Decision Degree-based Decision Tree Technology for Rule Extraction

Lin Sun, Jiucheng Xu, Zhan’ao Xue, Jinyu Ren
College of Computer & Information Technology, Henan Normal University, Xinxiang, Henan 453007, China
Email: linsunok@gmail.com

***Abstract*—Traditional rough set-based approaches to reduct have difficulties in constructing optimal decision tree, such as empty branches and over-fitting, selected attribute with more values, and increased expense of computational effort. It is necessary to investigate fast and effective search algorithms. In this paper, to address this issue, the limitations of current knowledge reduction for evaluating decision ability are analyzed deeply. A new uncertainty measure, called decision degree, is introduced. Then, the attribute selection standard of classical heuristic algorithm is modified, and the new improved significance measure of attribute is proposed. A heuristic algorithm for rule extraction from decision tree is designed. The advantages of this method for rule extraction are that it needn’t compute relative attribute reduction of decision tables, the computation is direct and efficient, and the time complexity is much lower than that of some existing algorithms. Finally, the experiment and comparison show that the algorithm provides more precise and simplified decision rules. So, the work of this paper will be very helpful for enlarging the application areas of rough set theory.**

***Index Terms*—granular computing, rough set, decision table, decision tree, decision degree, rule extraction**

I. Introduction

As a recently renewed research topic, granular computing is a new concept and computational model, and may be regarded as a label of family of theories, methodologies, and techniques [1-3]. Rough set theory [4-7], as a robust mathematical framework of granular computing, has been widely applied to extract rules from information tables or decision tables [8-12]. For example, An et al. [8] constructed the lower approximation operation and the upper approximation operation generated by a binary relation and its inverse relation in order to induce the minimal decision rules to support the decision task. Ziarko [12] proposed a method of rule extraction from decision tables by reducing boundary area in decision tables. There are at least three approaches to reducing boundary area. The first and simplest technique is to try to increase the decision table “resolution” by adding more attributes or by increasing the precision of existing ones. The second is to provide another layer of decision tables, by treating each sub-domain of objects matching the description of an elementary set of the boundary area of the original decision tables as a domain (the universe) by itself. The third is based on the idea of treating the sub-domain of the original domain corresponding to the whole boundary area as the new domain by itself. However, in fact, the classification accuracy (the approximation measure) is constrained according to decision requirements or preference of decision makers. An obvious question is how to extract much simpler decision rules on the basis of keeping an approximation measure [11]. That is to say, relative knowledge reduction must be obtained before rule extraction from decision tables. Many types of reduct were proposed in the area of rough set and each of the reductions aimed at some basic requirement. For example, by eliminating some rigorous conditions required by the distribution reduct, a maximum distribution reduct was introduced by Mi et al. in [13]. Unlike the possible reduct [14], the maximum distribution reduct can derive decision rules that are compatible with the original system. However, the complexity of these approaches above is much worse, which is inconvenient to extract decision rules from decision tables.

In recent years, how to evaluate the decision performance of a decision rule has become a very important issue in rough set theory. For example, Greco et al. [15] used some well-known confirmation measures within the rough set approach to discover relationships in data in terms of decision rules. For a decision rule set induced from a decision table, three parameters are traditionally associated as follows: the strength, the certainty factor and the coverage factor of the rule. In many practical decision problems, we always adopt several rule extraction methods for the same decision table. In this case, it is very important to check whether or not each of the rule extraction approaches is suitable for the given decision table. In other words, it is desirable to evaluate the decision performance of the decision rule set extracted by each of the rule extraction approaches. This strategy can help a decision maker to determine which of rule extraction methods suits for a given decision table. To evaluate and compare different rule learning algorithms, many criteria, most notably predictive accuracy, simplicity and time complexity, have been employed [16]. Predictive accuracy aims at generating rule sets with a low misclassification error on unseen test data. Simplicity aims at finding rule sets that are as small as possible. This can be measured through the number of rules and literals in the rule set. Simplicity has served in the machine learning literature as the most prominent measure of

doi:10.4304/jcp.7.7.1769-1779

human comprehensibility as it is generally agreed that the smaller the rule set, the easier it is to understand. One would like to have low time complexity so that algorithms scale well on large data sets.

A number of different classification techniques have been extensively studied [17, 18] and the induction of decision trees is a well-known approach for knowledge discovery in databases [18-25]. Decision trees can systematically analyze information contained in a large amount of data source to extract valuable rules and relationships. It is a representation of a decision procedure for determining the class of a given instance. This process for systemizing complicated decision problems and turning them into manageable knowledge structures is to find the major appealing feature of decision tree induction learning. To construct a decision tree, we have to select appropriate attributes as the tree nodes. Many methods are available for attribute selection, such as the entropy-based methods, Bayesian networks, Gini index methods, etc [22]. Among the classification algorithms implemented with the decision tree induction, the ID3 [23] and its follow-up revisions are well recognized as the prime stream of research. For example, C4.5, ID4, and ID5 build the decision trees by the prepruning technology, while GID3 and DID3 focus more on fault tolerance and dealing with some specific attributes. Pang et al. [18] presented a constructive method for association rule extraction, where the knowledge of data was encoded into a Support Vector Machine classification tree (SVMT), and linguistic association rule was extracted by decoding of the trained SVMT. The method of rule extraction over the SVMT, in the spirit of decision-tree rule extraction, achieved rule extraction not only from SVM, but also over the decision-tree structure of SVMT.

Since the prepruning approach of decision tree induction is disadvantageous in terms of scope, depth, and accuracy when compared with ID3, and the other approaches are still based on the algorithm of ID3, the structure of these induction learning approaches is the same [24]. The ID3 algorithm is well known to be more suitable to deal with nominal attributes, since its formation of decision branches depends on the discretized values of attributes, and nominal attributes usually provide a more convenient differentiability of classification. However, in the process of inducing a decision tree [25], the rough set-based approach tends to partition instances too excessively, and thus will construct a large decision tree and reveal trivial details in the data. As a result, the comprehensive abilities of some leaf nodes will be decreased for that they contain too few instances. This is usually called over-fitting when inducing a classifier and consequently the constructed decision tree needs further pruning to enhance the generalization ability. In [22], the authors presented a new approach for inducing decision tree based on Variable Precision Rough Set Model. It aims to handle uncertain information during the process of inducing decision tree and generalizes the rough set-based approach to decision tree construction by allowing some extent misclassification when classifying objects.

But until now, the research on constructing decision tree from decision tables, especially in large-scale data sets, has few literatures reported on simplified rule extraction in the view of rough set theory. Our research aims to find a method for decision tree-based rule extraction without computing relative attribute reduction of a decision table in rough set theory. To mitigate this problem, the decision tree is a good rule extraction example that is recommended here for decision rule extraction because every rule generated by a decision tree represents a certain decision path that has a comprehensible rule antecedent and rule consequence. Then, the main objectives of this paper are to establish the decision degree by introducing the notions of the certainty factor and the coverage factor of rule in [15], investigate some of its important properties and propositions, and apply them to decision tree for rule extraction from decision tables. This paper focuses on creating such a solution.

The rest of this paper is organized as follows. In Section II, we review some basic concepts and propositions of rough set theory. The limitations of current knowledge acquisition approaches and the decision degree of a decision table are presented respectively, and some of useful propositions and properties are educed in Section III. In Section IV, the significance measure of attribute is improve, and a new method for decision tree-based rule extraction is proposed, following that is a theoretical analysis of time complexity of this algorithm, and illustrated with an example. Section V provides a discussion of the experimental methods and results in detail. In Section VI, we draw conclusions and outline our main directions for future work.

II. Preliminaries

In this section, we will review several basic concepts such as information table, decision table and partial relation. Detailed description and formal definitions of the theory can be found in [1, 4-7]. Throughout this paper, we suppose that the universe U is a non-empty finite set.

An information table (IT) is usually expressed in the following form: $IT = (U, A, \{V_a\}, f_a)_{a\in A}$, where,

(1) U is a non-empty finite set of objects, indicating a given universe;

(2) A is a non-empty finite set of attributes;

(3) V_a is a value set (domain) of $a\in A$;

(4) f_a is a function from U to V_a, denoted by a mapping f_a: $U \rightarrow V_a$, called the information function of the IT.

Also, $(U, A, \{V_a\}, f_a)_{a\in A}$ can be written more simply as (U, A), if V_a and f_a are understood.

Let $P \subseteq A$ determine a binary indistinguishable relation $IND(P)$, given by

$$IND(P) = \{(u, v)\in U\times U : f(a, u) = f(a, v),\ \forall\, a\in P\}. \quad (1)$$

Obviously, $IND(P) = \bigcap_{a\in P} IND(\{a\})$. It shows that $IND(P)$ is an equivalence relation on the set U. For $P\subseteq A$, the relation $IND(P)$ constitutes a partition of U, which is denoted by $U/IND(P)$, or just U/P. That is,

$U/P = \{[u_i]_P : u_i \in U\}$ is called an information on U, where $[u_i]_P = \bigcup \{u_j \in U : (u_i, u_j) \in IND(P)\} = \bigcup \{u_j \in U : f(a, u_i) = f(a, u_j), \forall a \in P\}$ is called an equivalence block (equivalence class) of u_i with reference to P, and every equivalence block is called an information granule.

In particular, if $U/P = \{X : X = \{u\}, u \in U\} = \omega$, it is called an identity relation, and if $U/P = \{X : X = \{U\}\} = \delta$, it is called a universal relation.

Now, we define a partial order on all partition sets of U. Let U/P and U/Q be two partitions of a finite set U, then we define that the partition U/Q is coarser than the partition U/P (or the partition U/P is finer than the partition U/Q), denoted by $P \preceq Q$, between partitions by $P \preceq Q \Leftrightarrow \forall P_i \in U/P, \exists Q_j \in U/Q \rightarrow P_i \subseteq Q_j$. 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 Q is strictly coarser than P (or P is strictly finer than Q) and write $P \prec Q$.

Proposition 1. Let $IT = (U, A)$ be an information table and $Q \subseteq P \subset A$. Then we have $P \preceq Q$.

Proof. Suppose $U/P = \{P_1, P_2, \ldots, P_t\}$, $U/Q = \{Q_1, Q_2, \ldots, Q_s\}$, for any $P_i = [u]_P \in U/P$, since $Q \subseteq P$, then one has that $P_i = [u]_P = \bigcup \{v \in U : f(a, u) = f(a, v), \forall a \in P\} \subseteq Q_j = [u]_Q = \bigcup \{v \in U : f(a, u) = f(a, v), \forall a \in Q\}$. Hence, since each P_i is selected randomly, then we have proved that $P \preceq Q$ always holds.

A decision table (DT) is a special case of an information table. A DT can be generated by dividing the attribute set A into two disjoint subsets, or by adding some attributes to A. A decision table is a tetrad $(U, A = C \bigcup D, G(V_{d \in D}), f_C)$, where,

(1) U is a non-empty, finite set of objects, indicating a given universe;
(2) C is a non-empty, finite set of attributes, called condition attribute set;
(3) D is a non-empty, finite set of attributes, called decision attribute set, and $C \cap D = \emptyset$;
(4) $G(V_d)$ is the power set of V_d, and V_d is the value set (domain) of decision attribute $d \in D$;
(5) f_C is the function f_C: $U \rightarrow G(V_d)$, called an information function of the DT.

Obviously, the properties derived in previous sections also hold for $(U, A = C \bigcup D, G(V_{d \in D}), f_C)$. The tetrad $(U, A = C \bigcup D, G(V_{d \in D}), f_C)$ is usually denoted as a triple (U, C, D) for short, that is, $DT = (U, C, D)$.

Proposition 2. Let $DT = (U, C, D)$ be a decision table. If there exists $IND(C) \subseteq IND(D)$, then the DT is referred to as a consistent decision table. Otherwise, the DT is referred to as an inconsistent decision table.

Proof. It is straightforward.

Let $DT = (U, C, D)$ be a decision table. If P, $Q \subset C \bigcup D$ are two equivalent relations belonging to U, then P positive region of Q is defined as

$$POS_P(Q) = \bigcup \{\underline{P}Y : Y \in U / Q\}, \tag{2}$$

where $\underline{P}Y = \bigcup \{X \in U / P : X \subseteq Y\}$, and U/Q is a partition of U by Q.

Thus, D depends on P in a degree. That is, the formulation of approximation quality of P is defined as

$$\gamma_P(D) = \frac{|POS_P(D)|}{|U|}. \tag{3}$$

Let $DT = (U, C, D)$ be a decision table and for any $P \subseteq C$, to make arbitrary $r \in P$, r in P is unnecessary for D, if $POS_P(D) = POS_{P-\{r\}}(D)$. Otherwise, r is necessary. Then, P is independent relative to D, if every element in P is necessary for D.

Proposition 3. Let $DT = (U, C, D)$ be a decision table, then, the positive region of the partition U/D with respect to C, denoted by $POS_C(D) = \{x \in U : x$ is a consistent instance$\}$, that is,

$$POS_C(D) = \bigcup \{X : X \in U / C \wedge \forall x, y \in X \Rightarrow f(x, d) = f(y, d), \exists d \in D\}. \tag{4}$$

Proof. It is straightforward.

Proposition 4. Let $DT = (U, C, D)$ be a decision table and $POS_C(D) = \{x \in U : x$ is a consistent instance$\}$. If the DT is a consistent decision table, we have $POS_C(D) = U$.

Proof. It is straightforward.

Definition 1. Let $DT = (U, C, D)$ be a decision table and $U / (C \bigcup D) = \{[U'_1]_{C \bigcup D}, [U'_2]_{C \bigcup D}, \ldots, [U'_n]_{C \bigcup D}\}$, where $U = \{U_1, U_2, \ldots, U_m\}$, $n \leq m$, and $U'_i \in U$, then $U' = \{U'_1, U'_2, \ldots, U'_n\}$. F': $U' \times (C \bigcup D) \rightarrow V'$ is called a new information function. It is said that the 6-tuple (U', C, F', D, G, V') is a simplified decision table (SDT). The 6-tuple (U', C, F', D, G, V') is usually denoted as (U', C, D) for short, that is, $SDT = (U', C, D)$.

Thus, from Definition 1, it is obvious that by virtue of this technology of simplicity lots of redundancy information is deleted, and then the space complexity of the DT is decreased. The time-space complexity for computing core and attribute reduction is also cut down more efficiently. Therefore, the simplified decision table introduced is necessary.

III. Decision Degree

In this section, the limitations of current knowledge acquisition approaches are analyzed, to deal with this issue, and then we introduce a new uncertainty measure, called decision degree. Some of its important properties and propositions are discussed as well.

A. Limitations of Knowledge Acquisition Approaches

In the following, we analyze the limitations of reduction algorithms based on the positive region [7, 26, 27] and the conditional information entropy [7, 27].

Firstly, in a DT, let $P \subseteq C$, and if the quality of approximation of P with respect to D is equal to the quality of approximation of C with respect to D, i.e. $\gamma_P(D) = \gamma_C(D)$, and there doesn't exist $P^* \subset P$ such that $\gamma_{P*}(D) = \gamma_C(D)$, then P is called the reduct of C with respect to D [26]. Therefore, whether or not any condition attribute is redundant depends on whether or not the lower approximation quality corresponding to decision block is changed. Accordingly, if new inconsistent objects are added to the *DT*, it is not considered whether the probability distribution generated by the primary inconsistent objects is changed in their corresponding decision blocks. Hence, if the generated deterministic decision rules are the same, then, they will support the same important standards for evaluating decision ability. Suppose that the deterministic decision rules generated are the same, that is, the prediction of these rules has not change. Thus, these presented algorithms above only reflect whether or not the prediction of deterministic decision rules has change after reduction.

Secondly, in a *DT*, let $P \subseteq C$, and if $H(D|P) = H(D|C)$ and *P* is independent relative to *D*, then *P* is called the reduct of *C* with respect to *D* [7, 27]. Therefore, whether or not any condition attribute is redundant depends on whether or not the conditional information entropy value of decision table is changed. Furthermore, it is known that the value of conditional information entropy generated by $POS_C(D)$ is 0, thus $U - POS_C(D)$ may change the value of conditional information entropy of the *DT*. There must exist some new added and primary inconsistent objects in their corresponding decision blocks, hence, if their probability distribution is changed, then it will change the value of conditional information entropy of the *DT*. Therefore, we come to the conclusion that the main criterions of algorithms above in evaluating decision ability include two aspects as follows: the number of deterministic decision rules and the certainty factor of non-deterministic decision rules.

Thus, the researchers above only think about the change of certainty factor for all decision rules after reduction. However, in decision application, besides the certainty factor of rule, the object coverage factor of rule is also one of the most important standards of evaluating decision ability. Then, we draw the conclusion that these current knowledge acquisition approaches above cannot reflect the change of decision ability objectively. Therefore, it is necessary to investigate a new uncertainty measure and effective search algorithm.

B. Representation of Decision Degree

Let $SDT = (U', C, D)$ be a simplified decision table. $U'/C = \{X_1, X_2, \ldots, X_n\}$ and $U'/D = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions on U' induced by the equivalence relations $IND(C)$ and $IND(D)$, respectively. $Des_C(X_i) \rightarrow Des_D(Y_j)$ is called the decision rule in *SDT*, where $Des_C(X_i)$ and $Des_D(Y_j)$ are unique descriptions of the blocks X_i (i = 1, 2, ..., n) and Y_j (j = 1, 2, ..., m), respectively. The set of decision rules (r_{ij}) for each block Y_j can be defined as

$$\{r_{i,j}\} = \{Des_C(X_i) \rightarrow Des_D(Y_j) : X_i \cap Y_j \neq \emptyset\}. \qquad (5)$$

A decision rule r_{ij} is deterministic if and only if $Y_j \cap X_i = X_i$, and r_{ij} is non-deterministic otherwise.

The certainty factor and coverage factor of decision rule r_{ij} are defined respectively as

$$\alpha_{X_i}(Y_j) = \frac{|Y_j \cap X_i|}{|X_i|}, \qquad (6)$$

$$\kappa_{X_i}(Y_j) = \frac{|Y_j \cap X_i|}{|Y_j|}. \qquad (7)$$

It is notable that $\alpha_{X_i}(Y_j)$ measures the degree of sufficiency of a proposition, $Des_C(X_i) \rightarrow Des_D(Y_j)$, and that $\kappa_{X_i}(Y_j)$ measures the degree of its necessity.

Proposition 5. Let $IT = (U, A)$ be an information table and $P, Q \subseteq A$, then one has that $IND(P) \cap IND(Q) = IND(P \cup Q)$.

Proof. Let $IND(P) = \bigcap_{a \in P} IND(\{a\})$ and $IND(Q) = \bigcap_{a \in Q} IND(\{a\})$, then we find that

$$\begin{aligned}
&IND(P) \cap IND(Q)\\
&= (\textstyle\bigcap_{a \in P} IND(\{a\})) \cap (\bigcap_{a \in Q} IND(\{a\}))\\
&= (\textstyle\bigcap_{a \in P - P\cap Q} IND(\{a\})) \cap (\bigcap_{a \in P\cap Q} IND(\{a\})) \cap\\
&\quad (\textstyle\bigcap_{a \in Q - P\cap Q} IND(\{a\})) \cap (\bigcap_{a \in P\cap Q} IND(\{a\}))\\
&= (\textstyle\bigcap_{a \in P - P\cap Q} IND(\{a\})) \cap (\bigcap_{a \in Q - P\cap Q} IND(\{a\})) \cap\\
&\quad (\textstyle\bigcap_{a \in P\cap Q} IND(\{a\}))\\
&= \textstyle\bigcap_{a \in (P - P\cap Q)\cup(Q - P\cap Q)\cup(P\cap Q)} IND(\{a\})\\
&= \textstyle\bigcap_{a \in P\cup Q} IND(\{a\})\\
&= IND(P \cup Q).
\end{aligned}$$

Hence the proposition holds.

Proposition 6. Let U/P and U/Q be two classifications with the respective indistinguishable relations *P* and *Q* on *U*. The intersection "∩" between two classifications U/P and U/Q is denoted as follows: $U/P \cap U/Q = U/(P \cup Q) = U/(Q \cup P)$ (also called classification U/P AND U/Q).

Proof. It is straightforward from Proposition 5.

Thus, from Proposition 6, it is easy to obtain the following property in a simplified decision table.

Property 1. Let $SDT = (U', C, D)$ be a simplified decision table and $P \subseteq C$, $U'/P = \{X_1, X_2, \ldots, X_n\}$. Then for any $a \in C - P$, one has that $U'/(P \cup \{a\}) = U'/P \cap U'/\{a\} = X_1/\{a\} \cup X_2/\{a\} \cup \ldots \cup X_n/\{a\} = \cup \{X_i/\{a\} : i = 1, 2, \ldots, n\}$, i.e. $U'/(P \cup \{a\}) = \cup \{X/\{a\} : X \in U'/P\}$.

After arbitrary $X_i \in U'/P$ is further partitioned on *a*, and $U' = U' - X_i$, the searched space of the *SDT* is gradually reduced. Therefore, the several simplifications above will be helpful to increase computational efficiency. Thus, using the idea of simplifications, we define the concept of decision degree as follows.

Definition 2. Let $SDT = (U', C, D)$ be a simplified decision table and $P \subseteq C$, $U'/P = \{X_1, X_2, \ldots, X_n\}$, $U'/D = \{Y_1, Y_2, \ldots, Y_m\}$.Then, $S(D|P)$ denotes the decision degree of *D* with reference to *P* as follows

$$\begin{aligned}
S(D \mid P) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{|Y_j \cap X_i|}{|X_i|} \times \frac{|Y_j \cap X_i|}{|Y_j|}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{X_i}(Y_j)\kappa_{X_i}(Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{|Y_j \cap X_i|^2}{|X_i||Y_j|}\right).
\end{aligned} \qquad (8)$$

From the formulas (5)-(7), we know that $S(D|P)$ includes the degree of sufficiency of a proposition and the degree of its necessity.

Thus, from the formula (8), we easily obtain the following properties in a simplified decision table.

Property 2. $0 < S(D|P) \leq |U'|$.

Property 3. If $U'/P = \omega$ and $U'/D = \omega$, then the decision degree of D with reference to P achieves its maximum value $|U'|$.

Property 4. If $U'/P = \omega$ and $U'/D = \delta$, then the decision degree of D with reference to P achieves its minimum value 1.

Property 5. If $U'/P = \delta$ and $U'/D = \delta$, then the decision degree of D with reference to P achieves its minimum value 1.

Proposition 7. Let $SDT = (U', C, D)$ be a simplified decision table and $A_1, A_2 \subseteq C$, $U'/A_1 = \{X_1, X_2, \ldots, X_n\}$, $U'/D = \{Y_1, Y_2, \ldots, Y_m\}$. Assume that $U'/A_2 = \{X_1, X_2, \ldots, X_{p\text{-}1}, X_{p+1}, \ldots, X_{q\text{-}1}, X_{q+1}, \ldots, X_n, X_p \bigcup X_q\}$ is another partition generated through combining equivalence blocks X_p and X_q to $X_p \bigcup X_q$, where X_p and X_q are two equivalence blocks randomly selected from U'/A_1,

$$S(D \mid A_1) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{|Y_j \cap X_i|^2}{|X_i||Y_j|}\right), \text{ and}$$

$$S(D \mid A_2) = S(D \mid A_1) - \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_p|^2}{|X_p||Y_j|}\right) - \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_q|^2}{|X_q||Y_j|}\right) + \sum_{j=1}^{m}\left(\frac{|Y_j \cap (X_p \bigcup X_q)|^2}{|X_p \bigcup X_q||Y_j|}\right).$$

Then one has that $S(D|A_1) \geq S(D|A_2)$.

Proof. From Definition 2, we find that

$$\begin{aligned} S_\Delta &= S(D \mid A_1) - S(D \mid A_2) \\ &= \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_p|^2}{|X_p||Y_j|}\right) + \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_q|^2}{|X_q||Y_j|}\right) - \sum_{j=1}^{m}\left(\frac{|Y_j \cap (X_p \bigcup X_q)|^2}{|X_p \bigcup X_q||Y_j|}\right) \\ &= \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_p|^2}{|X_p||Y_j|}\right) + \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_q|^2}{|X_q||Y_j|}\right) - \sum_{j=1}^{m}\left(\frac{|(Y_j \cap X_p) \bigcup (Y_j \cap X_q)|^2}{|X_p \bigcup X_q||Y_j|}\right) \\ &= \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_p|^2}{|X_p||Y_j|}\right) + \sum_{j=1}^{m}\left(\frac{|Y_j \cap X_q|^2}{|X_q||Y_j|}\right) - \sum_{j=1}^{m}\left(\frac{(|Y_j \cap X_p| + |Y_j \cap X_q|)^2}{(|X_p| + |X_q|)|Y_j|}\right). \end{aligned}$$

Let $|X_p| = x$, $|X_q| = y$, $|Y_j \cap X_p| = ax$, and $|Y_j \cap X_q| = by$. There must be $x > 0$, $y > 0$, $0 \leq a \leq 1$, and $0 \leq b \leq 1$ such that one has that

$$\begin{aligned} S_\Delta &= S(D \mid A_1) - S(D \mid A_2) \\ &= \sum_{j=1}^{m}\left(\frac{a^2x^2}{x|Y_j|} + \frac{b^2y^2}{y|Y_j|} - \frac{(ax+by)^2}{(x+y)|Y_j|}\right) \\ &= \sum_{j=1}^{m}\left(\frac{xy(a-b)^2}{(x+y)|Y_j|}\right). \end{aligned}$$

Assume that a function $f_j = \frac{xy(a-b)^2}{(x+y)|Y_j|}$, for any j = 1, 2, …, m. Above all, if $a = b$, i.e. $\frac{|Y_j \cap X_p|}{|X_p|} = \frac{|Y_j \cap X_q|}{|X_q|}$ then one has that $S_\Delta = S(D \mid A_1) - S(D \mid A_2)$ = 0. If $a \neq b$, then one has that $S_\Delta = S(D \mid A_1) - S(D \mid A_2)$ > 0. Hence, $S(D|A_1) \geq S(D|A_2)$ and the proposition holds.

Proposition 7 states that X_p and X_q are randomly selected to construct equivalence block $X_p \bigcup X_q$ combined from U'/A_1, so as to obtain a new partition U'/A_2. Furthermore, in the practical operation, there must exist a good many equivalence blocks, which also need to be combined. Therefore, the coalition of some equivalence blocks is regarded as an automatic, repeatable combination of two equivalence blocks. Thus, the new partition U'/A_2 of U', induced by the equivalent relation A_2, is coarser than U'/A_1.

According to Proposition 7, we know that the combination of blocks induced by condition attributes will decrease the decision degree monotonously, and the decision degree will remain unchanged only if the memberships of the combined blocks for all decision blocks are the same. Thus, the memberships of all equivalence blocks, induced by condition attributes for all decision blocks, will remain unchanged after the combination.

Proposition 8. Let $SDT = (U', C, D)$ be a simplified decision table and for any $Q \subset P \subseteq C$, then $S(D|P) > S(D|Q)$ holds.

Proof. Let U' be a given simplified universe and $U'/P = \{X_1, X_2, \ldots, X_n\}$, $U'/Q = \{Z_1, Z_2, \ldots, Z_k\}$. Since $Q \subset P$, it follows from Proposition 1 that $P \prec Q$ and $n > k$, and then there exists a partition $\{I_1, I_2, \ldots, I_k\}$ of $\{1, 2, \ldots, n\}$ such that $Z_i = \bigcup \{X_j : j \in I_i, i = 1, 2, \ldots, k\}$. Thus, from Proposition 7, we easily find that the decision degree generated by $Z_i = \bigcup \{X_j : j \in I_i, i = 1, 2, \ldots, k\}$ is less than that generated by X_j. Hence, $S(D|P) > S(D|Q)$ and the proposition holds.

Proposition 8 states that in a simplified decision table decision degree of knowledge monotonically increases as the granularity of knowledge, produced by the equivalence relation, becomes small through finer classification with the increase of attributes in knowledge.

Proposition 9. Let $SDT = (U', C, D)$ be a simplified decision table and for any $P \subseteq C$, then $S(D|P) \leq S(D|C)$ holds.

Proof. The proof is similar to that of Proposition 8.

Definition 3. Let $SDT = (U', C, D)$ be a simplified decision table. An attribute a in P is said to be relatively dispensable (relatively reducible or relatively superfluous) for D in $P \subseteq C$, if $S(D|P) = S(D|P - \{a\})$, and relatively indispensable in P otherwise. For any $P \subseteq C$, if each attribute in P is relatively indispensable in P, then P is called relatively orthogonal to D.

Let $SDT = (U', C, D)$ be a simplified decision table and $P \subset C$, $U'/P = \{X_1, X_2, \ldots, X_n\}$, $U'/D = \{Y_1, Y_2, \ldots,$

$Y_m\}$, then the conditional information entropy of D with reference to P is defined as

$$H(D \mid P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log p(Y_j \mid X_i)\,, \quad (9)$$

where $p(X_i) = \frac{|X_i|}{|U'|}$, $p(Y_j \mid X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$, i = 1, 2, …, n, and j = 1, 2, …, m.

Proposition 10. Let $SDT = (U', C, D)$ be a simplified decision table and for any $A_2 \subseteq A_1 \subseteq C$. Then one has that $H(D|A_1) \leq H(D|A_2)$. The necessary and sufficient condition of equation is that for any X_i, $X_j \in U'/A_1$, $X_i \neq X_j$, if $X_i \cup X_j \subseteq U'/A_2$, then $\frac{|X_i \cap D_r|}{|X_i|} = \frac{|X_j \cap D_r|}{|X_j|}$ always holds, where $D_r \in U'/D$.

Proof. The proof is similar to that of Lemma 4.1 in [7].

Proposition 11. Let $SDT = (U', C, D)$ be a simplified decision table and for any $P \subseteq C$. Then $H(D|P) \geq H(D|C)$ if and only if $S(D|P) \leq S(D|C)$.

Proof. It is straightforward from Proposition 9 and 10.

Hence, it is easy to obtain the following properties from Proposition 11.

Property 6. Let U' be a given simplified universe and for any $a \in P \subseteq C$. If $S(D|P) = S(D|P-\{a\})$, then $POS_P(D) = POS_{P-\{a\}}(D)$.

Property 7. Let U' be a given simplified universe and for any $a \in P \subseteq C$. If $H(D|P) = H(D|P-\{a\})$, then $POS_P(D) = POS_{P-\{a\}}(D)$.

Property 8. Let U' be a given simplified universe and for any $a \in P \subseteq C$, then $S(D|P) = S(D|P-\{a\}) \Leftrightarrow H(D|P) = H(D|P-\{a\})$.

For a general information table, the definition of reducts in the algebra view is equivalent to its definition in the information view [7]. Thus, it is stated from these representations above, we find that the definition of the relative reducts of a consistent decision table (i.e., there are no conflicts or inconsistent objects in decision tables) in the algebra view is also equivalent to its definition in the information view. However, inconsistent decision tables often occur in real life. We need to calculate the reducts of inconsistent decision tables. Then, the decision degree of a relative reduct defined above does not remain unchanged. On the other hand, according to Proposition 11 and Property 6-8, if an attribute cannot provide any additional information for an existing attribute set to make a decision table, then it is reducible. That is, the definition of relative reducts of a decision table in the information view includes its definition in the algebra view. Any relative reduct of a decision table in the decision degree view must be its relative reduct in the algebra view. Furthermore, for decision degree of knowledge in decision tables, it not only decreases the complexity of arithmetic, especially in the logarithmic calculation, and saves the accounting time, but also is easy to program and calculate simply. Therefore, this paper focuses on creating such a heuristic algorithm developed further on this result.

IV. Decision Tree Approach to Rule Extraction

In this section, based on decision degree, we propose a new significance measure of attribute. Then, as an application of decision tree, we apply this measure to decision rule extraction in decision tables.

A. Representation of improved significance measure

It is known that decision tree is one of the most effective solutions to classification tasks [18-24]. This technique sets up classification models by constructing trees. How to construct a robust and efficient tree is still a challenge met in the research and application of decision tree.

Generally speaking, decision tree can be compared and evaluated according to predictive accuracy, speed, robustness, scalability and interpretability. It has been approved that ID3 and C4.5, introduced by Quinlan [23], satisfies with the above criteria. It also provides high predictive accuracy and efficiency among the compared main memory algorithms for classification. It is affirmed by experts in data mining and is always taken as the benchmark for researches of classifiers. Among decision tree with univariant splits, ID3 and C4.5 have the advantages in predictive accuracy and speed, however, they have the disadvantage in constructing a smaller tree, because ID3 and C4.5 have connatural limitations, such as the problem of empty branches and over-fitting.

To solve the problems, a new uncertainty measure as heuristic information is introduced, and this measure improves on attribute selection and partition methods. Such that the improved approach can reflect not only the importance of the entire training sample set, but also that of the related training sample sets of branch nodes. It combines branches, which have high attribute significance in divide and conquer process. According to the definition of attribute significance, the higher it is, the lower the purity of partition is. The significance measure of attribute considers the branches contributing nothing to classification as one branch, so it can reduce insignificant branches, avoid the problem of fragmentation, control the size of trees, and have high predictive accuracy. Therefore, the significance measure of attribute avoids automatically the tendency to select attributes with more values as test attributes.

Proposition 12. Let $SDT = (U', C, D)$ be a simplified decision table and $P \subseteq C$, $U'/P = \{X_1, X_2, \ldots, X_n\}$, $U'/D = \{Y_1, Y_2, \ldots, Y_m\}$, then $U'/(P \cup D) = U'/P \cap U'/D = \{Z_1, Z_2, \ldots, Z_k\}$. That is, for any $Z_t \in U'/(P \cup D)$, there exist $X_i \in U'/P$ and $Y_j \in U'/D$ such that $Z_t = X_i \cap Y_j$.

Proof. It is straightforward from Proposition 6 and Property 1.

Definition 4. Let $SDT = (U', C, D)$ be a simplified decision table, where U' is called the whole training sample sets. If $P \subseteq C$, for any testing attribute $a \in C - P$, $U'/\{a\} = \{U'_1, U'_2, \ldots, U'_t\}$, and $U'/(\{a\} \cup D) = \{Z_1, Z_2, \ldots, Z_k\}$, then the significance measure of a on U'_i with reference to D is defined as

$$Sig(a, P, U'_i, D) = \frac{S(D \mid P \cup \{a\}) - S(D \mid P)}{k}, \quad (10)$$

where $U'_i \in U'/\{a\}$ ($i = 1, 2, \ldots, t$) is a related training sample set on a from the branch node of sub-tree. For example, if the testing attribute a is the root node, its related training sample set is U'. k is the number of equivalence blocks from the partition $\{Z_1, Z_2, \ldots, Z_k\}$, and it shows the amount of decision rules created by the root or sub-root node a.

It shows from Definition 4 that the significance measure $Sig(a, P, U'_i, D)$ indicates the importance of attribute a added to P with reference to D in decision tables, offering the powerful reference to the decision. The bigger the significance measure of attribute is, the higher its position in the decision table is, and otherwise the lower its position is. Therefore, if each of the significance measure of attribute is calculated, then the attribute with the zero or lower significance measure is removed, and the knowledge reduction can be finished.

In the process of calculating $Sig(a, P, U'_i, D)$, it can be easily seen that every time to calculate any testing attribute a with the maximum of $Sig(a, P, U'_i, D)$ is in fact to calculate $\frac{S(D \mid P \cup \{a\})}{k}$ with the maximum. Because $S(D|P)$ is a constant when we calculate $Sig(a, P, U'_i, D)$. Meanwhile, to improve the limitations of ID3 and C4.5, when the branch nodes of sub-tree are generated by the testing attributes, all of their parent nodes and the corresponding attributes are no longer involved with the nodes of their sub-trees.

Therefore, the factors, which include their parent nodes, the corresponding attribute set P, and the related training sample sets, should not be considered. Thus, to calculate $\frac{S(D \mid P \cup \{a\})}{k}$ is in fact to calculate the corresponding $\frac{S(D \mid \{a\})}{k}$, that is, to calculate the corresponding equivalence blocks. When we introduce the technology of radix sorting in [26] to calculate equivalence blocks effectively, thus, all of the policies above will be helpful to reduce the quantity of computation and the time-space complexity of search.

B. Decision Tree-based Rule Extraction Algorithm

Towards dealing with both the above some difficulties of rule extraction, based on the improved significance measure of attribute, our idea is to put a method without computing relative attribute reduction of a decision table in rough set theory into a decision-tree structure. Motivated by this, based on decision degree, we propose the method over depth-first spanning for rule extraction from decision tree. The following is a summary of the steps of decision tree-based rule extraction algorithm, which consists of classification decision tree construction, and followed by three steps of rule extraction over depth-first spanning, including rule extraction over the tree structure, the one-class node, and deleting dispensable nodes from rule, respectively. Individual steps of the rule extraction are detailed in the subsections below and shown schematically in Fig. 1.

It is well known that rough set theory is always used to mine some patterns in the form of “if …, then …” decision rules from decision tables. More exactly, the decision rules say that if some condition attributes have given values, then some decision attributes have other given values. Then, we will not only consider how to discretize numerical attributes and construct a decision tree for rule extraction, but also focus on how to improve computational efficiency in the context of large-scale data sets. The rule extraction algorithm using decision tree, called RE-DT, is designed in a decision table, its time complexity is analyzed, and an illustrative example is employed to show the mechanism of algorithm RE-DT as follows.

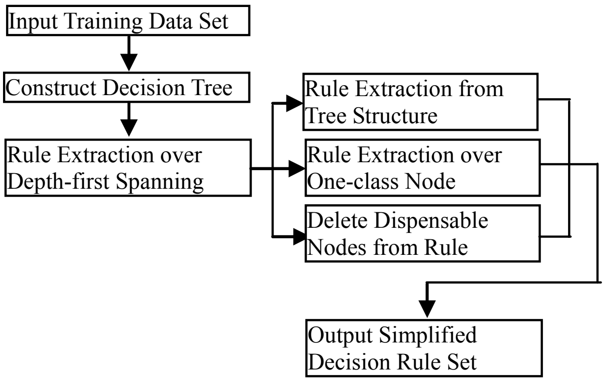
Input Training Data Set
Construct Decision Tree
Rule Extraction over Depth-first Spanning
Rule Extraction from Tree Structure
Rule Extraction over One-class Node
Delete Dispensable Nodes from Rule
Output Simplified Decision Rule Set

Figure 1. Steps of decision tree-based rule extraction algorithm.

Algorithm RE-DT

```
Input: $U$ (Training sample set), $C$ (Condition attribute set), $D$ (Decision attribute set).
Output: $T$ (a decision tree), decision rule set.
(1) Initialize $T$ as the root node, and current training sample set $U$.
(2) Perform $U/(C \cup D)$ with radix sorting, and obtain related training sample set $U'$.
(3) If $U'/D = \{U'\}$, then append $V_d$ to $T$.
(4) If $|C| = 1$, then append $U'_i \in U'/C$ to $T$ as terminal nodes, depth-first span path (rule) from terminal node to root down-up (or root to terminal node up-down), and start the following loop operations:
    (4.1) If the current node of $T$ is an internal node, then delete it.
    (4.2) If the path is only, then go to the next node, and return to (4.1).
    (4.3) Otherwise, add it to the path, go to the next node, and return to (4.1).
    (4.4) Mark the path as a simplified decision rule.
(5) Perform Eq. (8) to obtain $\frac{S(D \mid \{a\})}{k}$ ($a \in C$, $k = |U'_i/(\{a\} \cup D)|$, and $U'_i \in U'/\{a\} = \{U'_1, U'_2, \ldots, U'_t\}$) with radix sorting, and select the testing attribute $a$ with the maximum $\frac{S(D \mid \{a\})}{k}$ as the branch node.
    (5.1) If the testing attribute is not only, then select one with the minimum $k$.
    (5.2) If the selected is not only, then select the front.
(6) For every related training sample set $U'_i \in U'/\{a\}$, do the following operations:
    (6.1) If $U'_i$ belongs to one-class, then train a one-
```

class node to T, and append V_a to the corresponding arcs.

(6.2) Set the sub-tree T_i as (U'_i, $C = C - \{a\}$, D), where $i = 1, 2, \ldots, t$.

(7) If no new node is added to T, then output T and simplified decision rule set, otherwise, adjust current partitioning scale, and go to (4).

Obviously, the generation of decision rules is not based on a reduct of a decision table but the equivalence blocks and depth-first traversal in Algorithm RE-DT. Furthermore, in order to extract decision rules efficiently, we use the technology of twice-hash [26] to calculate all of the equivalence blocks, whose time complexity will be cut down to $O(|U|)$.

By using algorithm RE-DT, the time complexity to extract decision rules from a decision table is polynomial.

At step (2), the time complexity of computing the partition $U/(C \bigcup D)$ is $O(|U'|)$.

At step (4), since $|C| - 1$ is the maximum value for the circle times, the time complexity for depth-first traversal is

$$O((|C|-1)|U'_1|) + O((|C|-2)|U'_2|) + \ldots + O(|U'_{|U'/C|}|) \leq O(|C|^2\,|U'|).$$

At step (5), since $|C|$ is the maximum value for the circle times, the time complexity of selecting the testing attribute is

$$O(|C||U'|) + O((|C|-1)|U'|) + O((|C|-2)|U'|) + \ldots + O(|U'|) = O(\frac{|C||C+1|}{2}|U'|).$$

At step (6), the time complexity is also $O(|U'|)$.

Thus, in the worst case the time complexity of Algorithm RE-DT is

$$O(|U'|) + O(|C|^2|U'|) + O(\frac{|C||C+1|}{2}|U'|) =$$

$$O(\frac{3|C|^2+|C|+2}{2}|U'|) = O(|C|^2|U'|).$$

Therefore, Algorithm RE-DT is different from existing algorithms based on attribute reduct for extracting decision rules from decision tables. The time complexity of a rule-extracting algorithm based on attribute reduct is $O(|C|^3|U|^2)$. Obviously, the time complexity of algorithm RE-DT is much lower than those of the existing algorithms in [8, 11, 12, 15, 22-25, 27]. Furthermore, its worst space complexity is $O(|C||U'|)$.

Thus, the time complexity of algorithm RE-DT is largely reduced relative to those of the existing algorithms based on attribute reduct for extracting decision rules from decision tables. Therefore, this means that the proposed algorithm for finding the simplified decision rule set requires less computation and memory.

Example 1. $SDT = (U', C, D)$ is a simplified decision table in [27], described in Table I, where $U' = \{1, 2, \ldots, 14\}$, $C = \{a, b, c, d\}$.

From Table I, by using algorithm RE-DT, we construct a decision tree, shown in Fig. 2, and extract its corresponding simplified decision rules.

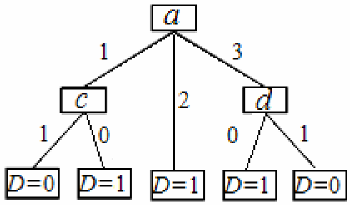

Figure 2. A decision tree.

By computing, it follows that $U'/D = \{\{1, 2, 6, 8, 14\}, \{3, 4, 5, 7, 9, 10, 11, 12, 13\}\} = \{D_0, D_1\}$. Therefore, five deterministic decision rules can be extracted as follows:

$Des_C(\{1, 2, 8\}) \rightarrow Des_D(D_0)$, i.e., $(a = 1, c = 1) \rightarrow (D = 0)$;
$Des_C(\{9, 11\}) \rightarrow Des_D(D_1)$, i.e., $(a = 1, c = 0) \rightarrow (D = 1)$;
$Des_C(\{3, 7, 12, 13\}) \rightarrow Des_D(D_1)$, i.e., $(a = 2) \rightarrow (D = 1)$;
$Des_C(\{4, 5, 10\}) \rightarrow Des_D(D_1)$, i.e., $(a = 3, d = 0) \rightarrow (D = 1)$;
$Des_C(\{6, 14\}) \rightarrow Des_D(D_0)$, i.e., $(a = 3, d = 1) \rightarrow (D = 0)$.

V. Experimental Methods and Results

In this section, we give a real world example used to explain the validity of the proposed rule extraction algorithm, and apply the proposed approach and other decision tree approaches to several data sets from the UCI Machine Learning Repository (http://www.ics.uci.edu), so as to evaluate the proposed approach.

Example 2. Consider descriptions of the career work in university. This is a decision table, described in Table II, where $U = \{1, 2, ..., 24\}$. The condition attributes A, B, C, and D stand for Political Landscape, Major Score, Practical Ability, and English Proficiency, respectively. E stands for Job Hunting. The values 1, 2, and 3 of Political Landscape stand for People, Member, and Party Member, respectively. The values 1 and 2 of other attributes stand for General and Good, respectively.

TABLE I.
A Simplified Decision Table

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a	1	1	2	3	3	3	2	1	1	3	1	2	2	3
b	1	1	1	2	3	3	3	2	3	2	2	2	1	2
c	1	1	1	1	0	0	0	1	0	0	0	1	0	1
d	0	1	0	0	0	1	1	0	0	0	1	1	0	1
D	0	0	1	1	1	0	1	0	1	1	1	1	1	0

TABLE II.
DECISION TABLE DESCRIPTIONS OF THE CAREER WORK IN UNIVERSITY

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3
B	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
C	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
D	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
E	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	2	2	2	2	1	2	1	2	2

From Table II, based on the above measure of significance, we can construct two decision trees, based on decision degree and ID3, shown in Fig. 3 and 4, respectively. It can be seen from Fig. 3 and 4 that nine decision rules are generated from each of the two decision trees. In Fig. 3, *B* and *C*, earlier than *A*, are selected. In Fig. 4, *D* is firstly selected, following that, *A* is selected secondly. However, through analyzing knowledge-based systems of employment expert, it is known that the significance of *B* and *C* is greater than that of *A* with more values. Therefore, the proposed algorithm avoids automatically the tendency to select attributes with more values as test attributes.

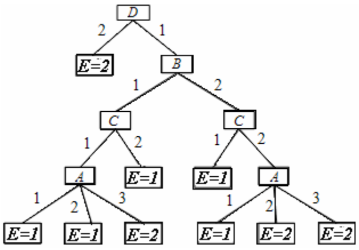

Figure 3. Decision degree-based decision tree.

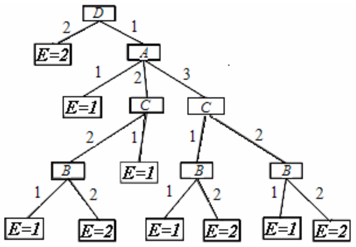

Figure 4. ID3-based decision tree.

In the following, we run our experiments on four data sets from the UCI Machine Learning Repository [28], which represent a wide range of domains and data characteristics. The description of the four data sets is shown in Table III. We use C 4.5-based algorithm in [23], rough set-based approach in [25], and Algorithm RE-DT to design C4.5, RS, and RE-DT, respectively. Then, we conduct empirical experiments to compare in terms of number of rules, running time and classification accuracy. In all experiments, 10-fold cross validation is conducted on all data sets to calculate the classification accuracy of the three methods. That is to say, we can obtain ten results with respect to each data set. Our experiments are performed on AMD Quad-core 3.1 GHz CPU, 4GB RAM, Windows XP, and program in Visual C++ 6.0. The detailed experimental results are summarized in Table IV. In the table, "NR" indicates the average number of rules, the running time of each algorithm, indicated by "RT", is the averaged CPU time in millisecond, and "Accu" indicates the average accuracy of the ten values of accuracy with respect to a data set and is assigned as the accuracy of the corresponding decision tree. Furthermore, this difference can be illustrated by plotting the ratios of their running time, C4.5/RS and C4.5/RE-DT, shown in Fig. 5.

According to Fig. 5, it is observed that the slope of the curve shown tends to increase with size of data sets, so that just shows the proposed method are thus much more suitable for large data sets, but the curve fluctuates distinctly. In fact, the main reason is that the attribute number of data sets is different.

From all of the experimental results, we can see that the decision trees constructed by the presented method tend to have simplified decision rule set, smaller running time and higher classification accuracy than that constructed by the methods in [23, 25] on most of the four datasets. It should be mentioned here that the number of condition attributes and the number of values one attribute can have can vary arbitrarily without limit, the tendency of all results is similar.

TABLE III.
DESCRIPTION OF DATA SETS

No.	Data sets	Size	Attributes (*C*/*D*)	Classes
1	Iris	150	4/1	3
2	Soybean	683	35/1	19
3	Breast Cancer Wisconsin	699	9/1	2
4	Tic-tac-toe	958	9/1	2

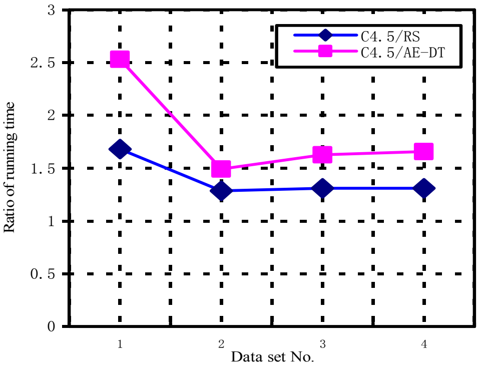

Figure 5. Ratio of the running time (C4.5/RS, C4.5/RE-DT).

TABLE IV.
PERFORMANCE COMPARISONS OF C4.5, RS AND RE-DT

No.	Data sets	C4.5			RS			RE-DT		
		NR	RT	Accu	NR	RT	Accu	NR	RT	Accu
1	Iris	4.8	810	94.73	2.1	482	95.53	2.1	320	96.26
2	Soybean	48.2	3100	91.78	41.5	2410	94.54	38.4	2080	95.12
3	Breast cancer Wisconsin	21.6	3010	95.01	20.8	2300	96.08	18.9	1850	98.00
4	Tic-tac-toe	32.6	3610	84.02	31.5	2758	92.16	30.8	2180	95.70

VI. CONCLUSIONS AND FUTURE WORK

In recent years, the rough set-based approaches to inducing decision tree are testified to be a simplified and feasible way for constructing decision tree. These current approaches also have some disadvantages, however, they can do well only in accurate classification where objects are strictly classified according to equivalence blocks. Therefore, the induced classifiers lack the ability to tolerate possible noises in real world data sets. This is an important problem to be handled in applications. In this paper, to offset and improve the limitations of current knowledge acquisition approaches, we introduce a new uncertainty measure, called decision degree, and a new significance measure for rules extraction of decision tree is designed. As an application of rough set, an algorithm, called RE-DT, has been proposed for extracting decision rules in decision tables. Two illustrative examples and four data sets from the UCI Machine Learning Repository have been employed to show the validity of this algorithm. Moreover, the time complexity of algorithm RE-DT is much lower than that of the existing approaches to decision tree. Thus, we believe that the use of more sophisticated methods, such as the accuracy estimates based on *k*-fold cross-validation or leave-one-out, could improve the performance of the proposed method and make its advantage stronger in efficiently extracting rules from practical large-scale data sets. This is the main research direction for our future work.

ACKNOWLEDGMENT

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

This work is supported by the National Natural Science Foundation of China under Grant (No. 60873104, No. 61040037), 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 (No. 12A520027), and the Education Fund for Youth Key Teachers of Henan Normal University.

REFERENCES

[1] A. Skowron, J. Stepaniuk, and R. Swiniarski, “Modeling Rough Granular Computing Based on Approximation Spaces”, *Information Sciences*, vol. 184, no.1, pp.20–43, 2012.

[2] 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, no.9, pp.1525–1532, 2011.

[3] L. Sun, J. C. Xu, and S. Q. Li, “Knowledge Reduction Based on Granular Computing from Decision Information Systems”, *Lecture Notes in Computer Science*, vol.6401, pp.46–53, 2010.

[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, no.12, pp.2259–2268, 2011.

[5] L. Sun, J. C. Xu, S. Q. Li, and Y. P. Gao, “Rough Entropy Extensions for Feature Selection under Incomplete Decision Information Systems”, *Advances in Information Sciences and Service Sciences*, vol.3, no.11, pp.264–274, 2011.

[6] J. C. Xu and L. Sun, “A New Knowledge Reduction Algorithm Based on Decision Power in Rough Set”, *Transactions on Rough Sets*, vol.XII, pp.76–89, 2010.

[7] G. Y. Wang, “Rough Reduction in Algebra View and Information View”, *International Journal of Intelligent Systems*, vol.18, no.6, pp.679–688, 2003.

[8] L. P. An and L. Y. Tong, “Binary Relations as A Basis for Rule Induction in Presence of Quantitative Attributes”, *Journal of Computers*, vol.5, no.3, pp. 440–447, 2010.

[9] 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, no.9, pp.129–139, 2011.

[10] R. Q. Chang, Z. Pei, and C. Zhang, “A Modified Editing K-Nearest Neighbor Rule”, *Journal of Computers*, vol.6, no.7, pp.1493–1500, 2011.

[11] Y. H. Qian, J. Y. Liang, and C. Y. Dang. “Converse Approximation and Rule Extraction from Decision Tables in Rough Set Theory”, *Computers and Mathematics with Applications*, vol.55, no.8, pp. 1754–1765, 2008.

[12] W. Ziarko, “Acquisition of Hierarchy-Structured Probabilistic Decision Tables and Rules from Data”, *Expert Systems*, vol.20, no.5, pp.305–310, 2003.

[13] J. S. Mi, Y. Leung, and W. Z. Wu, “Dependence-Space-Based Attribute Reduction in Consistent Decision Tables”, *Soft Computing*, vol.15, no.2, pp. 261−268, 2011.

[14] D. Y. Li, B. Zhang, and Y. Leung, “On Knowledge

Reduction in Inconsistent Decision Information Systems", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol.12, no.5, pp.651–672, 2004.

[15] S. Greco, Z. Pawlak, and R. Slowinski, "Can Bayesian Confirmation Measures be Useful for Rough Set Decision Rules?", *Engineering Applications of Artificial Intelligence*, vol.17, no.4, pp.345–361, 2004.

[16] U. Rückert, L. D. Raedt, "An Experimental Evaluation of Simplicity in Rule Learning", Artificial Intelligence, vol.172, no.1, pp.19–28, 2008.

[17] G. H. Chen, Y. Tang, Y. Pan, and Q. Deng, "Question Classification Using Multiple Kernel Learning and Semantic Information", *Journal of Computers*, vol.6, no.11, pp.2325–2334, 2011.

[18] S. N. Pang and N. Kasabov, "Encoding and Decoding the Knowledge of Association Rules over SVM Classification Trees", *Knowledge and Information Systems*, vol.19, no.1, pp.79–105, 2009.

[19] S. Badr and A. Bargiela, "Case Study of Inaccuracies in The Granulation of Decision Trees", *Soft Computing*, vol.15, no.6, pp.1129–1136, 2011.

[20] J. H. Zhai, "Fuzzy Decision Tree Based on Fuzzy-Rough Technique", *Soft Computing*, vol.15, no.6, pp.1087–1096, 2011.

[21] L. X. Jiang and C. Q. Li, "Scaling up the Accuracy of Decision Tree Classifiers: A Naïve Bayes Combination", *Journal of Computers*, vol.6, no.7, pp. 1325–1331, 2011.

[22] J. M. Wei, S. Q. Wang, M. Y. Wang, J. P. You, and D. Y. Liu, "Rough Set Based Approach for Inducing Decision Trees", *Knowledge-Based Systems*, vol.20, no.8, pp.695–702, 2007.

[23] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1993.

[24] Y. S. Huang and B. C. Huang, "Recognition of Multi–Interval Rules in Dataset with Continuous–Valued Attributes", *Expert Systems with Applications*, vol.36, no.2, pp.1485–1492, 2009.

[25] J. M. Wei, "Rough Set Based Approach to Selection of Node", *International Journal of Computational Cognition*, vol.1, no.2, pp.25–40, 2003.

[26] Y. Liu, R. Xiong, and J. Chu, "Quick Attribute Reduction Algorithm with Hash", *Chinese Journal of Computers*, vol.32, no.8, pp.1493–1499, 2009.

[27] G. Y. Wang, "Rough Set Theory and Knowledge Acquisition", Xi'an Jiaotong University Press, China, 2001.

[28] UCI Repository of Machine Learning Databases, 1998. Available: http: // ftp.ics.uci.edu/pub/machine-learning-data bases

Lin Sun was born in Nanyang, Henan, China, in 1979. He received his B.S. and M.S. degree in computer science and technology from Henan Normal University, Henan, China, in 2003, and 2007, respectively. He is a Ph.D. candidate at Beijing University of Technology.

His major is Pattern Recognition and Intelligent System. His main research interests include granular computing, rough set, intelligent information processing, web intelligent.

Jiucheng Xu was born in Luoyang, Henan, China, in 1964. He received his B.S. degree in mathematics from Henan Normal University, Henan, China, in 1986, the M.S. degree and the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Shanxi, China, in 1995, and 2004, respectively.

Currently, he is a professor in the College of Computer & Information Technology, Henan Normal University. His main research interests include granular computing, rough set, data mining, intelligent information processing, image retrieval.