A Feature Selection Approach of Inconsistent Decision Systems in Rough Set

Lin Sun
College of Computer & Information Engineering, Henan Normal University, Xinxiang 453007, P. R. China
Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China
International WIC Institute, Beijing University of Technology, Beijing 100124, P. R. China
Email: linsunok@gmail.com

Jiucheng Xu, Yuhui Li
College of Computer & Information Engineering, Henan Normal University, Xinxiang 453007, P. R. China
Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China
Email: xjc@htu.cn

***Abstract*—Feature selection has been widely discussed as an important preprocessing step in data mining applications since it reduces a model's complexity. In this paper, limitations of several representative reduction methods are analyzed firstly, and then by distinguishing consistent objects form inconsistent objects, decision inclusion degree and its probability distribution function as a new measure are presented for both inconsistent and consistent simplified decision systems. New definitions of distribution reduct and maximum distribution reduct for simplified decision systems are proposed. Many important propositions, properties, and conclusions for reduct are drawn. By using radix sorting and hash techniques, a heuristic distribution reduct algorithm for feature selection is constructed. Finally, compared with other feature selection algorithms on six UCI datasets, the proposed approach is effective and suitable for both consistent and inconsistent decision systems.**

***Index Terms*—feature selection, rough set, decision system, decision inclusion degree, distribution reduct**

I. INTRODUCTION

Rough set theory, originated by Pawlak [1] in 1980s, is a powerful mathematical tool to deal with inexact, uncertain, and vague knowledge in information systems [2-5]. It has been widely used for feature selection because it is completely data-driven and does not require any auxiliary information [6]. The selection of relevant and significant features is an important problem particularly for data sets with large number of features [7-9]. But those irrelevant features can deplete the storage space, deteriorate the computational performance, and even decrease the generalization power of the induced patterns [2, 3, 6, 7]. It is, thus, desirable to search for a feature subset that has the same predictive capability as that of the original feature set.

In the last two decades, as an important successful application of rough set models in a variety of problems such as artificial intelligence, machine learning, data mining, and so on, feature selection or attribute reduction in information systems has been drawing wide attention [2, 3, 10]. There are many techniques for feature selection developed in rough set theory [2-12]. These types of feature selection have been proposed in the analysis of information systems, each of which aimed at some basic real-world requirements. Unfortunately, it has been proved that finding all reducts or finding an optimal reduct (a reduct with the least number of attributes) is an NP-complete problem [13]. Many researchers devote themselves to finding an efficient reduct by optimization techniques [2-20]. A distribution reduct [21] was a subset of the feature set that preserved the degree to which every object belonged to each decision class. Kryszkiewicz [22] described two methods of knowledge reduction for inconsistent decision systems, namely assignment reduction and distribution reduction. In an inconsistent decision system, assignment reduction maintains unchanged with the possible decisions for arbitrary object. In comparison, distribution is characterized by preserving the class membership distribution and is a more complete knowledge reduction for all objects in an inconsistent decision system. In other words, the distribution reduction preserves not only all of the deterministic information but also the non-deterministic information of an inconsistent decision system. Yao and Zhao [23] thought that the partition based on the membership distribution vector was more complex, which allowed the distribution reduction to preserve the quality of the decisions. However, it can be a concern that the distribution reduction has strict requirements, and the decision rules derived from distribution reduction are usually less compact and more complicated. For this reason, the maximum distribution reduction in [24] proposed by Zhang et al. remains unchanged with the maximum decision classes for all of the objects in a decision system, which is a good compromise between the compactness of derived rules and the capability of preserving information with respect to decisions. Mi et al. [25] introduced β-reduct on the basis of variable precision rough set model. This type of reduct preserved the sum of objects in the β lower approximations of all decision classes. Wu et al. [26] proposed the concepts of β lower

doi:10.4304/jcp.9.6.1333-1340

distribution reduct and β upper distribution reduct. Ye et al. [27] presented an algorithm for finding a maximum distribution reduct of an inconsistent decision system. Liu et al. [28] introduce a new type of reducts called the λ-Fuzzy-Reduct. However, some of these current algorithms for feature selection have still their own some limitations. Based on the mutual information, Miao and Hu [29] constructed a heuristic algorithm costing time complexity $O(|C||U|^2)$ + $O(|U|^3)$. Hence, the main disadvantage of these methods is much time-space cost. Xu et al. [30] designed a new and relatively reasonable formula for an efficient reduction algorithm, whose worst time complexity was cut down to Max($O(|C||U|)$, $O(|C^2|U/C|)$). Liu et al. [31] presented a hash-based algorithm to calculating positive region, and its time complexity decreased to $O(|U|)$, and a reduction algorithm with twice-hash was presented, whose time complexity was $O(|C|^2|U/C|)$. So far its efficiency is fortunate. However, because of various factors such as noise in the data, lack of critical knowledge, compact representation, and prediction capability, most of decision systems are inconsistent. Inconsistent decision system is a common information system in realistic decision analysis problems, as well as is the focus of study in information systems reduction processing [32]. Discernible matrix that was used for seeking core attribute set of inconsistent decision system in [33] is defective [34]. It may not obtain the right attribute set. Moreover, when several reductions are achieved, their advantages and disadvantages of every reduction cannot be compared in actual applications. Algorithms in [35] used for seeking core attribute set with discernible matrix for incompatible decision system are also defective. Distribution reduction of both consistent and inconsistent decision systems were defined in [33], and their equivalent forms were discussed there. But there was no further study for these two kinds of knowledge reduction methods. Qin et al. [36] proved conditional information entropy reduction in [34] and distribution reduction in [37] were equivalent, and they cannot only ensure decision-making ability of invariant consistent decision rules, but also can guarantee decision-making ability of invariant inconsistent decision rules. Although the heuristic approaches above can avoid the exponential computation in exhaustive methods, they still suffer from intensive computation of either discernibility functions or partitions of universe. Therefore, it is necessary to propose an effective heuristic feature selection algorithm in inconsistent decision systems with less time-space complexity. This paper focuses on creating such a solution.

The remainder of this paper is structured as follows. In Section II, some basic concepts are recalled. In Section III, some concepts, properties and propositions about decision inclusion degree and probability distribution function are presented for both inconsistent and consistent simplified decision systems. An effective heuristic distribution reduct algorithm for feature selection is put forward in Section IV. Section V gives the applications and experimental evaluations. Finally, the conclusions are described in Section VI.

II. Preliminaries

In this section, we review briefly some notions and results related to information systems and decision systems in rough sets. Detailed description of concepts can be found in [1, 2, 5, 10].

The notion of information system (*IS*) has been studied by many authors as a simple knowledge representation method. Formally, an information system is a quadruple $IS = (U, A, V, f)$, where U is a finite nonempty set of objects indicating a given universe; A is a finite nonempty set of features; V is the union of feature domains such that $V = \bigcup_{a\in A} V_a$ for V_a denoting the value domain of feature a; $f: U \times A \to V$ is an information function which associates a unique value of each feature with every object belonging to U, such that for any $a \in A$ and $u \in U$, $f(u, a) \in V_a$. Also, $IS = (U, A, V, f)$ can be written more simply as $IS = (U, A)$.

With every subset $P \subseteq A$, there is an associated indistinguishable relation $IND(P)$ as follows:

$$IND(P) = \{(u, v) \in U \times U \mid \forall a \in P, f(u, a) = f(v, a)\}.$$

It can be easily shown that $IND(P)$ is an equivalence relation on the set U and $IND(P) = \bigcap_{a\in P} IND(\{a\})$. For $P \subseteq A$, the equivalence relation $IND(P)$ partitions U into some equivalence classes given by $U/IND(P) = \{[u]_P \mid u \in U\}$, for simplicity, $U/IND(P)$ will be replaced by U/P, where $[u]_P$ denotes the equivalence class (block) determined by u with respect to P, i.e., $[u]_P = \{v \in U \mid (u, v) \in IND(P)\}$. Each $[u]_P$ is viewed as an information granule consisting of indistinguishable elements.

We define a partial order on all partition sets of U. Let P and Q be two equivalence relations of U, $U/P = \{P_1, P_2, \ldots, P_m\}$ and $U/Q = \{Q_1, Q_2, \ldots, Q_n\}$ be partitions of the finite set U. Then, we define that the partition U/P is finer than the partition U/Q (or the partition U/Q is coarser than the partition U/P), denoted by $P \preceq Q$ (or $Q \succeq P$), between partitions by $P \preceq Q \Leftrightarrow \forall P_i \in U/P, \exists Q_j \in U/Q \to P_i \subseteq Q_j$. If $P \preceq Q$ and $P \succeq Q$, then we say that $P = Q$. If $P \preceq Q$ and $P \neq Q$, then we say that U/Q is strictly coarser than U/P (or U/P is strictly finer than U/Q) and write $P \prec Q$ (or $Q \succ P$).

An information system $IS = (U, A)$ is also called a decision system (*DS*) if $A = C \cup D$, and $C \cap D = \emptyset$, where C is the finite set of condition features and D is the finite set of decision features. Obviously, the previous properties derived hold for $DS = (U, A = C \cup D, V, f)$. The quadruple $DS = (U, A = C \cup D, V, f)$ is usually denoted by a triple (U, C, D) for short, that is, $DS = (U, C, D)$.

Theorem 1. Let $DS = (U, C, D)$ be a decision system with $P, Q \subseteq C \cup D$. If $Q \subseteq P$, then $P \preceq Q$.

Proof. Suppose $U/P = \{P_1, P_2, \ldots, P_m\}$, $U/Q = \{Q_1, Q_2, \ldots, Q_n\}$, for any $P_i = [x]_P \in U/P$, since $Q \subseteq P$, then one has that $P_i = [x]_P = \{y \mid f(x, a) = f(y, a), \forall a \in P\} \subseteq Q_j = [x]_Q = \{y \mid f(x, a) = f(y, a), \forall a \in Q\}$. Hence, since each P_i selected randomly, then $P \preceq Q$ holds. This completes the proof.

Let $DS = (U, C, D)$ be a decision system with $P \subseteq C$. For any $x_i, x_j \in U$, x_i and x_j conflict with each other from P to D if and only if $f(x_i, a) = f(x_j, a)$ for any $a \in P$, and $f(x_i, d) \neq f(x_j, d)$, where $d \in D$. An instance $x \in U$ is a consistent instance in the DS if and only if there does not exist an instance $y \in U$, which conflicts with $x \in U$. Hence, we have the conclusion that the DS is a consistent decision system if and only if each instance $x \in U$ is a consistent instance.

Let $DS = (U, C, D)$ be a decision system with $X \subseteq U$ a subset of universe, attribute subsets P, $Q \subseteq C \cup D$, then $\underline{P}(X) = \cup \{[x]_P | [x]_P \subseteq X\}$ is called P-lower approximation of X. The P-positive region of Q is denoted by

$$POS_P(Q) = \cup \{\underline{P}(X) \mid X \in U/Q\}.$$

Let $DS = (U, C, D)$ be a decision system with any attribute subsets $P \subseteq C \cup D$, to make $a \in P$, and a in P is dispensable for D, if $POS_P(D) = POS_{P-\{a\}}(D)$. Otherwise a is necessary. Then, P is independent relative to D, if every element in P is indispensable for D.

Let $DS = (U, C, D)$ be a decision system. The elements in $POS_C(D)$ are regarded as the objects of consistent set, and the elements in $U - POS_C(D)$ are regarded as the objects of inconsistent set [20].

Let $DS = (U, C, D)$ be a decision system with $U/(C \cup D) = \{[U'_1]_{C \cup D}, [U'_2]_{C \cup D}, \ldots, [U'_n]_{C \cup D}\}$, where $U = \{U_1, U_2, \ldots, U_m\}$, $n \leq m$, and $U'_i \in U$, then $U' = \{U'_1 \cup U'_2 \cup \ldots \cup U'_n\}$. Then (U', C, D) is called a simplified decision system (*SDS*). It is obvious that by virtue of this technology of simplicity lots of redundancy information is deleted, and then the space complexity of the DS is decreased [5].

Theorem 2. Let $SDS = (U', C, D)$ be a simplified decision system. If there exists $IND(C) \subseteq IND(D)$, then the *SDS* is referred to as a consistent simplified decision system (*CSDS*). Otherwise, the *SDS* is referred to as an inconsistent simplified decision system (*ISDS*).

Proof. It is straightforward.

Theorem 3. Let $SDS = (U', C, D)$ be a simplified decision system. If $POS_C(D) = U'$, then we say that the S is consistent, otherwise the S is inconsistent.

Proof. It can be derived directly from the definition of positive region and Theorem 2.

Let $DS = (U, C, D)$ be a decision system with P, $Q \subset C \cup D$, $U/P = \{X_1, X_2, \ldots, X_n\}$, $U/Q = \{Y_1, Y_2, \ldots, Y_m\}$, then the conditional information entropy of knowledge Q with reference to P in [39] is denoted by

$$H(Q \mid P) = -\sum_{i=1}^{n} \frac{|X_i|}{|U|} \sum_{j=1}^{m} \frac{|Y_j \cap X_i|}{|X_i|} \log \frac{|Y_j \cap X_i|}{|X_i|}.$$

III. Decision Inclusion Degree

A. Limitations of Representative Reduction Methods

Firstly, it is known that matrix theory is the core content in advanced algebra. Many of ideas and methods in matrix theory have greatly enriched algebraic theory of mathematics. With deepening of people researching in science, application of matrix theory becomes wider. However, in classical reduction methods, core attribute set can be found by discernible matrix method, and then matrix elements containing core attributes are deleted from discernible matrix. Then, matrix elements free of core attributes with disjunctive form are turned into conjunctive form expression. At last, one makes reduction in this expression, and then it becomes a disjunctive normal form. But the process of transforming conjunctive normal form into disjunctive normal form is very complicated, and it always causes that the time complexity of discernible matrix method does exponential growth with the increase of system size. Each of disjunctive normal form with core attribute set is reduction of decision systems, such that all reducts can be obtained. However, it is unnecessary to compute all reductions in practical problems because people usually only concern about how to find minimum reduction. Therefore, it is concluded that matrix methods in classical rough set theory cannot search the minimum or suboptimal reduction effectively.

Secondly, in a decision system $DS = (U, C, D)$, a reduct of the *DS*, named as a positive region reduct for convenience, is presented in [30, 38] as follows: for any $P \subseteq C$ and $D = \{d\}$, if $POS_P(D) = POS_C(D)$ and $POS_Q(D) \neq POS_C(D)$ for any $Q \subset P$, then P is a positive region reduct of the *DS*. That is, whether or not any condition attribute is redundant depends on whether or not the positive region is changed. Thus, these presented algorithms in [30, 38] only reflect whether or not the prediction of deterministic decision rules has change after reduction [5]. Therefore, if new inconsistent objects are added to the *DS*, it is not considered whether the probability distribution generated by the primary inconsistent objects is changed in their corresponding decision blocks.

Thirdly, in a decision system $DS = (U, C, D)$, a reduct of the *DS*, named as an information entropy reduct for convenience, is presented in [39] as follows: for any $P \subseteq C$ and $D = \{d\}$, if $H(D|P) = H(D|C)$ and $H(D|Q) \neq H(D|C)$ for any $Q \subset P$, then P is an information entropy reduct of the *DS*. That is, whether or not any condition attribute is redundant depends on whether or not the conditional information entropy value of decision system is changed. However, in practical application, there exist new added and primary inconsistent objects in decision blocks, hence, if their probability distribution is changed [5]. Thus, the main criterions of algorithms in [30, 38, 39] in evaluating decision ability only think about the change of certainty factor for all decision rules after reduction.

B. Decision Inclusion Degree and Probability Distribution Function

Inclusion degree is a kind of soft computing method to deal with fuzzy and uncertain knowledge [40]. Data analysis based on inclusion degree is one of main application technologies in rough set theory, which is mainly used to analyze rough classification, attribute dependency, attribute significance, and so on. Uncertainty reasoning methods can be summed up in a special kind of inclusion degrees [20, 40]. Set X is a universe, A and B are two subsets of X. Degree of collection A included in

set B is $D(B/A)$ which is called inclusion degree. In the meantime, knowledge acquisition from a large number of cases for some rules, and rules of the before and after parts relationship is also a kind of actually closed inclusion, so you can use inclusion degree theory to study uncertain rules. In decision systems, decision rules can be extracted. Certain rules can be extracted from consistent decision systems, but uncertain rules or possible rules only can be extracted from inconsistent decision systems.

Theorem 4. Let $SDS = (U', C, D)$ be a simplified decision system with $P, Q \subseteq C$, $U'/D = \{D_1, D_2, \ldots, D_m\}$. Then $POS_P(D) = POS_Q(D)$ if and only if $\underline{P}(D_i) = \underline{Q}(D_i)$, where $i = 1, 2, \ldots, m$.

Proof. Suppose that $\underline{P}(D_i) \neq \underline{Q}(D_i)$, $i = 1, 2, \ldots, m$, if $POS_P(D) = POS_Q(D)$, it follows from the definition of positive region that one has $\underline{P}(D_i) = \underline{Q}(D_i)$. This yields a contradiction. Thus, $POS_P(D) = POS_Q(D) \Leftrightarrow \underline{P}(D_i) \neq \underline{Q}(D_i)$, where $i = 1, 2, \ldots, m$. This completes the proof.

In a simplified decision system $SDS = (U', C, D)$, suppose that $D_0 = U' - POS_C(D)$, it follows that $\underline{C}D_0 = D_0$. That is, all inconsistent objects $U' - POS_C(D)$ detached form the unattached set D_0. Then, suppose that $\underline{C}D_i \neq \emptyset$, one has another decision partition $\{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$ of C on U, and then a new equivalent relation can be constructed, denoted by R_D. Similar to [17], it follows that there exists $U'/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$. It can be concluded that the presented decision partition U'/R_D has not only detached consistent objects from different decision blocks in U, but also distinguished consistent objects from inconsistent objects.

Definition 1. Let $ISDS = (U', C, D)$ be an inconsistent simplified decision system with $P \subseteq C$, $U'/D = \{D_1, D_2, \ldots, D_m\}$, $U'/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$, and $\forall u \in U'$. The decision inclusion degree is denoted by $D(\underline{C}D_i/[u]_P)$, is defined as

$$D(\underline{C}D_i / [u]_P) = \frac{|\underline{C}D_i \cap [u]_P|}{|[u]_P|},$$

where $i = 1, 2, \ldots, m$.

Definition 2. Let $ISDS = (U', C, D)$ be an inconsistent simplified decision system with $P \subseteq C$, $U'/D = \{D_1, D_2, \ldots, D_m\}$, $U'/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$, and $\forall u \in U'$. The probability distribution function P with respect with u in U' is denoted by $\mu_P(u)$, is defined as

$$\mu_P(u) = \left(\frac{|\underline{C}D_0 \cap [u]_P|}{|[u]_P|}, \frac{|\underline{C}D_1 \cap [u]_P|}{|[u]_P|}, \ldots, \frac{|\underline{C}D_m \cap [u]_P|}{|[u]_P|}\right).$$

From Definition 2, it can be obtained the following property immediately.

Property 1. Let $ISDS = (U', C, D)$ be an inconsistent simplified decision system with $\forall u, v \in U'$. Then

$$\forall u, v \in [u]_P \Rightarrow \mu_P(u) = \mu_P(v).$$

Property 1 states that in an inconsistent simplified decision system $ISDS = (U', C, D)$, for any $X \in U/C$, the probability distribution function of each equivalence class X only need to be calculated.

Let $DS = (U, C, D)$ be a decision system and $P \subseteq C$. If $U/P = \{X_1, X_2, \ldots, X_n\}$, $D = \{d\}$, $U/D = \{Y_1, Y_2, \ldots, Y_m\}$, and $U/R_D = \{\underline{C}Y_0, \underline{C}Y_1, \underline{C}Y_2, \ldots, \underline{C}Y_m\}$, then let $H(R_D|P)$ denote the conditional rough entropy of D with reference to P of DS in [17] as follows

$$H(R_D \mid P) = \sum_{i=1}^{n} \frac{|X_i|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}Y_j \cap X_i|}{|X_i|} \log_2 \frac{|X_i|}{|\underline{C}Y_j \cap X_i|}.$$

Theorem 5. Let $ISDS = (U', C, D)$ be an inconsistent simplified decision system with $P \subseteq C$, $U'/P = \{X_1, X_2, \ldots, X_n\}$, $U'/C = \{Y_1, Y_2, \ldots, Y_k\}$, $U/D = \{D_1, D_2, \ldots, D_m\}$, $U'/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$, and $\forall u \in U'$. $H(R_D \mid P) = H(R_D \mid C) \Leftrightarrow \forall u \in U' \Rightarrow \mu_P(u) = \mu_C(u)$.

Proof. ($\Rightarrow$) Suppose that $P \subseteq C$, it follows from the above partial order that $C \preceq P$. Then, it can draw from Proposition 10 in [17] that $H(R_D|C) \leq H(R_D|P)$. For $\forall u \in U'$, when $H(R_D|P) = H(R_D|C)$, assume that $\mu_P(u) = \mu_C(u)$ is not true, then there exists u_0 such that $\mu_P(u_0) \neq \mu_C(u_0)$. Thus one has that

$$\left(\frac{|\underline{C}D_0 \cap [u_0]_P|}{|[u]_P|}, \frac{|\underline{C}D_1 \cap [u_0]_P|}{|[u]_P|}, \ldots, \frac{|\underline{C}D_m \cap [u_0]_P|}{|[u]_P|}\right) \neq$$
$$\left(\frac{|\underline{C}D_0 \cap [u_0]_C|}{|[u]_C|}, \frac{|\underline{C}D_1 \cap [u_0]_C|}{|[u]_C|}, \ldots, \frac{|\underline{C}D_m \cap [u_0]_C|}{|[u]_C|}\right).$$

Hence, it can be obtained that $[u_0]_P \neq [u_0]_C$. One has that $[u_0]_C \subset [u_0]_P$ from the above partial order. Assume that $[u_0]_P = \bigcup \{[u_i]_C \mid u_i \in U', i = 0, 1, 2, \ldots, l, \text{ and } 1 \leq l \leq |U'|\}$, for $\forall s \neq t$ ($s, t \in \{0, 1, 2, \ldots, l\}$), and $[u_s]_C \cap [u_t]_C = \emptyset$. Then there exist at least $s_0, t_0 \in \{0, 1, 2, \ldots, l\}$ and $s_0 \neq t_0$ such that $\mu_C(u_{s_0}) \neq \mu_C(u_{t_0})$, otherwise $\mu_P(u_0) = \mu_C(u_0)$. Thus one has that

$$\left(\frac{|\underline{C}D_0 \cap [u_{s_0}]_C|}{|[u_{s_0}]_C|}, \frac{|\underline{C}D_1 \cap [u_{s_0}]_C|}{|[u_{s_0}]_C|}, \ldots, \frac{|\underline{C}D_m \cap [u_{s_0}]_C|}{|[u_{s_0}]_C|}\right) \neq$$
$$\left(\frac{|\underline{C}D_0 \cap [u_{t_0}]_C|}{|[u_{t_0}]_C|}, \frac{|\underline{C}D_1 \cap [u_{t_0}]_C|}{|[u_{t_0}]_C|}, \ldots, \frac{|\underline{C}D_m \cap [u_{t_0}]_C|}{|[u_{t_0}]_C|}\right).$$

It can be obtained from Proposition 8 in [17] that for $\forall X_i, X_j \in U'/P$, $X_i \neq X_j$, $\forall \underline{C}D_k \in U'/R_D$, if $X_i \bigcup X_j \in U'/C$ and $\frac{|X_i \cap \underline{C}D_k|}{|X_i|} = \frac{|X_j \cap \underline{C}D_k|}{|X_j|}$ always holds, then $H(R_D|P) = H(R_D|C)$. Thus, it is obvious from $\mu_C(u_{s_0}) \neq \mu_C(u_{t_0})$ that $H(R_D|P) > H(R_D|C)$, which contradicts with the above hypothesis that $H(R_D|P) = H(R_D|C)$. Therefore, it can be obtained that $\forall u \in U' \Rightarrow \mu_P(u) = \mu_C(u)$. ($\Leftarrow$) Suppose that $P \subseteq C$, if $\forall u \in U' \Rightarrow \mu_P(u) = \mu_C(u)$, it follows from the above partial order that $C \preceq P$, and then one has that $[u]_P = [u_1]_C \bigcup [u_2]_C \bigcup \ldots \bigcup [u_l]_C$, where $u_1, u_2, \ldots, u_l \in U'$, $1 \leq l \leq |U'|\}$, $\forall s \neq t$ ($s, t \in \{0, 1, 2, \ldots, l\}$), and $[u_s]_C \cap [u_t]_C = \emptyset$. Thus, it can be obtained that $\mu_P(u_1) = \mu_P(u_2) = \ldots = \mu_P(u_l)$. It follows from Definition 4 in [17] that

$$\sum_{i=1}^{l} \frac{|[u_i]_C|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap [u_i]_C|}{|[u_i]_C|} \log_2 \frac{|[u_i]_C|}{|\underline{C}D_j \cap [u_i]_C|}$$
$$= \sum_{i=1}^{l} \frac{|[u_i]_C|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap [u]_B|}{|[u]_B|} \log_2 \frac{|[u]_B|}{|\underline{C}D_j \cap [u]_B|}$$

$$= \frac{|[u_1]_C \cup [u_2]_C \cup ... \cup [u_l]_C|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap [u]_B|}{|[u]_B|} \log_2 \frac{|[u]_B|}{|\underline{C}D_j \cap [u]_B|}$$
$$= \frac{|[u]_B|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap [u]_B|}{|[u]_B|} \log_2 \frac{|[u]_B|}{|\underline{C}D_j \cap [u]_B|}.$$

Hence, one has that

$$H(R_D \mid C) = \sum_{i=1}^{k} \frac{|Y_i|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap Y_i|}{|Y_i|} \log_2 \frac{|Y_i|}{|\underline{C}D_j \cap Y_i|}$$
$$= \sum_{i=1}^{n} \frac{|X_i|}{|U|} \sum_{j=0}^{m} \frac{|\underline{C}D_j \cap X_i|}{|X_i|} \log_2 \frac{|X_i|}{|\underline{C}D_j \cap X_i|}$$
$$= H(R_D \mid P),$$

Thus, $H(R_D|P) = H(R_D|C)$ holds. This completes the proof.

Theorem 5 shows that in an inconsistent simplified decision system $ISDS = (U', C, D)$, the conditional rough entropy of D with reference to $P \subseteq C$ has the same discriminating capability as the probability distribution function P with respect with u in U' when their equations hold.

IV. Feature Selection Of Inconsistent Decision Systems

A. Decision Inclusion Degree-based Distribution Reduction

Distribution reduction which is based on inclusion degree [40] can obtain all reducts of a decision system, and we can always use discernible matrix method to find core attribute set in both consistent and inconsistent decision system. Since core and reduction are the most important concepts of knowledge reduction in decision systems and the main goal of using discernible matrix is also for calculating core and reduction, but both time complexity and space complexity of existing methods on the basis of discernible matrix for core are not ideal. In order to overcome the limitations of these above methods which seek core and reduction through discernible matrix to improve operational efficiency, it is necessary to find a new heuristic method. Attribute significance based on positive region in classical rough set theory only makes a quantitative description for positive region cardinality, while attribute significance based on conditional information entropy only describes separation of objects belonging to different decision classes in equivalence classes of condition attribute subset without considering separation of consistent and inconsistent objects that have the same decision attribute values. Due to inconsistent objects in inconsistent decision systems, the existing methods based on positive region and conditional information entropy cannot equally express knowledge reduction [5, 17, 30, 39]. Then, if all inconsistent objects are separated from consistent objects, it is helpful to search for the minimum or suboptimal reduction.

Definition 3. Let $ISDS = (U', C, D)$ be an inconsistent simplified decision system with $P \subseteq C$, $U'/D = \{D_1, D_2, \ldots, D_m\}$, $U'/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, \ldots, \underline{C}D_m\}$. For $\forall u_i \in U'$, $\mu_P(u_i) = (\underline{C}D_0^P(u_i), \underline{C}D_1^P(u_i), ..., \underline{C}D_m^P(u_i))$, and $\gamma_P(u_i) = \{\underline{C}D_h \mid \underline{C}D_h^P(u_i) = \underset{0 \le j \le m}{Max} \underline{C}D_j^P(u_i)\}$, where $\underline{C}D_j^P(u_i) = \frac{|\underline{C}D_j \cap [u_i]_P|}{|[u_i]_P|}$, $j = 1, 2, \ldots, m$. Then

(1) P is called a distribution set of the $ISDS$ if $\mu_P(u_i) = \mu_C(u_i)$ for $i = 1, 2, \ldots, |U'|$. P is called a distribution reduct of the $ISDS$ if and only if $\mu_P(u_i) = \mu_C(u_i)$ for $i = 1, 2, \ldots, |U'|$, and for $\forall P' \subset P$, there exist $u_j \in U'$ such that $\mu_{P'}(u_j) \neq \mu_C(u_j)$.

(2) P is called a maximum distribution set of the $ISDS$ if $\gamma_P(u_i) = \gamma_C(u_i)$ for $i = 1, 2, \ldots, |U'|$. P is called a maximum distribution reduct of the $ISDS$ if and only if $\gamma_P(u_i) = \gamma_C(u_i)$ for $i = 1, 2, \ldots, |U'|$, and for $\forall P' \subset P$, there exist $u_j \in U'$ such that $\gamma_{P'}(u_j) \neq \gamma_C(u_j)$.

Definition 3 states that in an inconsistent simplified decision system $ISDS = (U', C, D)$, if attribute subset $P \subseteq C$ is a distribution reduct of an inconsistent decision system, then rules coming from P and C have the same reliability. Then, from Definition 3, it can be obtained the following property immediately.

Property 2. Let $CSDS = (U', C, D)$ be a consistent simplified decision system. Since $D_0 = U' - POS_C(D) = \emptyset$, one has that $\underline{C}D_0 = \emptyset$, then the probability distribution function P with respect with u in U' degenerates into the general probability distribution function, and the conditional rough entropy degenerates into the conditional information entropy in [39].

Property 2 illustrates that the probability distribution function in consistent decision systems is a special instance of that in inconsistent decision systems. This means that the definition of probability distribution function in consistent decision systems is a consistent extension in inconsistent decision systems. It follows that the decision inclusion degree in an inconsistent decision system is suitable for measuring the uncertainty of both inconsistent and consistent decision systems. Therefore, the distribution reduct and the maximum distribution reduct are suitable for both inconsistent and consistent decision systems. In what follows, the inconsistent or consistent simplified decision systems can be unified into the simplified decision systems.

Definition 4. Let $SDS = (U', C, D)$ be a simplified decision system with $P \subseteq C$. For $\forall a \in P$ and $\forall u \in U'$, the significance measure of a in P with reference to D is denoted by $SIG^{inner}(a, P, D)$, defined as

$$SIG^{inner}(a, P, D) = \frac{\{\forall u \in U' \mid \mu_{P-\{a\}}(u) \neq \mu_P(u)\}}{|U'|}.$$

Definition 5. Let $SDS = (U', C, D)$ be a simplified decision system with $P \subseteq C$. For $\forall a \in C - P$ and $\forall u \in U'$, the significance measure of a in P with reference to D is denoted by $SIG^{outer}(a, P, D)$, defined as

$$SIG^{outer}(a, P, D) = \frac{\{\forall u \in U' \mid \mu_{P \cup \{a\}}(u) \neq \mu_P(u)\}}{|U'|}.$$

According to Definitions 4 and 5, it can be obtained the following properties immediately.

Property 3. $0 \le SIG^{inner}(a, P, D) \le 1$.

Property 4. $0 \le SIG^{outer}(a, P, D) \le 1$.

Property 5. when $P = C$, $SIG^{outer}(\ , C, D) = 0$.

Property 6. $\forall\ a \in C - P$ is a dispensable attribute if and only if $SIG^{outer}(a, P, D) = 0$.

Definition 5 shows that the significance measure $SIG^{outer}(a, P, D)$ indicates the importance of attribute a added to $P \subseteq C$ with reference to D in a simplified decision system $SDS = (U', C, D)$, offering the powerful reference to the decision. Furthermore, the bigger the significance measure of attribute is, the higher its position in the decision system is, otherwise the lower its position is. Thus, all the definitions above are used as heuristic information for feature selection algorithm to select a reduct from consistent or inconsistent data sets. It is known that the intersection of all attribute reducts is said to be indispensable and is called the core in a decision system. Each attribute in the core must be in every attribute reduction of the decision system. Then, the significance measures above can be used to find the core attributes. The following properties are of interest with this regard.

Property 7. Let $SDS = (U', C, D)$ be a simplified decision system with $P \subseteq C$. $\forall\ a \in P$ is indispensable in P with reference to D if and only if $SIG^{inner}(a, P, D) > 0$.

Property 8. Let $SDS = (U', C, D)$ be a simplified decision system. For $\forall\ a \in C$, if $SIG^{inner}(a, C, D) > 0$, then a is a core attribute of the SDS, i.e., $CORE = \{a \in C \mid SIG^{inner}(a, C, D) > 0\}$.

Theorem 6. Let $SDS = (U', C, D)$ be a simplified decision system with $P \subseteq C$. P is a distribution reduct of C relative to D if $\mu_P(u) = \mu_C(u)$ for $\forall\ u \in U'$ and $SIG^{inner}(a, P, D) > 0$ for $\forall\ a \in P$.

Proof. It can be derived directly from Definition 3 and Properties 6 and 7.

B. Feature Selection Algorithm of Inconsistent Decision Systems

In the following, we focus on how to improve computational efficiency of a heuristic feature selection algorithm. Then we introduce the idea of radix sorting in [30] and hash in [31] to calculate equivalence blocks and positive region effectively. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure using a granulation order. Thus, through the decomposition of $SIG^{outer}(a, P, D)$, it can be seen easily that every time to calculate any attribute a with the maximum of $SIG^{outer}(a, P, D)$ is in fact to calculate that with the maximum of $\mu_{P\cup\{a\}}(u)$, because $\mu_P(u)$ is a constant when we calculate $SIG^{outer}(a, P, D)$. Therefore, we only need calculate $\mu_{P\cup\{a\}}(u)$ except $\mu_P(u)$. Thus, the above policies will help to reduce the quantity of computation and the time-space of search. Formally, we can now construct a distribution reduct algorithm, also called an efficient feature selection algorithm based on decision inclusion degree (FSDID) for inconsistent decision systems as follows.

Algorithm 1. FSDID

Input: An inconsistent decision system $IDS = (U, C, D)$, where $C = \{c_1, c_2, \ldots, c_{|C|}\}$, and $D = \{d\}$

Output: *reduct*, a reduct of *IDS*

(1) Let $CORE = \emptyset$, $R = \emptyset$

(2) Calculate U/C, U/D and $U/(C \cup D)$ incrementally to get U' by radix sorting, and obtain $POS_C(D)$ and $U' - POS_C(D)$ by hash, then get U'/R_D

(3) Calculate $\mu_C(u)$ and $\mu_{C-\{c_i\}}(u)$ to get $CORE = \{c_i \in C \mid SIG^{inner}(a, C, D) > 0\}$ for $\forall\ u \in U'$ and $i = 1, 2, \ldots, |C|$, then let $R = CORE$ and go to (5)

(4) Select a_i with $max\{\mu_{P\cup\{a_i\}}(u)\}$ by radix sorting to put a_i into H, where $\forall\ a_i \in C - R$
// Select a_i with $max\{SIG^{outer}(a_i, P, D)\}$
(4.1) If $|H| \ne 1$, select $a_i \in H$ with $min\{|U/(P \cup \{a_i\})|\}$
(4.2) If the selected is not only, then select the front
(4.3) $R = R \cup \{a_i\}$

(5) If $\mu_R(u) \ne \mu_C(u)$, then go to (4), else
(5.1) Let $R = R - CORE$;
(5.2) $t = |R|$;
(5.3) For ($i = 1$; $i \le t$; i++)
(5.3.1) $a_i \in R$;
(5.3.2) $R = R - \{a_i\}$;
(5.3.3) If $\mu_{R\cup CORE}(u) \ne \mu_C(u)$, then $R = R \cup \{a_i\}$

(6) $reduct = R \cup CORE$

(7) End

Remark. The above steps for feature selection algorithm of distribution reduct should be of reference in obtaining the maximum distribution reduct in both inconsistent and consistent decision systems. It can be easily seen that Step 5 in FSDID algorithm ensures that the distribution reduct is complete, which can ensure that the final reduct will be obtained. By calculation and analysis, the total worst time complexity of FSDID algorithm is $O(|C||U|) + O((|C| - 1)|U|) + O((|C| - 2)|U|) + \ldots + O(|U|) = O(|C|^2|U|)$, which is below the time complexity of these methods in [22, 24, 27-29, 32, 33, 35, 38, 39]. After comparison, it can be easily known that the algorithm proposed in this paper is effective and available. Furthermore, the worst space complexity of FSDID algorithm is $O(|C||U|)$.

V. Experimental Results

In this section, we apply the proposed approach and other feature selection approaches in several data sets from the UCI Repository of machine learning databases, to evaluate the proposed approach. In the following, their advantages and disadvantages can be further found easily through comparing roundly the Algorithm 4 in [38] and the Algorithm CEBARKCC in [39] with the proposed FSDID algorithm, shortly denoted by Alg_a, Alg_b, and Alg_c, respectively. Here we choose six discrete databases from UCI datasets and use three algorithms above to do more experiments on PC (Inter(R) Pentium(R) D CPU 3.4 GHz, 2 GB memory, Windows XP). Then the comparison results of three feature selection algorithms are outlined in Table I.

TABLE I.
COMPARISON RESULTS FOR DIFFERENT FEATURE SELECTION ALGORITHMS

NO.	Dataset	Objects	Attributes	Consistent or not	Selected attributes		
					Alg_a	Alg_b	Alg_c
1	Liver-disorders	345	7	Yes	3	3	3
2	Zoo	101	17	No	10	9	8
3	Vehicle	946	20	Yes	4	4	4
4	Mushroom	8124	23	Yes	5	4	3
5	Voting-records	435	17	Yes	10	9	9
6	Breast cancer-wisconsin	683	10	Yes	5	4	4

VI. CONCLUSIONS

Dataset dimensionality is one of the primary impediments to data analysis areas. An important step prior to constructing a classifier for a very large data set is feature selection. In this regard, by distinguishing consistent objects form inconsistent objects, the decision inclusion degree, the probability distribution function, the distribution reduct and the maximum distribution reduct are presented for both inconsistent and consistent simplified decision systems. Furthermore, many important properties and propositions are discussed as well. An effective heuristic feature selection algorithm in inconsistent decision systems with less time-space complexity are put forward as a distribution reduct. The theoretical analyses show that the time complexity of this method is lower than that of existing representative feature selection methods. Meanwhile, the experiment results are consistent with our theoretical analysis. In sum, the proposed method is an effective means of feature selection for both inconsistent and consistent decision systems, especially large ones.

ACKNOWLEDGMENT

We are highly grateful to the anonymous reviewers, referees and Editor-in-Chief for their valuable comments and hard work.

This work was supported by the National Natural Science Foundation of China (Nos. 60873104, 61370169), the Key Project of Science and Technology Department of Henan Province (No. 112102210194), the Science and Technology Research Key Project of Educational Department of Henan Province (Nos. 12A520027, 13A52 0529), and the Education Fund for Youth Key Teachers of Henan Normal University.

REFERENCES

[1] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic Publishers, 1991.

[2] J. C. Xu, L. Sun, and Q. Q. Zhang, *Theories and Methods of Granular Computing and Its Uncertainty Information Measures*, Beijing: Academic Press, 2013.

[3] L. Sun, J. C. Xu, and Y. Tian, "Feature selection using rough entropy-based uncertainty measures in incomplete decision systems", *Knowledge-Based Systems*, vol. 36, pp. 206−216, 2012.

[4] L. Sun, J. C. Xu, S. Q. Li, X. Z. Cao, and Y. P. Gao, "New approach for feature selection by using information entropy", *Journal of Information and Computational Science*, vol. 8, pp. 2259−2268, 2011.

[5] L. Sun, J. C. Xu, Z. A. Xue, and J. Y. Ren. "Decision degree-based decision tree technology for rule extraction", *Journal of Computers*, vol. 7, pp. 1769−1779, 2012.

[6] Z. C. Lu, Z. Qin, Y. Q. Zhang, and J. Fang, "A fast feature selection approach based on rough set boundary regions", *Pattern Recognition Letters*, vol. 36, pp. 81−88, 2014.

[7] L. Sun and J. C. Xu, "A granular computing approach to gene selection", *Bio-Medical Materials and Engineering*, vol. 24, pp. 1307−1314, 2014.

[8] L. Sun, J. C. Xu, J. Y. Ren, and T. H. Xu, "Granularity partition-based feature selection and its application in decision systems", *Journal of Information and Computational Science*, vol. 9, pp. 3487−3500, 2012.

[9] L. Sun, J. C. Xu, Y. W. Hu, and L. N. Du, "Granular space-based feature selection and its applications", *Journal of Software*, vol. 8, pp. 817−826, 2013.

[10] G. Y. Wang, Y. Y. Yao, and H. Yu. "A survey on rough set theory and its application", *Chinese Journal of Computers*, vol. 32, pp. 1229−1246, 2009.

[11] L. Sun and J. C. Xu, "Feature selection using mutual information based uncertainty measures for tumor classification", *Bio-Medical Materials and Engineering*, vol. 24, pp. 763−770, 2014.

[12] L. Sun, J. C. Xu, Z. A. Xue, and L. J. Zhang, "Rough entropy-based feature selection and its application", *Journal of Information and Computational Science*, vol. 8, pp. 1525−1532, 2011.

[13] S. K. M. Wong and W. Ziarko, "On optimal decision rules in decision tables", *Bulletin of the Polish Academy of Sciences*, vol. 33, pp. 693−696, 1985.

[14] J. C. Xu and L. Sun, "Knowledge entropy and feature selection in incomplete decision systems", *Applied Mathematics & Information Sciences*, vol. 7, pp. 829−837, 2013.

[15] J. C. Xu and L. Sun, "A new knowledge reduction algorithm based on decision power in rough set", *Transactions on Rough Sets*, vol. 12, pp. 76−89, 2010.

[16] L. Sun, J. C. Xu, and Y. P. Song, "Information quantity-based decision rule acquisition from decision tables", *Journal of Convergence Information Technology*, vol. 7, pp. 57−67, 2012.

[17] L. Sun, J. C. Xu, and L. J. Zhang, "Approaches to knowledge reduction of decision systems based on conditional rough entropy", *International Journal of Advancements in Computing Technology*, vol. 3, pp. 129−139, 2011.
[18] L. Zhang and B. Zhang, "The quotient space theory of problem solving", *Fundamenta Informaticae*, vol. 59, pp. 287−298, 2004.
[19] L. Sun, J. C. Xu, C. Wang, T. H. Xu, and J. Y. Ren, "Granular computing-based granular structure model and its application in knowledge retrieval", *Information Technology Journal*, vol. 11, pp. 1714−1721, 2012.
[20] L. Sun, J. C. Xu, and Y. Y. Ma, "New reduction method based on inclusion degree in inconsistent decision table", *Chinese Computer Engineering and Applications*, vol. 43, pp. 166–168, 2007.
[21] M. Beynon, "Reducts within the variable precision rough sets model: a further investigation", *European Journal of Operational Research*, vol. 134, pp. 592−605, 2001.
[22] M. Kryszkiewicz, "Comparative study of alternative type of knowledge reduction in inconsistent systems", *International Journal of Intelligent Systems*, vol. 16, pp. 105−120, 2001.
[23] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models", *Information Sciences*, vol. 178, pp. 3356−3373, 2008.
[24] W. X. Zhang, J. S. Mi, and W. Z. Wu, "Approaches to knowledge reductions in inconsistent systems", *International Journal of Intelligent Systems*, vol. 18, pp. 989−1000, 2003.
[25] J. S. Mi, W. Z. Wu, and W. X. Zhang, "Approaches to knowledge reductions based on variable precision rough sets model", *Information Sciences*, vol. 159, pp. 255−272, 2004.
[26] W. Z. Wu, M. Zhang, H. Z. Li, and J. S. Mi, "Knowledge reduction in random information systems via dempster-shafer theory of evidence", *Information Sciences*, vol. 174, pp. 143−164, 2005.
[27] D. Y. Ye, Z. J. Chen, and C. Y. Yu, "A novel maximum distribution reduction algorithm for inconsistent decision tables", in *Proceedings of International Conference on Knowledge Science, Engineering and Management*, J. Lang, F. Z. Lin, and J. Wang, Eds., Lecture Notes in Computer Science, vol. 4092, 2006, pp. 548−555.
[28] Q. Liu, L. T. Chen, J. Z. Zhang, and F. Min, "Knowledge reduction in inconsistent decision tables", in *Proceedings of International Conference on Advanced Data Mining and Applications*, X. Li, O. R. Zaïane, and Z. H. Li, Eds., Lecture Notes in Computer Science, vol. 4093, 2006, pp. 626–635.
[29] D. Q. Miao and G. R. Hu, "A heuristic algorithm for reduction of knowledge", *Chinese Journal of Computer Research and Development*, vol. 36, pp. 681–684, 1999.
[30] Z. Y. Xu, Z. P. Liu, B. R. Yang, et al., "A quick attribute reduction algorithm with complexity of max ($O(|C||U|)$,$O(|C|^2|U/C|)$)", *Chinese Journal of Computers*, vol. 29, pp. 391–399, 2006.
[31] Y. Liu, R. Xiong, and J. Chu, "Quick attribute reduction algorithm with hash", *Chinese Journal of Computers*, vol. 32, pp. 1493–1499, 2009.
[32] J. S. Mi, W. Z. Wu, and W. X. Zhang, "An approach to approximation deduction in inconsistent decision tables", in *Proceedings of Rough Sets, Fuzzy set, Data Mining, and Granular*, 2003, pp. 283–286.
[33] K. Li, Y. S. Liu, and L. Wang, "An attribute reduction algorithm of rough set", *Chinese Computer Engineering and Applications*, vol. 38, pp. 15–16, 2002.
[34] G. Y. Wang, "The calculation method of core properties in decision table", *Chinese Journal of Computers*, vol. 26, pp. 611–615, 2003.
[35] T. Bin and L. L. Li, "The discussion about attribute reduction algorithm based on clear matrix", *Chinese Computer Engineering and Applications*, vol. 40, pp. 184–186, 2004.
[36] K. Y. Qin, P. Zheng, and W. F. Du, "The relationship among knowledge reduction approaches", in *Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery*, Changsha, China, 2005, pp. 1232–1241.
[37] W. X. Zhang, Y. Leung, and W. Z. Wu, *Information System and Knowledge Discovery*, Beijing: Science Press, 2003.
[38] S. H. Liu, Q. J. Sheng, B. Wu, et al., "Research on efficient algorithms for rough set methods", *Chinese Journal of Computers*, vol. 26, pp. 524–529, 2003.
[39] G. Y. Wang, H. Yu, and D. C. Yang, "Decision table reduction based on conditional information entropy", *Chinese Journal of Computers*, vol. 25, pp. 759–766, 2002.
[40] W. X. Zhang, W. Z. Wu, J. Y. Liang, and D. Y. Li, *Rough Set Theory and Methods*, Beijing: Science Press, 2001.

Lin Sun works at College of Computer & Information Engineering, Henan Normal University. He is currently a Ph.D. Candidate at Beijing University of Technology. He received his B.S. and M.S. degree in Computer Science and Technology, Henan Normal University in 2003 and 2007, respectively. His main research interests include rough set, granular computing, bioinformatics, and data mining.