	0.3433 ± 0.1220	0.2528 ± 0.1840
	0.2	47.6219 ± 6.1172	23.1859 ± 0.3763	0.3754 ± 0.2024	0.5052 ± 0.2870
Breast-cancer-wisconsin	0.0	23.3094 ± 3.2563	15.6906 ± 3.2757	0.0700 ± 0.1552	0.0700 ± 0.1552
	0.1	24.0297 ± 4.2501	10.3016 ± 3.6133	0.0733 ± 0.1172	0.1200 ± 0.1833
	0.2	24.6734 ± 3.7963	9.8859 ± 3.7257	0.2343 ± 0.2185	0.1433 ± 0.2914

computational time of feature selection while producing the same results or much better ones as those original algorithms. The results show that the positive approximation is an efficient accelerator and can effectively select a feature subset from an incomplete data set.

Acknowledgments

We are very grateful to the anonymous reviewers for their valuable comments and suggestions.

This work was supported by National Natural Science Fund of China (Nos. 71031006, 60903110, 60773133, 70971080), National Key Basic Research and Development Program of China (973) (No. 2007CB311002), Foundation of Doctoral Program Research of Ministry of Education of China (No. 20101401110002), Natural Science Fund of Shanxi Province, China (No. 2009021017-1, No. 2008011038), and CityU 113308 of RGC of Hong Kong SAR Government.

Appendix A. Related proof

Lemma 1. *Let A,B,C,A',B',C' be six finite sets, where $A'=A\cup C$, $B'=B\cup C'$. If $A'\subseteq B'$ and $C'\cap(A\cup B)=\emptyset$, then $A\subseteq B$.*

Proof. Let $a\in A$, then $a\in A'$. From $A'\subseteq B'$, it follows that $a\in B'$. Since $C'\cap(A\cup B)=\emptyset$, so $C'\cap A=\emptyset$, then $a\notin C'$. From $a\in B'$, $a\notin C'$ and $B'=B\cup C'$, it follows that $a\in B$. That is to say, $A\subseteq B$. □

Lemma 2. *Let $S=(U,C\cup D)$ be an incomplete decision table, $B\subseteq C$ and $U'=U-POS_B^U(D)$. If $S_{B\cup\{a\}}^U(u')\subseteq S_D^U(u')$, $u'\in U'$, then $S_{B\cup\{a\}}^{U'}(u')\subseteq S_D^{U'}(u')$.*

Proof. For $u'\in U'$ and $U'=U-POS_B^U(D)$, we denote by $X=\{u \mid u\in S_{B\cup\{a\}}^U(u'),u\in POS_B^U(D)\}$ and $Y=\{v \mid v\in S_D^U(u'),v\in POS_B^U(D)\}$. Then, we have that $S_{B\cup\{a\}}^U(u')=S_{B\cup\{a\}}^{U'}(u')\cup X$ and $S_D^U(u')=S_D^{U'}(u')\cup Y$. From the formula of Y, it follows that $Y\subseteq POS_B^U(D)$.Thus, $Y\cap S_{B\cup\{a\}}^{U'}(u')=\emptyset$ and $Y\cap S_D^{U'}(u')=\emptyset$, that is $Y\cap(S_{B\cup\{a\}}^{U'}(u')\cup S_D^{U'}(u'))=\emptyset$. And, from $S_{B\cup\{a\}}^U(u')\subseteq S_D^U(u')$ and Lemma 1, it easily follows that $S_{B\cup\{a\}}^{U'}(u')\subseteq S_D^{U'}(u')$. □

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

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 $\gamma_B(D)=|POS_B(D)|/|U|$. Since $U'=U-POS_B^U(D)$, one can know $POS_B^{U'}(D)=\emptyset$. From Lemma 2, one can know that $S_{B\cup\{a\}}^{U'}(u')\subseteq S_D^{U'}(u')$ if $S_{B\cup\{a\}}^U(u')\subseteq S_D^U(u')$, $u'\in U'$. Hence, it has that $POS_{B\cup\{a\}}^{U'}(D)=POS_{B\cup\{a\}}^U(D)-POS_B^U(D)$. Therefore,

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

$$=\frac{|U'|}{|U|}\frac{|POS_{B\cup\{a\}}^U(D)|-|POS_B^U(D)|}{|POS_{B\cup\{a\}}^U(D)|-|POS_B^U(D)|}=\frac{|U'|}{|U|}.$$

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

Lemma 3. *Let $S=(U,C\cup D)$ be an incomplete decision table, $B\subseteq C$ and $U'=U-POS_B^U(D)$, then*

$$|S_B^U(u')|-|S_B^U(u')\cap S_D^U(u')|=|S_B^{U'}(u')|-|S_B^{U'}(u')\cap S_D^{U'}(u')|,u'\in U'.$$

Proof. For $u' \in U'$ and $U' = U - POS_B^U(D)$, we denote by $X = \{u \mid u \in S_B^U(u'), u \in POS_B^U(D)\}$ and $Y = \{v \mid v \in S_D^U(u'), v \in POS_B^U(D)\}$. Then, we have that $S_B^U(u') = S_B^{U'}(u') \cup X$ and $S_D^U(u') = S_D^{U'}(u') \cup Y$. From the formulas of X and Y, it follows that $X \subseteq POS_B^U(D)$ and $Y \subseteq POS_B^U(D)$, respectively. Thus, $Y \cap S_B^{U'}(u') = \emptyset$ and $X \cap S_D^{U'}(u') = \emptyset$. Since $\forall u \in X$, one can know that $u \in S_B^U(u')$. According to the symmetry of a tolerance relation, it easily follows that $u' \in S_B^U(u)$. In addition, from the definition of positive region, we know that $S_B^U(u) \subseteq S_D^U(u)$, thus $u' \in S_D^U(u)$. Similarly, one has that $u \in S_D^U(u')$. Furthermore, from $\forall u \in X$ and $\forall u \in POS_B^U(D)$, one can obtain that $u \in Y$, i.e., $X \subseteq Y$. Therefore, one has that

$$\begin{aligned} S_B^U(u') \cap S_D^U(u') &= (S_B^{U'}(u') \cup X) \cap (S_D^{U'}(u') \cup Y) \\ &= (S_B^{U'}(u') \cap S_D^{U'}(u')) \cup (S_B^{U'}(u') \cap Y) \cup (S_D^{U'}(u') \cap X) \cup (X \cap Y) \\ &= (S_B^{U'}(u') \cap S_D^{U'}(u')) \cup X. \end{aligned}$$

And since $X \subseteq POS_B^U(D)$, we have that $= (S_B^{U'}(u') \cap S_D^{U'}(u')) \cap X = \emptyset$ and $|S_B^U(u') \cap S_D^U(u')| = |S_B^{U'}(u') \cap S_D^{U'}(u')| + |X|$. Therefore, it follows that

$$\begin{aligned} |S_B^{U'}(u')| - |S_B^{U'}(u') \cap S_D^{U'}(u')| &= (|S_B^U(u')| - |X|) - (|S_B^U(u') \cap S_D^U(u')| - |X|) \\ &= |S_B^U(u')| - |X| - (|S_B^U(u') \cap S_D^U(u')| + |X|) \\ &= |S_B^U(u')| - |S_B^U(u') \cap S_D^U(u')|. \end{aligned}$$

That is, $|S_B^U(u')| - |S_B^U(u') \cap S_D^U(u')| = |S_B^{U'}(u')| - |S_B^{U'}(u') \cap S_D^{U'}(u')|$, $u' \in U'$. This completes the proof. □

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

Proof. Let $U/SIM(B) = \{S_B^U(u_1), S_B^U(u_2), \ldots, S_B^U(u_q), S_B^U(u_{q+1}), \ldots, S_B^U(u_{|U|})\}$, $U/SIM(D) = \{S_D^U(u_1), S_B^D(u_2), \ldots, S_D^U(u_q), S_D^U(u_{q+1}), \ldots, S_D^U(u_{|U|})\}$, where $u_i \in POS_B^U(D)$ $(i = 1,2,\ldots,q)$. Let us denote Liang's conditional entropy in the universe U by $E^U(D|B)$. Then it follows that

$$\begin{aligned} E^U(D|B) &= \frac{1}{|U|^2} \sum_{i=1}^{|U|} (|S_B^U(u_i)| - |S_B^U(u_i) \cap S_D^U(u_i)|) \\ &= \frac{1}{|U|^2} \sum_{i=1}^{q} (|S_B^U(u_i)| - |S_B^U(u_i) \cap S_D^U(u_i)|) + \frac{1}{|U|^2} \sum_{i=q+1}^{|U|} (|S_B^U(u_i)| \\ &\quad - |S_B^U(u_i) \cap S_D^U(u_i)|) \\ &= \frac{1}{|U|^2} \sum_{i=1}^{q} (|S_B^U(u_i)| - |S_B^U(u_i)|) + \frac{1}{|U|^2} \sum_{i=q+1}^{|U|} (|S_B^U(u_i)| \\ &\quad - |S_B^U(u_i) \cap S_D^U(u_i)|) \\ &= \frac{1}{|U|^2} \sum_{i=q+1}^{|U|} (|S_B^U(u_i)| - |S_B^U(u_i) \cap S_D^U(u_i)|). \end{aligned}$$

Furthermore, from Lemma 3, it follows that

$$\begin{aligned} &\frac{1}{|U|^2} \sum_{i=q+1}^{|U|} (|S_B^U(u_i)| - |S_B^U(u_i) \cap S_D^U(u_i)|) \\ &\quad = \frac{|U'|^2}{|U|^2} \frac{1}{|U'|^2} \sum_{i=q+1}^{|U|} (|S_B^U(u_i)| - |S_B^U(u_i) \cap S_D^U(u_i)|) \\ &\quad = \frac{|U'|^2}{|U|^2} \frac{1}{|U'|^2} \sum_{j=1}^{|U'|} (|S_B^{U'}(u_j)| - |S_B^{U'}(u_j) \cap S_D^{U'}(u_j)|) = \frac{|U'|^2}{|U|^2} E^{U'}(D|B). \end{aligned}$$

Therefore, we have that $Sig_2^{outer}(a,B,D,U)/Sig_2^{outer}(a,B,D,U') = |U'|^2/|U|^2$. Thus, if $Sig_2^{outer}(a,B,D,U) \geq Sig_2^{outer}(b,B,D,U)$, $\forall a, b \in C - B$, then $Sig_2^{outer}(a,B,D,U') \geq Sig_2^{outer}(b,B,D,U')$. This completes the proof. □

Theorem 6. *Let $S = (U, C \cup D)$ be an incomplete decision table, $B \subseteq C$, $U' = U - POS_B^U(D)$, and $\beta = 0$. For $\forall a, b \in C - B$, if $Sig_3^{outer}(a,B,D,U) \geq Sig_3^{outer}(b,B,D,U)$, then $Sig_3^{outer}(a,B,D,U') \geq Sig_3^{outer}(b,B,D,U')$.*

Proof. Similar to the proof in Theorem 4, it can be easily proved. □

References

[1] I. Guyon, A. Elisseeff, An introduction to variable feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[2] J. Yu, General c-means clustering model, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1197–1211.
[3] Z.H. Zhou, Three perspectives of data mining, Artificial Intelligence 143 (1) (2003) 139–146.
[4] N. Kwak, C.H. Choi, Input feature selection for classification problems, IEEE Transactions on Neural Networks 13 (2002) 143–159.
[5] 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.
[6] T. Pavlenko, On feature selection, curse-of-dimensionality and error probability in discriminant analysis, Journal of Statistical Planning and Inference 115 (2003) 565–584.
[7] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[8] C.K. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Information Processing Manage 42 (2006) 155–165.
[9] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[10] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of AAAI-92, 1992, pp. 129–134.
[11] M. Modrzejewski, Feature selection using rough set theory, in: Proceedings of European Conference on Machine Learning, 1993, pp. 213–226.
[12] 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.
[13] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.
[14] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognition 35 (2002) 825–834.
[15] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (2004) 1351–1363.
[16] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Letters 26 (2005) 965–975.
[17] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters 26 (2005) 1632–1640.
[18] 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.
[19] 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.
[20] 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.
[21] H. Liu, R. Setiono, Feature selection via discretization, IEEE Transactions on Knowledge Data Engineering 9 (4) (1997) 642–645.
[22] Y. Leung, D.Y. Li, Maximal consistent block technique for rule acquisition in incomplete information systems, Information Sciences 153 (2003) 85–106.
[23] J.Y. Liang, Z.Z. Shi, D.Y. Li, M.J. Wierman, The information entropy, rough entropy and knowledge granulation in incomplete information systems, International Journal of General Systems 35 (6) (2006) 641–654.
[24] M. Kryszkiewicz, Rough set approach to incomplete information systems, Information Sciences 112 (1998) 39–49.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] Y.H. Qian, J.Y. Liang, C.Y. Dang, Interval ordered information systems, Computer & Mathematics with Applications 56 (2008) 1994–2009.
[30] A. Skowron, Extracting laws from decision tables: a rough set approach, Computational Intelligence 11 (1995) 371–388.
[31] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, International Journal of Computer Intelligence 11 (2) (1995) 323–338.
[32] 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.

[33] 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.
[34] D. Slezak, Approximate entropy reducts, Foundamenta Informaticae 53 (3–4) (2002) 365–390.
[35] 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.
[36] G.Y. Wang, J. Zhao, J.J. An, A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Foundamenta Informaticae 68 (3) (2005) 289–301.
[37] 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 Science and Information 2 (3) (2004) 557–562.
[38] 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.
[39] C.S. Yang, L. Shu, Attribute reduction algorithm of incomplete decision table based on tolerance relation, Computer Technology and Development 16 (9) (2006) 68–69 72.
[40] Y.H. Qian, J.Y. Liang, F. Wang, A new method for measuring the uncertainty in incomplete information systems, Fuzziness and Knowledge-Based Systems 17 (6) (2009) 855–880.
[41] B. Huang, X.Z. Zhou, R.R. Zhang, Attribute reduction based on information quantity under incomplete information systems, Systems Engineering—Theory & Practice 4 (2005) 55–60.
[42] Y.H. Qian, J.Y. Liang, W. Wei, Accelerating incomplete feature selection, in: Proceedings of the 8th IEEE International Conference on Machine Learning and Cybernetics, Baoding, China, 2009, pp. 350–358.
[43] Y.H. Qian, J.Y. Liang, C.Y. Dang, Incomplete multigranulation rough set, IEEE Transactions on Systems, Man and Cybernetics: Part A 40 (2) (2010) 420–431.
[44] W. Ziarko, Variable precision rough set model, Journal of Computer and System Science 46 (1993) 39–59.
[45] Y.H. Qian, J.Y. Liang, C.Y. Dang, Knowledge structure, knowledge granulation and knowledge distance in a knowledge base, International Journal of Approximate Reasoning 50 (1) (2009) 174–188.

Yuhua Qian is an Assistant Professor of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China, and is also a Senior Research Associate of Department of Manufacturing Engineering and Engineering Management at the City University of Hong Kong. He received the M.S. degree in Computers with applications at Shanxi University (2005). He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing and artificial intelligence. He has published more than 20 articles in international journals. He also served on the Editorial Board of International Journal of Knowledge-Based Organizations. He is also a member of IEEE.

Jiye Liang is a professor of School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. He received the Ph.D degree in Applied Mathematics from Xi'an Jiaotong University. He also has a B.S. in computational mathematics from Xi'an Jiaotong University. His research interests include artificial intelligence, granular computing, data mining and knowledge discovery. He has published more than 30 articles in international journals.

Witold Pedrycz is a Professor and Director of Computer Engineering and Software Engineering in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. He is actively pursuing research in Computational Intelligence, fuzzy modeling, knowledge discovery and data mining, fuzzy control including fuzzy controllers, pattern recognition, knowledge-based neural networks, relational computation, and Software Engineering. He has published numerous papers in this area. He is also author of 7 research monographs covering various aspects of Computational Intelligence and Software Engineering. Dr. Pedrycz is an IEEE Fellow. He currently serves as the Chief Editor of Information Sciences, the Chief Editor of IEEE Transactions on Systems Man and Cybernetics Part A, and an Associate Editor of IEEE Transactions on Fuzzy Systems. He also served on the Editorial Board of IEEE Transactions on Neural Networks.

Chuangyin Dang received a Ph.D. degree in operations research/economics from the University of Tilburg, The Netherlands, in 1991, a M.S. degree in applied mathematics from Xidian University, China, in 1986, and a B.S. degree in computational mathematics from Shanxi University, China, in 1983. He is Associate Professor at the City University of Hong Kong. He is best known for the development of the D1-triangulation of the Euclidean space and the simplicial method for integer programming. His current research interests include computational intelligence, optimization theory and techniques, applied general equilibrium modeling and computation. He is a senior member of IEEE and a member of INFORS and MPS.