International Journal of Computational Intelligence Systems, Vol. 4, No. 4 (June, 2011), 655-671

# Feature Selection in Decision Systems Based on Conditional Knowledge Granularity

**Tingquan Deng [1,2] , Chengdong Yang [2] , Qinghua Hu [3]**

[1] *College of Science, Harbin Engineering University, Harbin 150001, P.R. China*

*E-mail: Deng.tq@hrbeu.edu.cn*

[2] *College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, P.R. China*

*E-mail: yangchengdong2008@163.com*

[3] *School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, P.R. China*

*E-mail: huqinghua@hcms.hit.edu.cn*

Received 18 June 2010
Accepted 15 February 2011

**Abstract**

Feature selection is an important technique for dimension reduction in machine learning and pattern recognition communities. Feature evaluation functions play essential roles in constructing feature selection algorithms. This paper introduces a new notion of knowledge granularity, called conditional knowledge granularity, reflecting relationship between conditional attributes and decision attribute. An evaluation function to measure significance of conditional attributes is proposed and equivalent characterization of attribute reduction is established based on the conditional knowledge granularity. An optimal algorithm for feature selection is developed on the basis of the proposed evaluation function. Furthermore, a novel approach to performing feature selection in an inconsistent decision system is put forward through establishing a rough communication between the inconsistent decision system and a consistent decision system. Simulated experiments verifies feasibility and efficiency of the proposed technique.

*Keywords:* Conditional knowledge granularity, Rough sets, Attribute reduction, Feature selection, Rough communication.

## 1. Introduction

As a very popular mathematical tool to deal with incomplete, inexact and vague knowledge, rough set theory [29] was originated in 1982 and has attracted much attention from theory and application domains. Various extensions of rough sets, such as variable precision rough sets [54], rough fuzzy set [3], fuzzy rough sets [5,25,52], etc., have been developed to meet their applications in decision support systems [9,24], image processing [6,32], attribute reduction [43,47], feature selection [18], data mining [36], neural computing [53], conflict analysis [31], knowledge discovery [12,22], rule extracting [42], fault prognosis [2], information communication [44], classifier designs [15], and so on.

Among applications of rough sets, attribute reduction is one of the most essential and useful tech-

niques of data analysis and processing in machine learning [26], pattern recognition [11,41], and artificial intelligence [19]. Since redundant information usually covers a number of attributes or features in real world applications, which may confuse learning algorithms, cause distinguish slowdown in learning process and increase risk of learned classifiers to over-fit training data [50], removing superfluous or irrelevant features is necessary in classification modeling. The aim of attribute reduction focuses on solution to such a problem and obtains a compact data set preserving the same discrimination capability as the original one.

In rough set theory, there are two basic approaches to finding attribute reducts of an information system. One is based on discernibility matrices [39]. All reducts of attribute set can be obtained by means of this method. However, it was proved that finding all reducts or finding an optimal reduct (a reduct with the smallest number of attributes) is an NP-complete problem [1,46]. Another method is to achieve an approximate reduct where the selected attributes have higher significance degrees than the others by an appropriate optimal algorithm. This method is dynamic and is referred to as feature selection technique [38].

A key issue on feature selection lies in establishing an efficient criterion, called an evaluation function, to assess quality of attributes. The evaluation function based on dependency is the most widely used one [30]. It measures the proportion of the number of samples in positive regions of classification in the whole universe and the corresponding feature selection strategy has been successfully applied in many domains, such as fuzzy rough sets based feature selection [4,13,17,38], neighborhood rough sets based feature selection[14] and tolerance rough sets based feature selection [28]. However, this evaluation function is not always effective in finding a reduct of information system, see Example 1 for detailed illustration. It is, therefore, necessary to develop other evaluation functions to measure significance of attributes. Mutual information based evaluation function is such a classical uncertainty measure applied to feature selection.

Mutual information reflects relevance between conditional attributes and decision one. The induced evaluation function has been extensively applied to fuzzy rough sets based feature selection [37,48], variable precision rough sets based feature selection [8], dominance rough sets based feature selection [16], rough sets and Bayesian networks based feature selection [40], dynamic mutual information based feature selection [23], and normalized mutual information based feature selection [7]. However, mutual information of knowledge relies strongly on prior probability of knowledge that is unknown in information systems. In practice, the prior probability has to be replaced by posterior probability. Errors will be inevitably caused in feature selection and the classification performance will decrease in machine learning.

The concept of information granularity, initialed by L.A. Zadeh [51], regards discontinuous information as a mass of information granularity. It reflects the degree of legibility of information. The notion of granularity of knowledge or knowledge granularity, derived from information granularity, can describe classification performance of knowledge and has been extensively applied to feature selection in information tables and evaluation of decision performance [20,21,33,34,35]. However, it cannot be used directly to perform attribute reduction or feature selection in decision systems.

In order to extend the concept of knowledge granularity from information tables to decision information systems, a new notion of knowledge granularity, called conditional knowledge granularity, is introduced to characterize relationship between two attribute subsets. An evaluation function is developed based on the new notion to assess significance of features in consistent decision systems to perform attribute reduction and feature selection. Equivalent characterizations for attribute reduction are established in information systems as well as in inconsistent decision systems. An optimal algorithm of feature selection is designed according to the proposed evaluation function.

It is well-known that inconsistent decision systems usually appear in data mining and machine learning community. Unfortunately, the proposed method of feature selection cannot be directly used

in inconsistent decision systems. This paper designs a rough communication [44] between an inconsistent decision system and a consistent decision systems to realize the problem of feature selection in an inconsistent decision system. Such a mapping from an inconsistent decision system to a consistent decision system transforms the problem of finding a reduct of inconsistent decision system to that of finding a reduct of the induced consistent decision system.

The rest of this paper is organized as follows. Section 2 reviews some basic concepts on rough sets, as well as the method of feature selection based on dependency. In Section 3, the notion of conditional knowledge granularity is introduced in information systems. Its properties are wholly investigated. Section 4 concerns the issue on feature selection based on the proposed conditional knowledge granularity in consistent decision systems. An equivalent characterization of attribute reduction and an optimal algorithm for feature selection are established in decision information systems. Section 5 considers the problem of feature selection in inconsistent decision systems. A rough communication between an inconsistent decision system and a consistent decision system is designed to deal with this problem. Experiments and analysis are shown in Section 6 and conclusions follow in Section 7.

## 2. Preliminaries

We firstly recall some basic concepts on rough sets. An information system or an information table [49] is a four-tuple $S = (U, A, V, f)$ satisfying

(1) $U$ is a non-empty finite set of objects;
(2) $A$ is a non-empty finite set of attributes or features;
(3) $V = \cup_{a \in A} V_a$ and $V_a$ is the value set of attribute $a$; and
(4) $f$ is a mapping from $U \times A$ to $V$ such that, for any $x \in U$ and for any $a \in A$, $f(x,a) \in V_a \subseteq V$.

For any subset $B$ of attribute set $A$, an indiscernibility relation $R_B$ on $U$ is defined as

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

It is obvious that $R_B$ is an equivalence relation induced by $B$ and we denote by $[x]_B$ the equivalence class of $x$ with respect to $R_B$. $U/R_B = \{[x]_B \mid x \in U\}$ is therefore a partition of $U$ induced by $R_B$, denoted by $U/B$ if there is no confusion arisen. Each equivalence relation $R_B$, or alternatively, the subset $B$, is called a piece of knowledge in $S$.

For a subset $X \subseteq U$, the lower approximation and the upper approximation of $X$ with respect to knowledge $B$ are, respectively, defined by

$$\begin{aligned} \underline{R_B}(X) &= \{x \in U \mid [x]_B \subseteq X\} \\ \overline{R_B}(X) &= \{x \in U \mid [x]_B \cap X \neq \emptyset\} \end{aligned}$$

$\underline{R_B}(X)$, denoted also by $Pos_B(X)$, is called the positive region of $X$ with respect to $B$, whereas $U \setminus \overline{R_B}(X)$ is the negative region, where $X \setminus Y$ denotes the set minus of $X$ by $Y$. If $\underline{R_B}(X) \neq \overline{R_B}(X)$, $X$ is referred to be a rough set in $S = (U, A, V, f)$ and it can be represented generally by the pair $(\underline{R_B}(X), \overline{R_B}(X))$. $BN_B(X) = \overline{R_B}(X) \setminus \underline{R_B}(X)$ is referred to as rough boundary set of $X$ with respect to $B$. The objects in $\underline{R_B}(X)$ can be totally perceived and precisely classified by $B$, but the objects in $BN_B(X)$ cannot.

**Definition 1.** Let $S = (U, A, V, f)$ be an information system. An attribute $a \in A$ is said to be dispensable in $A$ if $Pos_{A\setminus\{a\}}(X) = Pos_A(X)$ for all $X \subseteq U$, whereas $a \in A$ is called not dispensable or independent in $A$ if $Pos_{A\setminus\{a\}}(X) \neq Pos_A(X)$ for some $X \subseteq U$.

In data mining, one of main purposes is to remove all dispensable attributes and to generate a reduct of information system, which preserves the same classification performance for objects as the original one. In mathematics the definition of reduct can be presented as follows.

**Definition 2.** [29] Let $S = (U, A, V, f)$ be an information system. A subset $B \subseteq A$ is called a reduct of $A$ in $S$ if for all $X \subseteq U$,

(1) $Pos_B(X) = Pos_A(X)$;
(2) $Pos_{B\setminus\{a\}}(X) \neq Pos_B(X)$ for any $a \in B$.

It is verified that in classical rough set theory the equivalent statement that $R_B = R_A$ if and only if $Pos_B(X) = Pos_A(X)$ for all $X \subseteq U$ is true. With this

consequence a simplified characterization for reduct can be reached.

**Proposition 1.** *Let $S=(U,A,V,f)$ be an information system. A subset $B\subseteq A$ is a reduct of $A$ in $S$ if and only if*

(1) $R_B=R_A$;
(2) $R_{B\setminus\{a\}}\neq R_B$ *for any* $a\in B$.

An information system $S=(U,A,V,f)$ is referred to be a decision information system, or a decision system (decision table), when $A$ is divided into two nonempty subsets $C$ and $D$, called the conditional attribute set and the decision attribute set, respectively.

In a decision system $S=(U,C\cup D,V,f)$, the degree of dependency of $D$ on $C$ is defined by

$$\gamma(D|C)=\frac{|POS_C(D)|}{|U|}$$

where $POS_C(D)=\cup_{D_i\in U/D}\underline{R_C}(D_i)$ is the positive region of $D$ with respect to $C$ and $|X|$ denotes the cardinality of set $X$. $S$ is called a consistent decision system if $\gamma(D|C)=1$, e.g., $D$ depends totally on $C$, whereas $S$ is an inconsistent decision system if $\gamma(D|C)<1$, e.g., $D$ depends partially on $C$.

**Proposition 2.** *Let $S=(U,C\cup D,V,f)$ be a decision system and $B\subseteq C$. Then $\gamma(D|B)=\gamma(D|C)$ if and only if $Pos_B(D)=Pos_C(D)$.*

**Proof** Let $B\subseteq C$, then $[x]_C\subseteq[x]_B$ for all $x\in U$, and so $Pos_B(D)\subseteq Pos_C(D)$.

$\Rightarrow$: If $\gamma(D|B)=\gamma(D|C)$, one has $|Pos_B(D)|=|Pos_C(D)|$. It is natural that the equation $Pos_B(D)=Pos_C(D)$ holds.

$\Leftarrow$: It is obvious. □

The measure of dependency can be used to evaluate significance of attributes in decision systems. Based on this measure, redundant attributes can be removed and a reduct of attribute set can be reached.

**Definition 3.** Let $S=(U,C\cup D,V,f)$ be a decision system. An attribute $a\in C$ is called relative dispensable with respect to $D$ if $\gamma(D|C\setminus\{a\})=\gamma(D|C)$. $B\subseteq C$ is called a relative reduct of $C$ with respect to $D$ in $S$ if it is the greatest subset of $C$ that every element in $B$ is not dispensable.

For short, a relative reduct of a decision system is called a reduct if there is no confusion arisen. With Definition 3 an equivalent characterization for reduct of a decision system can be obtained.

**Proposition 3.** *Let $S=(U,C\cup D,V,f)$ be a decision system. A subset $B\subseteq C$ is a reduct of $C$ with respect to $D$ in $S$ if and only if*

(1) $\gamma(D|B)=\gamma(D|C)$; *and*
(2) $\gamma(D|B\setminus\{a\})<\gamma(D|B)$ *for any* $a\in B$.

Proposition 3 provides ones a theoretical warrant to find reducts of $C$ in $S$. In large scale systems, however, it is difficult or even impossible to find all reducts of $C$, though to find all reducts is unnecessary in theoretical investigation as well as in application community. To find an optimal reduct of attribute set is more practical. To proceed with such a technique, an evaluation function is very essential in establishing such an optimal algorithm for attribute reduction or feature selection.

**Definition 4.** (Evaluation function based on dependency) Let $S=(U,C\cup D,V,f)$ be a decision system and $B\subseteq C$. The evaluation function based on dependency is defined as, for any attribute $a\in C\setminus B$,

$$Sig_\gamma(a,B,D)=\gamma(D|B\cup\{a\})-\gamma(D|B)$$

It is a fact that the evaluation function based on dependency is a classical measure in assessing the significance degree of attributes in decision systems, where the value of evaluation function at attribute $a$ is designated as the significance measure of $a$ in $S$. It is evident that if $\gamma(D|B\cup\{a\})=\gamma(D|B)$, the significance degree $Sig_\gamma(a,B,D)$ of $a$ is 0 and $a$ is therefore dispensable in $B$. In addition, the larger the significance degree of a feature, the more significant the feature is. The subset of selected features is expected to be a family of attributes in which each element has the largest significance degree in $S$. However, a reduct can not always be obtained by this method in decision systems.

**Example 1.** Consider a decision information system $S=(U,C\cup D,V,f)$ shown in Table 1, where $U=\{x_1,x_2,\ldots,x_{18}\}$ is the universe of discourse,

Table 1: A decision information system

| *Events* | *Outlook* | *Temp* | *Humidity* | *Windy* | *Decision* |
|---|---|---|---|---|---|
| $x_1$ | Sunny | Med | Low | True | Play |
| $x_2$ | Rain | Med | High | True | Play |
| $x_3$ | Sunny | High | Med | True | Don't play |
| $x_4$ | Sunny | High | Med | False | Don't play |
| $x_5$ | Overcast | High | Med | False | Play |
| $x_6$ | Rain | Med | High | True | Don't play |
| $x_7$ | Rain | Low | Low | True | Don't play |
| $x_8$ | Overcast | Med | High | True | Don't play |
| $x_9$ | Sunny | High | Med | True | Play |
| $x_{10}$ | Overcast | Med | High | True | Play |
| $x_{11}$ | Overcast | High | Med | False | play |
| $x_{12}$ | Rain | Med | Med | False | Play |
| $x_{13}$ | Overcast | Low | Low | True | Play |
| $x_{14}$ | Rain | Low | Med | False | Play |
| $x_{15}$ | Rain | Med | High | False | Don't play |
| $x_{16}$ | Sunny | Med | High | False | Don't play |
| $x_{17}$ | Sunny | Med | Low | False | Play |
| $x_{18}$ | Sunny | High | Med | False | Play |

$C = \{a,b,c,d\}$ is the set of conditional attributes with $a =$ *Outlook*, $b =$ *Temperature*, $c =$ *Humidity*, $d =$ *Windy*, and $D = \{$*Decision*$\}$ is the set of decision attribute. Each event $x_i \in U$ is described by four attributes and is classified to either *Don't Play* or *Play*, the values of decision attribute.

It is clear that $\gamma(D|\{a\}) = \gamma(D|\{b\}) = \gamma(D|\{c\}) = \gamma(D|\{d\}) = 0$. No conditional attribute can be sorted out and the reduct of $S$ cannot be efficiently obtained by the method of feature selection based on dependency.

This example indicates that the evaluation function based on dependency is not an effective measure for assessing significance of attributes. In the following an approach to feature selection is proposed by introducing an efficient evaluation function.

## 3. Knowledge granularity and attribute reduction in information systems

The notion of knowledge granularity, derived from information granularity, can be used to describe classification performance for the universe of discourse by a given knowledge. In this section, we present some basic results about knowledge granularity in information systems and propose a revised version of knowledge granularity so as to meet more applications.

**Definition 5.** [21] Let $S = (U, A, V, f)$ be an information system, the knowledge granularity, *GK*, is a mapping from the powerset of $A$ to $[0,1]$ satisfying

(1) Non-negativity. $GK(P) \geqslant 0$ for any $P \subseteq A$;

(2) Invariability. For any $P, Q \subseteq A$ with $|U/P| = |U/Q|$, if there exists a bijection $g : U/P \to U/Q$ such that for any $X \in U/P$, there is $Y \in U/Q$ with $g(X) = Y$, then $GK(P) = GK(Q)$; and

(3) Monotonicity. $GK(P) \leqslant GK(Q)$ for any $P, Q \subseteq A$ with $Q \subseteq P$.

Let $U/P = \{X_1, X_2, \ldots, X_n\}$, it is verified that

$$GK(P) = \sum_{i=1}^{n} \frac{|X_i|^2}{|U|^2}$$

is an expression of knowledge granularity of $P$. The knowledge granularity $GK(P)$ of $P$ can be revised as

follows.

$$\begin{aligned} GK(P) &= \sum_{i=1}^{n} \frac{|X_i|^2}{|U|^2} \\ &= \sum_{i=1}^{n} \frac{|[x_{i_1}]_P| + |[x_{i_2}]_P| + \ldots + |[x_{i_{s_i}}]_P|}{|U|^2} \\ &= \frac{|[x_1]_P| + |[x_2]_P| + \ldots + |[x_{|U|}]_P|}{|U|^2} \\ &= \sum_{x \in U} \frac{|[x]_P|}{|U|^2} \end{aligned}$$

where $X_i = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{s_i}}\}$ with $|X_i| = s_i$ and $\sum_{i=1}^{n} s_i = |U|$.

In order to characterize relationship between two attribute sets, two new concepts, namely joint knowledge granularity and conditional knowledge granularity, are put forward as follows.

**Definition 6.** Let $S = (U, A, V, f)$ be an information system. For $P, Q \subseteq A$, the joint knowledge granularity of $P$ together with $Q$ is defined by

$$GK(P:Q) = \sum_{x \in U} \frac{|[x]_P \cap [x]_Q|}{|U|^2}$$

and the conditional knowledge granularity of $P$ under $Q$ is defined as

$$GK(P|Q) = \frac{\sum_{x \in U} |[x]_P \cap [x]_Q|}{\sum_{x \in U} |[x]_Q|}$$

$GK(P|Q)$ can be considered as an alternative version of inclusion degree of $U/Q$ being included in $U/P$. The joint knowledge granularity and conditional knowledge granularity have many properties, parts of which are listed as follows.

**Proposition 4.** *Let $S = (U, A, V, f)$ be an information system. For any $P, Q \subseteq A$, one has*

(1) $GK(P:Q) = GK(P|Q)GK(Q)$;
(2) $GK(P:Q) \leqslant \min\{GK(P), GK(Q)\}$;
(3) $GK(P:Q) = GK(P \cup Q)$;
(4) $\frac{1}{|U|} \leqslant GK(P) \leqslant 1$ *and* $\frac{1}{|U|} \leqslant GK(P:Q) \leqslant 1$.

*Furthermore, if $P \subseteq Q$, then*

(5) $GK(P:Q) = GK(Q)$.

**Proof** (1) – (3) and (5) are obvious from Definition 6. It is sufficient to verify (4).

It is clear that $\frac{1}{|U|} \leqslant GK(P) \leqslant 1$ as shown in Property 1-3 in the reference[21].

In view of the fact $GK(P:Q) = GK(P \cup Q)$, $GK(P:Q)$ has the same range as $GK(P)$. □

Since $GK(P:Q) \leqslant \min\{GK(P), GK(Q)\} \leqslant GK(Q)$, one has that $GK(P|Q) = GK(P:Q)/GK(Q) \leqslant 1$. By the fact that $\frac{1}{|U|} \leqslant GK(P)$ and $\frac{1}{|U|} \leqslant GK(P:Q) \leqslant 1$, it is clear that $GK(P|Q) = GK(P:Q)/GK(Q) \geqslant \min_{P,Q \subseteq A}\{GK(P:Q)\}/\max_{Q \subseteq A}\{GK(Q)\} = \frac{1}{|U|}$. Hence, the following proposition holds.

**Proposition 5.** *Let $S = (U, A, V, f)$ be an information system. For any $P, Q \subseteq A$, one has*

(1) $\frac{1}{|U|} \leqslant GK(P|Q) \leqslant 1$;
(2) $GK(P|Q) = 1$ *if* $P \subseteq Q$;
(3) $GK(P|Q) = 1/|U|$ *iff* $U/P = \omega$ *and* $U/Q = \delta$, *where* $\omega = \{\{x\} | x \in U\}$ *and* $\delta = \{U\}$.

From Proposition 5, we conclude that the conditional knowledge granularity $GK(P|Q)$ reflects the close relationship of $P$ together with $Q$, and the larger the value of $GK(P|Q)$, $U/Q$ is included in $U/P$ with a higher degree.

Based on the definition and properties of conditional knowledge granularity one has the following equivalent statements.

**Theorem 6.** *Let $S = (U, A, V, f)$ be an information system and $a \in A$. The following statements are equivalent.*

(1) *$a$ is dispensable in $A$;*
(2) $GK(A \setminus \{a\}) = GK(A)$;
(3) $GK(\{a\} | A \setminus \{a\}) = 1$.

**Proof** (1) $\Rightarrow$ (2): If $a$ is dispensable in $A$, then $R_A = R_{A \setminus \{a\}}$, and so $[x]_A = [x]_{A \setminus \{a\}}$ for all $x \in U$. Thus

$$GK(A) = \sum_{x \in U} \frac{|[x]_A|}{|U|^2} = \sum_{x \in U} \frac{|[x]_{A \setminus \{a\}}|}{|U|^2} = GK(A \setminus \{a\})$$

(2) $\Rightarrow$ (3): If $GK(A\setminus\{a\}) = GK(A)$, then

$$\begin{aligned} GK(\{a\}|A\setminus\{a\}) &= \frac{GK(\{a\}:A\setminus\{a\})}{GK(A\setminus\{a\})} \\ &= \frac{GK(\{a\}\cup(A\setminus\{a\}))}{GK(A)} \\ &= 1 \end{aligned}$$

(3) $\Rightarrow$ (1): If $GK(\{a\}|A\setminus\{a\}) = 1$, then $GK(A) = GK(A\setminus\{a\})$, e.g.,

$$\sum_{x\in U}\frac{|[x]_A|}{|U|^2} = \sum_{x\in U}\frac{|[x]_{A\setminus\{a\}}|}{|U|^2}$$

In view of the fact that $[x]_A \subseteq [x]_{A\setminus\{a\}}$, there must be $[x]_A = [x]_{A\setminus\{a\}}$ for all $x\in U$. Consequently, $R_A = R_{A\setminus\{a\}}$, and so $a$ is dispensable in $A$. □

**Corollary 7.** *Let $S = (U, A, V, f)$ be an information system and $a \in B \subseteq A$. The following statements are equivalent.*

(1) *$a$ is independent in $B$;*
(2) *$GK(B\setminus\{a\}) > GK(B)$;*
(3) *$GK(\{a\}|B\setminus\{a\}) < 1$.*

**Theorem 8.** *Let $S = (U, A, V, f)$ be an information system and $B \subseteq A$. Then $B$ is a reduct of $A$ in $S$ if and only if*

(1) *$GK(A) = GK(B)$; and*
(2) *$GK(\{a\}|B\setminus\{a\}) < 1$ for any $a \in B$.*

**Proof** It is straightforward from Theorem 6 and Corollary 7. □

The following example shows a straightforward application of Theorem 8.

**Example 2.** Given an information system $S = (U, A, V, f)$, shown in Table 2, where $U = \{1,2,3,4,5,6,7\}$ is the set of objects and $A = \{a,b,c,d,e\}$ is the attribute set in which *a*, *b*, *c*, *d*, and *e* represent temperature, humidity, wind, fertilization, and pesticide, respectively. Let $B = \{a,b,c,e\}$, by calculation one has

$$\begin{aligned} GK(B) &= \sum_{x\in U}\frac{|[x]_B|}{|U|^2} \\ &= \frac{1+1+1+1+1+1+1}{7^2} \\ &= \frac{7}{49} = GK(A) \end{aligned}$$

$$\begin{aligned} GK(B\setminus\{a\}) &= GK(B\setminus\{c\}) = GK(B\setminus\{e\}) \\ &= \frac{2^2+1+1+1+1+1}{7^2} = \frac{9}{49} \\ GK(B\setminus\{b\}) &= GK(\{a,c,e\}) \\ &= \frac{2^2+1+1+2^2+1}{7^2} = \frac{11}{49} \end{aligned}$$

That is, for any $x\in B$, one has $GK(\{x\}|B\setminus\{x\}) = \frac{GK(B)}{GK(B\setminus\{x\})} < 1$. Therefore, $B = \{a,b,c,e\}$ is a reduct of $A$. The result coincides with practical desire.

## 4. Feature selection in consistent decision systems

**Definition 7.** Let $S = (U, C\cup D, V, f)$ be a consistent decision system and $B \subseteq C$. The conditional knowledge granularity of decision attribute set $D$ under the conditional attribute subset $B$ is defined by

$$GK(D|B) = \frac{\sum_{x\in U}|[x]_D \cap [x]_B|}{\sum_{x\in U}|[x]_B|}$$

One can see that the quantity $\frac{|[x]_D\cap[x]_B|}{|[x]_B|}$ characterizes the inclusion degree of $[x]_B$ being included in $[x]_D$. Therefore, the conditional knowledge granularity $GK(D|B)$ can be regarded as a generalization of inclusion degree of $U/B$ being included in $U/D$. In combination with Proposition 4 the following property is straightforward.

**Proposition 9.** *Let $S = (U, C\cup D, V, f)$ be a decision system, then*

(1) *$GK(D:B) \leqslant \min\{GK(B), GK(D|B)\}$ for any $B\subseteq C$;*
(2) *$GK(D:C) = 1$ iff $S$ is a consistent decision systems.*

Table 2: An information system

| $U$ | $a$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| 1 | High | Middle | Middle | fer1 | pes1 |
| 2 | Middle | Middle | Middle | fer2 | pes1 |
| 3 | Middle | High | Weak | fer1 | pes2 |
| 4 | Middle | High | Middle | fer1 | pes1 |
| 5 | High | Middle | Weak | fer1 | pes1 |
| 6 | High | Middle | Middle | fer1 | pes2 |
| 7 | High | High | Middle | fer1 | pes2 |

**Theorem 10.** *Let $S=(U,C\cup D,V,f)$ be a consistent decision system and $B\subseteq C$. Then $\gamma(D|B)=\gamma(D|C)$ iff $GK(D|B)=GK(D|C)$.*

**Proof** Since $S$ is a consistent decision system, it is trivial that $\gamma(D|C)=1$ and that $GK(D|C)=1$.

$\Rightarrow$: Let $B\subseteq C$, if $\gamma(D|B)=\gamma(D|C)$, then $Pos_B(D)=Pos_C(D)=U$. Therefore, for any $x\in U$, one has $[x]_B\subseteq[x]_D$. As a result,

$$GK(D|B)=\frac{\sum_{x\in U}|[x]_D\cap[x]_B|}{\sum_{x\in U}|[x]_B|}=1$$

$\Leftarrow$: Since $B\subseteq C$, for any $x\in U$, one has $[x]_C\subseteq[x]_B$. In view of the fact that $GK(D|B)=GK(D|C)$ and that $GK(D|C)=1$, one has $GK(D|B)=\frac{\sum_{x\in U}|[x]_D\cap[x]_B|}{\sum_{x\in U}|[x]_B|}=1$. Thus $|[x]_B\cap[x]_D|=|[x]_B|$, which implies that for any $x\in U$, $[x]_B\subseteq[x]_D$. Therefore, $Pos_B(D)=U$. The consequences $Pos_B(D)=Pos_C(D)$ and $\gamma(D|B)=\gamma(D|C)$ hold. □

**Corollary 11.** *Let $S=(U,C\cup D,V,f)$ be a consistent decision system. Then $a\in C$ is dispensable with respect to $D$ in $S$ if and only if $GK(D|C\setminus\{a\})=GK(D|C)$, whereas $a$ is independent if and only if $GK(D|C\setminus\{a\})<GK(D|C)$.*

Corollary 11 guarantees that when a dispensable attribute is deleted, the conditional knowledge granularity of decision attribute set under the set of rest conditional attributes keeps invariant. The following theorem characterizes attribute reduction of a consistent decision system, whose proof is straightforward.

**Theorem 12.** *Let $S=(U,C\cup D,V,f)$ be a consistent decision system and $B\subseteq C$. Then $B$ is a reduct of $C$ with respect to $D$ in $S$ if and only if*

(1) $GK(D|B)=GK(D|C)$*; and*
(2) $GK(D|B\setminus\{a\})<GK(D|B)$ *for any $a\in B$.*

**Definition 8.** (Evaluation function based on conditional knowledge granularity) Let $S=(U,C\cup D,V,f)$ be a consistent decision system and $B\subseteq C$. The evaluation function based on conditional knowledge granularity is defined by, for any $a\in C\setminus B$,

$$Sig_{GK}(a,B,D)=GK(D|B\cup\{a\})-GK(D|B)$$

It is noted that a larger quantity of $Sig_{GK}(a,B,D)$ indicates a stronger association between $B\cup\{a\}$ and $D$. Therefore, the evaluation function based on conditional knowledge granularity can characterize significance of every conditional attribute with respect to decision attribute set in consistent decision systems. Succedent examples show that this function can serve as a rational measure to find a reduct by an optimal algorithm for feature selection. Such an algorithm aims to find an attribute with the greatest amount of significance, or alternatively, to make the equivalence class of each sample be included in its corresponding decision class as much as possible. The forward greedy search algorithm can serve as evaluating feasibility and efficiency of the proposed measure for feature selection in consistent decision systems.

**Algorithm 13** *Forward greedy search algorithm of feature selection based on conditional knowledge granularity in consistent decision systems (FGS-CKG-CDS):*
*Input:* $(U,C\cup D,V,f)$

*Output: red*
*1:* $\emptyset \rightarrow red, 0 \rightarrow temp$
*2: for each* $a_i \in C \setminus red$
*3: compute the significance* $Sig_{GK}(a_i, red, D) = GK(D|red \cup \{a_i\}) - GK(D|red)$
*4: end for*
*5: select the attribute* $a_k$ *such that*
*6:* $Sig_{GK}(a_k, red, D) = \max_i Sig_{GK}(a_i, red, D)$
*7: if* $temp < Sig_{GK}(a_k, red, D)$
*8:* $temp = Sig_{GK}(a_k, red, D)$
*9:* $red \cup \{a_k\} \rightarrow red$
*10: goto 2*
*11: else*
*12: return red*
*13: end if*
*14: end*

In the first iteration, we start the set of selected features *red* with empty set and specify $GK(D|red) = 0$. Then adding an attribute $a$ with the greatest value of significance $Sig_{GK}(a, red, D)$ into *red*. The rest features in each iteration are all evaluated and the one with the greatest significance is chosen and added into *red*. The algorithm does not stop until adding any of the rest features to selected feature set will not bring any increment.

There exist two main operations of consuming time in this algorithm. One is to compute the value of $Sig_{GK}(a, red, D)$ and the other is to judge whether the value of $Sig_{GK}(a, red, D)$ is the greatest one or not. The first procedure is carried out in a time complexity $O(|U|^2)$, while the time complexity in the second step is $O(|C|^2)$. Therefore, the FGS-CKG-CDS approach to feature selection works in a straightforward way with overall time complexity $O(|U|^2|C|^2)$. In the worst case, the whole computational complexity of this algorithm is $|U|^2 \times |C| + |U|^2 \times (|C|-1) + \ldots + |U|^2 = (|C|+1) \times |C| \times |U|^2/2$.

The FGS-CKG-CDS algorithm has the same time complexity as the one based on dependence and as the one based on mutual information. In the following we present an example to show the procedure of feature selection by employing the FGS-CKG-CDS.

**Example 3.** Given a consistent decision system shown in Table 3, where $C = \{a, b, c, d, e\}$ is the conditional attribute set and $D$ is the decision attribute.

At first, let $red = \emptyset$, the corresponding knowledge granularity is computed as follows.

$$\begin{aligned} GK(\{a\}) &= \sum_{x \in U} \frac{|[x]_{\{a\}}|}{|U|^2} = 0.5000 \\ GK(D:\{a\}) &= GK(D \cup \{a\}) \\ &= \frac{1+2^2+2^2+3^2}{8^2} = 0.2813 \\ GK(D|\{a\}) &= \frac{GK(D:\{a\})}{GK(\{a\})} = 0.5625 \end{aligned}$$

In a similar way, one has

$$\begin{aligned} GK(D|\{b\}) &= GK(D|\{d\}) \\ &= \frac{1+2^2+1+4^2}{2^2+6^2} = 0.5500 \\ GK(D|\{c\}) &= \frac{1+2^2+5^2}{1+7^2} = 0.6000 \\ GK(D|\{e\}) &= \frac{3^2+2^2+3^2}{5^2+3^2} = 0.6471 \end{aligned}$$

It is obvious that $GK(D|\{e\}) = \max_{x \in C} GK(D|\{x\})$, so $e$ is firstly added to *red*, i.e., $red = red \cup \{e\} = \{e\}$.

Repeating the same steps we will obtain $red = \{e, c, b, d\}$.

Since $GK(D|red) = GK(D|C)$ and $GK(D|red \setminus \{x\}) < GK(D|C)$ for any $x \in red$, the set $red = \{e, c, b, d\}$ is therefore a reduct of $C$ with respect to $D$.

According to the evaluation function based on dependency for feature selection of Table 3, one obtains a reduct $\{e, b, d, c\}$ of $C$, the same result as aforementioned, but with different orders. Moreover, with an application of evaluation function based on mutual information to feature selection, one obtains a reduct $\{e, b, a, c, d\}$, which is larger than that obtained by the proposed measure.

## 5. Feature selection in inconsistent decision systems

Rough communication is initialized as a bridge between information systems in granular computing [44]. The purpose of this section is to find a reduct

Table 3: A consistent decision system

| $U$ | $a$ | $b$ | $c$ | $d$ | $e$ | $D$ |
|---|---|---|---|---|---|---|
| 1 | Middle | High | Middle | fer1 | pes1 | High |
| 2 | High | Middle | Middle | fer2 | pes1 | High |
| 3 | High | Middle | High | fer1 | pes1 | High |
| 4 | High | Middle | Middle | fer1 | pes1 | Low |
| 5 | Middle | Middle | Middle | fer1 | pes1 | Low |
| 6 | High | High | Middle | fer1 | pes2 | Low |
| 7 | Middle | Middle | Middle | fer2 | pes2 | Low |
| 8 | Middle | Middle | Middle | fer1 | pes2 | Low |

of an inconsistent decision system by means of constructing a proper rough communication and converting the inconsistent decision system to a consistent decision system.

For a decision system $S=(U,C\cup D,V,f)$, if we denote $V$ by $V_C\cup V_D$, where $V_C$ and $V_D$ denote the set of values of conditional attributes and that of values of decision attributes, respectively, then $S$ can be denoted by $S=(U,C\cup D,V_C\cup V_D,f)$. Meanwhile, the partition of $U$ under the decision attribute set $D$ is symbolled by $U/V_D$. With the idea that a decision system $S$ is inconsistent if and only if its boundary region $BN_C(D)=U\setminus POS_B(D)$ is nonempty, we develop the following rough communication between two decision systems.

**Definition 9.** Let $S=(U,C\cup D,V_C\cup V_D,f)$ be an inconsistent decision system and $S'=(U,C\cup D,V_C\cup V'_D,f')$ be another decision system with $V_D\neq V'_D$, and let $U/V_D=\{D_1,D_2,\ldots,D_N\}$ and $U/V'_D=\{D'_1,D'_2,\ldots,D'_M,D'_{M+1}\}$ be the partitions of $U$ with respect to the decision attribute set $D$ (corresponding to different decision attribute values). If there exists a mapping $F$ from $S$ to $S'$ satisfying, for each $i\in\{1,2,\ldots,N\}$, there exists a unique $j\in\{1,2,\ldots,M\}$, and for each $j\in\{1,2,\ldots,M\}$, there exists a unique $i\in\{1,2,\ldots,N\}$ such that

$$D'_j=F(D_i),D'_{M+1}=U\setminus\cup_{i=1}^{N}F(D_i)$$

then $F$ is called a rough communication between $S$ and $S'$. Meanwhile, $S'$ is called the induced decision system from $S$.

According to the definition of rough communication, it is verified that, for any inconsistent decision system, its induced decision system is undoubtedly a consistent decision system.

Given an inconsistent decision system, it is easy to construct such a mapping (rough communication) in Definition 9. For example, given a set $B\subseteq C$, let $F(X)$ be the positive region of $X$ with respect to $B$, $POS_B(X)$, then an inconsistent decision system $S$ can be converted to another decision system $S'$ by means of this rough communication.

In the following such a rough communication is employed to deal with the problem of feature selection in inconsistent decision system. It is observed that $D'_{M+1}$ is composed of the objects that do not fall into the positive region of $D$ and that the values of decision attributes of objects in each $D'_j$, $j=1,2,\ldots,M$, are the same as the ones in some unique $D_i$, $i=1,2,\ldots,N$, whereas the value of decision attributes of objects in $D'_{M+1}$ is assigned to another that is different from those corresponding to $D'_j$, $j=1,2,\ldots,M$. As a whole, the mapping of rough communication does not change the decision system $S$ except the value of decision attributes of objects in the boundary region $BN_C(D)$ in $S$.

For example, let $S$ be a decision system as shown in Table 4, one can obtain that $POS_C(D)=\{u_1,u_3,u_4,u_5,u_6,u_9\}$ and $BN_C(D)=\{u_2,u_7,u_8\}$. $S$ is therefore an inconsistent decision system. According to the rough communication offered above, $S$ can be converted to a new decision system $S'$ in which the value of decision attribute $D$ of objects in $BN_C(D)$ is changed to another value, 3, for example. That is, $\{u_2,u_7,u_8\}$ is labelled as a new decision class $D'_3$. A new decision system $S'$ is constructed and is shown in Table 5. It is evident that $S'$ is a

Table 4: An inconsistent decision system $S$

| $U$ | $a$ | $b$ | $c$ | $d$ | $e$ | $D$ |
|---|---|---|---|---|---|---|
| $u_1$ | 1 | 1 | 2 | 1 | 1 | 2 |
| $u_2$ | 1 | 2 | 1 | 2 | 2 | 1 |
| $u_3$ | 2 | 2 | 2 | 1 | 1 | 2 |
| $u_4$ | 2 | 2 | 1 | 1 | 1 | 1 |
| $u_5$ | 1 | 2 | 1 | 1 | 1 | 1 |
| $u_6$ | 2 | 1 | 1 | 1 | 2 | 1 |
| $u_7$ | 1 | 2 | 1 | 2 | 2 | 1 |
| $u_8$ | 1 | 2 | 1 | 2 | 2 | 2 |
| $u_9$ | 1 | 2 | 1 | 1 | 2 | 1 |

Table 5: The induced consistent decision system $S'$

| $U$ | $a$ | $b$ | $c$ | $d$ | $e$ | $D$ |
|---|---|---|---|---|---|---|
| $u_1$ | 1 | 1 | 2 | 1 | 1 | 2 |
| $u_2$ | 1 | 2 | 1 | 2 | 2 | 3 |
| $u_3$ | 2 | 2 | 2 | 1 | 1 | 2 |
| $u_4$ | 2 | 2 | 1 | 1 | 1 | 1 |
| $u_5$ | 1 | 2 | 1 | 1 | 1 | 1 |
| $u_6$ | 2 | 1 | 1 | 1 | 2 | 1 |
| $u_7$ | 1 | 2 | 1 | 2 | 2 | 3 |
| $u_8$ | 1 | 2 | 1 | 2 | 2 | 3 |
| $u_9$ | 1 | 2 | 1 | 1 | 2 | 1 |

consistent decision system.

In general, it is possible that there exists some $i \in \{1,2,\ldots,N\}$ satisfying $F(D_i) = \emptyset$. In which case, the objects in $D_i$ are all assigned to $D'_{M+1}$. However, without loss of generality, it is assumed that $M = N$, i.e., every positive region of $D_i\,(i = 1,2,\ldots,N)$ with respect to $B \subseteq C$ is nonempty. Thus $D'_1 = POS_B(D_1), D'_2 = POS_B(D_2), \ldots, D'_N = POS_B(D_N)$, and $D'_{N+1} = BN_B(D)$.

If one can verify the statement that a subset of condition attribute set is a reduct of an inconsistent decision system if and only if it is a reduct of the induced consistent decision system, then one can obtain a reduct of an inconsistent decision system by rough communication. In the sequel, we will investigate this issue. The following proposition is obvious.

**Proposition 14.** *Let $S = (U, A, V, f)$ be an information system, then for any $B \subseteq A$,*

(1) *$POS_B(X) \subseteq POS_B(Y)$ for all $X, Y \subseteq U$ with $X \subseteq Y$;*

(2) *$POS_B(X) = POS_B(POS_B(X))$ for all $X \subseteq U$.*

We firstly study the relationship between an inconsistent decision system and its induced consistent decision system.

**Theorem 15.** *Assume that $S = (U, C \cup D, V_C \cup V_D, f)$ is an inconsistent decision system, $S' = (U, C \cup D, V_C \cup V'_D, f')$ is its induced decision system, and $B \subseteq C$. Then $\gamma(V_D|B) = \gamma(V_D|C)$ if and only if $\gamma(V'_D|B) = \gamma(V'_D|C)$, where $\gamma(V_D|B)$ and $\gamma(V'_D|B)$ denote the dependency of $D$ on $B$ in $S$ and in $S'$, respectively.*

**Proof** $\Rightarrow$: Let $\gamma(V_D|B) = \gamma(V_D|C)$, then $\cup_{D_i \in U/V_D} POS_B(D_i) = \cup_{D_i \in U/V_D} POS_C(D_i)$. Therefore, $POS_B(D_i) = POS_C(D_i)$ for all $D_i \in U/V_D$.

For any $D'_i \in U/V'_D$, it is clear that

$$POS_B(D'_i) = POS_B(POS_B(D_i)) = POS_B(D_i) = D'_i$$

Thus $POS_B(D'_i) = POS_C(D'_i)$, for all $i = 1,2,\ldots,N$, due to the fact that $S'$ is a consistent decision system. Furthermore,

$$\begin{aligned} POS_B(D'_{N+1}) &= BN_B(D) \\ &= U \setminus POS_B(D) \\ &= U \setminus POS_C(D) \\ &= BN_C(D) \\ &= POS_C(D'_{N+1}) \end{aligned}$$

As a result, $\cup_{D'_i \in U/V'_D} POS_B(D'_i) = \cup_{D'_i \in U/V'_D} POS_C(D'_i)$, i.e., $\gamma(V'_D|B) = \gamma(V'_D|C)$.

$\Leftarrow$: If $\gamma(V'_D|B) = \gamma(V'_D|C)$, then, for all $D'_i \in U/V'_D$, $POS_B(D'_i) = POS_C(D'_i)$ due to the fact that $POS_B(D'_i) \subseteq POS_C(D'_i)$ and $\cup_{D'_i \in U/V'_D} POS_B(D'_i) = \cup_{D'_i \in U/V'_D} POS_C(D'_i)$. In particular, $POS_B(D'_{N+1}) = POS_C(D'_{N+1})$, i.e., $BN_B(D) = BN_C(D)$. Therefore, $\cup_{D_i \in U/V_D} POS_B(D_i) = U \setminus BN_B(D) = U \setminus BN_C(D) = \cup_{D_i \in U/V_D} POS_C(D_i)$, and the equality $\gamma(V_D|B) = \gamma(V_D|C)$ is implied. □

In term of Theorems 15 and the results in Section 4 the following corollaries hold.

**Corollary 16.** *Assume that $S = (U, C \cup D, V_C \cup V_D, f)$ is an inconsistent decision system, $S' = (U, C \cup D, V_C \cup V'_D, f')$ is its induced decision system, and $B \subseteq C$. Then $a \in B$ is dispensable with respect to $D$ in $S$ if and only if $\gamma(V'_D|B \setminus \{a\}) = \gamma(V'_D|B)$; whereas $a \in B$ is independent if and only if $\gamma(V'_D|B \setminus \{a\}) < \gamma(V'_D|B)$.*

**Corollary 17.** *Assume that $S = (U, C \cup D, V_C \cup V_D, f)$ is an inconsistent decision system, $S' = (U, C \cup D, V_C \cup V'_D, f')$ is its induced decision system, and $B \subseteq C$. Then $a \in B$ is dispensable with respect to $D$ in $S$ if and only if $GK(V'_D|B \setminus \{a\}) = GK(V'_D|B)$; whereas $a \in B$ is independent if and only if $GK(V'_D|B \setminus \{a\}) < GK(V'_D|B)$, where $GK(V'_D|B)$ denotes the conditional knowledge granularity of $D$ under $B$ in $S'$.*

The consequences presented above ensure that a subset of conditional attribute set is a reduct of an inconsistent decision system if and only if it is a reduct of the induced consistent decision system. This conclusion can be summarized as follows.

**Theorem 18.** *Assume that $S = (U, C \cup D, V_C \cup V_D, f)$ is an inconsistent decision system, $S' = (U, C \cup D, V_C \cup V'_D, f')$ is its induced decision system, and $B \subseteq C$, then $B$ is a reduct of $C$ with respect to $D$ (related to $V_D$) in $S$ if and only if $B$ is a reduct of $C$ with respect to $D$ (related to $V'_D$) in $S'$.*

In virtue of these conclusions, the problem of finding a reduct in an inconsistent decision system is equivalent to that in its induced consistent decision system.

**Theorem 19.** *Let $S = (U, C \cup D, V_C \cup V_D, f)$ be an inconsistent decision system, $S' = (U, C \cup D, V_C \cup V'_D, f')$ be the induced decision system of $S$, and $B \subseteq C$. Then $B$ is a reduct of $C$ with respect to $D$ in $S$ if and only if*

(1) $GK(V'_D|B) = GK(V'_D|C)$; *and*
(2) $GK(V'_D|B \setminus \{a\}) < GK(V'_D|B)$ *for any* $a \in B$.

Similar to Definition 8, an evaluation function based on conditional knowledge granularity in an inconsistent decision system $S = (U, C \cup D, V_C \cup V_D, f)$ can be defined by, for any $a \in C \setminus B$,

$$Sig_{GK}(a, B, D) = GK(V'_D|B \cup \{a\}) - GK(V'_D|B)$$

Following Theorem 19, an optimal algorithm to find a reduct of an inconsistent decision system, called FGS-CKG-IDS, can be designed analogously to Algorithm 13.

In comparison to FGS-CKG-CDS for feature selection in consistent decision system, the FGS-CKG-IDS algorithm is of only one additional step to judge which the system is consistent or not. In practical applications, it is not necessary to highlight inconsistent decision systems. If a decision system is inconsistent, it is sufficient to reclassify the decision classes and to construct its induced consistent decision system by means of rough communication.

In the end of this section, an example is presented showing validity of the proposed method of feature selection in inconsistent decision system.

Let $S = (U, C \cup D, V_C \cup V_D, f)$ be a decision information system, shown in Example 1 and Table 1, where $V_D = \{Play, Don't\ play\}$.

By computation one obtains that $POS_C(D) = \{x_1, x_3, x_4, x_5, x_7, x_9, x_{11}, x_{12}, x_{13}, x_{14}, x_{17}, x_{18}\}$ and $BN_C(D) = \{x_2, x_6, x_8, x_{10}, x_{15}, x_{16}\}$, so $S$ is an inconsistent decision system. We assign another

value, '*Maybe*' for instance, to that of decision attribute $D$ of elements in $BN_C(D)$. Then $S$ is converted to $S' = (U, C \cup D, V_C \cup V'_D, f')$, shown in Table 6, where $V'_D = \{Play, Maybe, Don't\ play\}$. It is clear that $S'$ is a consistent decision system.

The FGS-CKG-IDS algorithm is employed to perform feature selection for the above decision system. At first, let $red = \emptyset$ and $GK(V'_D|red) = 0$, then the conditional knowledge granularity can be computed as follows.

$$GK(V'_D|\{a\}) = 0.4182, GK(V'_D|\{b\}) = 0.5890$$
$$GK(V'_D|\{c\}) = 0.8103, GK(V'_D|\{d\}) = 0.4024$$

Clearly, $GK(V'_D|\{c\}) = \max_{x \in C} GK(V'_D|\{x\})$, so $c$ is added to $red$, i.e., $red = red \cup \{c\} = \{c\}$. Again,

$$\begin{aligned} GK(V'_D|\{c,a\}) &= 1.0000, \\ GK(V'_D|\{e,b\}) &= 0.8889, \\ GK(V'_D|\{c,d\}) &= 0.8235 \end{aligned}$$

Since $GK(V'_D|\{c,a\}) = \max_{x \in C \setminus red} GK(V'_D|\{c,x\})$, thus $red = red \cup \{a\} = \{c,a\}$.

Since $GK(V'_D|red) = GK(V'_D|C) = 1$ and $GK(V'_D|red \setminus \{x\}) < GK(V'_D|C)$ for any $x \in red$, the set $red = \{c,a\}$ is therefore a reduct of $C$.

## 6. Experiments and analysis

In order to verify effects of the proposed approach to feature selection, comparative experiments have been implemented to show the results with those based on dependency and based on mutual information.

Six standard data sets from UCI machine learning data repository, University of California at Irvine [27] are employed in our experiments. The number of samples in these data sets varies from 198 to 20000. Some of data sets are mixed with continuous attribute values and discrete ones. They are therefore preprocessed by discretizing the ranges of attributes and segmenting these ranges into several equal-width intervals. It is checked that three of them are consistent decision systems and the rest are inconsistent, shown in Table 7.

As a widely used technique to test and evaluate classification accuracies for data sets, the 10-fold cross validation [10] is used to divide the samples into 10 subsets at random and nine of them are used as training set and the rest one as the test set. After 10 rounds, the average value and variation are considered as the final classification accuracy. In our experiments, the popular CART algorithm, linear support vector machine algorithm (SVM) and C4.5 algorithm are used to test classification accuracies of raw data sets and of selected subsets.

Table 8, Table 9 and Table 10 show comparative results of classification accuracies of feature selection through the optimal algorithm with three different evaluation functions by using CART, SVM and C4.5, respectively. The columns with "$\gamma$", "$I$" and "$GK$" denote, separatively, the classification accuracies of feature selection based on dependency, on mutual information, and on the conditional knowledge granularity. "$\surd$" marks the results with the largest classification accuracy among those.

It is observed from the tables that the average classification accuracy of objects with respect to the selected attributes based on the conditional knowledge granularity is larger than that based on dependency and based on mutual information whenever the test algorithm is taken to be CART, SVM or C4.5. In detail, they are 0.8612 (based on dependency), 0.8593 (based on mutual information), 0.8726 (based on conditional knowledge granularity) with CART, 0.8554, 0.8572, 0.8579 with SVM, and 0.7557, 0.7555, 0.7686 with C4.5.

Furthermore, the number of selected subsets with the highest classification accuracy by the conditional knowledge granularity is larger than that by dependency and by mutual information with the test methods, CART and C4.5. They are 0 (based on dependency), 1 (based on mutual information) and 5 (based on conditional knowledge granularity) with the use of CART, and 3, 2 and 6 with the C4.5 algorithm, whereas the number of selected subsets with the highest classification accuracy by the proposed techniques is similar to that based on mutual information, but slightly weaker than that based on dependency with SVM.

In addition, the experimental results show that all

Table 6: The induced consistent decision system

| *Events* | *Outlook* | *Temp* | *Humidity* | *Windy* | *Decision* |
|---|---|---|---|---|---|
| $u_1$ | Sunny | Med | Low | True | Play |
| $u_2$ | Rain | Med | High | True | Maybe |
| $u_3$ | Sunny | High | Med | True | Don't play |
| $u_4$ | Sunny | High | Med | False | Don't play |
| $u_5$ | Overcast | High | Med | False | Play |
| $u_6$ | Rain | Med | High | True | Maybe |
| $u_7$ | Rain | Low | Low | True | Don't play |
| $u_8$ | Overcast | Med | High | True | Maybe |
| $u_9$ | Sunny | High | Med | True | Play |
| $u_{10}$ | Overcast | Med | High | True | Maybe |
| $u_{11}$ | Overcast | High | Med | False | play |
| $u_{12}$ | Rain | Med | Med | False | Play |
| $u_{13}$ | Overcast | Low | Low | True | Play |
| $u_{14}$ | Rain | Low | Med | False | Play |
| $u_{15}$ | Rain | Med | High | False | Maybe |
| $u_{16}$ | Sunny | Med | High | False | Maybe |
| $u_{17}$ | Sunny | Med | Low | False | Play |
| $u_{18}$ | Sunny | High | Med | False | Play |

Table 7: The raw data sets from UCI

| Data set | Abbreviation | Samples | Features | Class | Consistent |
|---|---|---|---|---|---|
| Australian credit approval | Credit | 690 | 15 | 2 | inconsistent |
| Horse colic | Horse | 368 | 23 | 2 | consistent |
| Mushroom | Mushroom | 8124 | 23 | 2 | consistent |
| Letter Recognition | Letter | 20000 | 17 | 26 | consistent |
| Wisconsin diagnostic breast cancer | Wdbc | 569 | 31 | 2 | inconsistent |
| Wisconsin prognostic breast cancer | Wpbc | 198 | 34 | 2 | inconsistent |

of the six selected feature sets obtained by the proposed technique have the highest classification accuracies with the use of the popular C4.5 test algorithm.

## 7. Conclusions

This paper generalizes the notion of knowledge granularity to conditional knowledge granularity so as to solve the problem of feature selection in decision systems. An evaluation function based on the proposed knowledge granularity is designed to measure significance of attributes. This function reflects the measure of a nonlinear combination of inclusion degrees of equivalence class of each sample being included in its decision class.

A rough communication between an inconsistent decision system and a consistent decision system has been established and the problem of feature selection in an inconsistent decision system is converted to that in its induced consistent system. Equivalent characterizations of attribute reduction have been established based on the proposed evaluation function. Optimal algorithms for feature selection have been realized in both decision systems.

Numerical experiments and comparative inves-

Table 8: Comparison of classification accuracies based on different methods with CART

| Data sets | raw data | $\gamma$ | $I$ | $GK$ |
|---|---|---|---|---|
| Credit | 0.8273±0.1486 | 0.8259±0.1474 | 0.8172±0.1428 | 0.8302±0.1507$\surd$ |
| Horse | 0.9592±0.0230 | 0.8911±0.0491 | 0.8996±0.0487 | 0.9320±0.0447$\surd$ |
| Mushroom | 0.9637±0.0990 | 0.9637±0.0990 | 0.9685±0.0996$\surd$ | 0.9637±0.0990 |
| Letter | 0.8656±0.0105 | 0.8569±0.0130 | 0.8599±0.0105 | 0.8654±0.0103$\surd$ |
| Wdbc | 0.9050±0.0455 | 0.9227±0.0334 | 0.9192±0.0380 | 0.9279±0.0337$\surd$ |
| Wpbc | 0.7063±0.0754 | 0.7068±0.0850 | 0.6911±0.0953 | 0.7166±0.0705$\surd$ |
| Average | 0.8712 | 0.8612 | 0.8593 | 0.8726 |

Table 9: Comparison of classification accuracies based on different methods with SVM

| Data sets | raw data | $\gamma$ | $I$ | $GK$ |
|---|---|---|---|---|
| Credit | 0.8548±0.1851 | 0.8548±0.1851$\surd$ | 0.8548±0.1851$\surd$ | 0.8548±0.1851$\surd$ |
| Horse | 0.8914±0.0443 | 0.8830±0.0532 | 0.9157±0.0471$\surd$ | 0.9076±0.0552 |
| Mushroom | 0.9234±0.1261 | 0.8736±0.1039$\surd$ | 0.8664±0.1886 | 0.8736±0.1039$\surd$ |
| Letter | 0.8226±0.0109 | 0.7805±0.0140$\surd$ | 0.7657±0.0139 | 0.7707±0.0140 |
| Wdbc | 0.9773±0.0248 | 0.9773±0.0234$\surd$ | 0.9773±0.0234$\surd$ | 0.9773±0.0234$\surd$ |
| Wpbc | 0.7737±0.0773 | 0.7632±0.0304$\surd$ | 0.7632±0.0304$\surd$ | 0.7632±0.0304$\surd$ |
| Average | 0.8739 | 0.8554 | 0.8572 | 0.8579 |

Table 10: Comparison of classification accuracies based on different methods with C4.5

| Data sets | raw data | $\gamma$ | $I$ | $GK$ |
|---|---|---|---|---|
| Credit | 0.8565±0.1852 | 0.8580±0.1829$\surd$ | 0.8464±0.1957 | 0.8580±0.1829$\surd$ |
| Horse | 0.9565±0.0479 | 0.8967±0.1314 | 0.9185±0.1139 | 0.9402±0.0818$\surd$ |
| Mushroom | 1.0000±0.0000 | 1.0000±0.0000$\surd$ | 1.0000±0.0000$\surd$ | 1.0000±0.0000$\surd$ |
| Letter | 0.8791±0.0102 | 0.8639±0.0116 | 0.8778±0.0106 | 0.8806±0.0103$\surd$ |
| Wdbc | 0.9332±0.0724 | 0.9438±0.0661$\surd$ | 0.9438±0.0661$\surd$ | 0.9438±0.0661$\surd$ |
| Wpbc | 0.7323±0.2980 | 0.7273±0.3027 | 0.7020±0.3280 | 0.7576±0.2842$\surd$ |
| Average | 0.7654 | 0.7557 | 0.7555 | 0.7686 |

tigation on feature selection based on the proposed evaluation function with that based on classical measures have been carried out in this paper. It is verified there the proposed approach leads to more data sets with higher average classification accuracy of feature selection. In addition, with the new technique, it is not necessary to distinguish whether a decision system is consistent or not in advance.

## Acknowledgements

The authors would like to thank reviewers for their valuable suggestion in revising this paper.

This paper is supported by the National Natural Science Foundation of China under Grant 10771043 and 60703013, the Research Fund from National Defence Key Laboratory of Autonomous Underwater Vehicle Technology under Grant 002010260730 and the Support Project for Young Scholars in General Institutions of Higher Learning of Heilongjiang Province under Grant 1151G076.

## References

1. E. Amaldi and V. Kann, "On the approximation of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, **209(1–2)**, 237–260 (1998).

2. J.A. Antonino-Daviu, M. Riera-Guasp, M. Pineda-Sanchez, J. Pons-Llinares, R. Puche-Panadero and J. Perez-Cruz, "Feature extraction for the prognosis of electromechanical faults in electrical machines through the DWT," *International Journal of Computational Intelligence Systems*, **2(2)**, 158–167 (2009).
3. M. Banerjee and S.K. Pal, "Roughness of a fuzzy set," *Information Sciences*, **93(3-4)**, 235–246 (1996).
4. R.B. Bhatt and M. Gopal, "On fuzzy-rough sets approach to feature selection," *Pattern Recognition Letters*, **26(7)**, 965–975 (2004).
5. T.Q. Deng, Y.M. Chen, W.L. Xu and Q.H. Dai, "A novel approach to fuzzy rough sets based on a fuzzy covering," *Information Sciences*, **177(11)**, 2308–2326 (2007).
6. T.Q. Deng and H.J.A.M. Heijmans, "Grey-scale morphology based on fuzzy logic," *Journal of Mathematical Imaging and Vision*, **16(2)**, 155–171 (2002).
7. P.A. Estevez, M. Tesmer, C.A. Perez and J. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, **20(2)**, 189–201 (2009).
8. S. Foitong, P. Rojanavasu, B. Attachoo and O. Pingern, "Estimating optimal feature subsets using mutual information feature selector and rough sets," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **5476** LNAI, 973–980 (2009).
9. P. Fortemps, S. Greco and R. Slowinski, "Multicriteria decision support using rules that represent rough-graded preference relations," *Fuzzy Sets and Systems*, **188(1)**, 206–223 (2008).
10. M. Gaag, T. Hoffman, M. Remijsen, R. Hijman, L. Haan, B. Meijel, P. Harten, L. Valmaggia, M. Hert, A. Cuijpers and D. Wiersma, "The five-factor model of the positive and negative syndrome scale ii: A tenfold cross-validation of a revised model," *Schizophrenia Research*, **85(1–3)**, 280–287 (2006).
11. X. Gong, Y. Yang, J. Lin and T. Li, "Expression detection based on a novel emotion recognition method," *International Journal of Computational Intelligence Systems*, **4(1)**, 44–53 (2011).
12. X. Hao, H. Fu and K. Shi, "S-rough sets and the discovery of f-hiding knowledge," *Journal of Systems Engineering and Electronics*, **19(6)**, 1171–1177 (2008).
13. Q. Hu, Z. Xie and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognition*, **40**, 1825–1844 (2007).
14. Q. Hu, D. Yu, J. Liu and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Information Sciences*, **178**, 3577–3594 (2007).
15. Q. Hu, D. Yu and Z. Xie, "Neighborhood classifiers," *Expert Systems with Applications*, **34(2)**, 866–876 (2008).
16. Q. Hu, M. Guo, D. Yu and J. Liu, "Information entropy for ordinal classification," *Science in China Series F: Information Sciences*, **53 (6)**, 1188–1200 (2010).
17. R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, **17(4)**, 824–838 (2009).
18. M.C. Lee, "Using support vector machine with a hybrid feature selection method to the stock trend prediction," *Expert Systems with Applications*, **36(8)**, 10896–10904 (2009).
19. T. Li, D. Ruan, W. Geert, J. Song and Y. Xu, "A rough sets based characteristic relation approach for dynamic attribute generalization in data mining," *Knowledge-Based Systems*, **20(5)**, 485–494 (2007).
20. J. Liang and Z. Shi, "The information entropy, rough entropy and knowledge granulation in rough set theory," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **12(1)**, 37–46 (2004).
21. J. Liang, J. Wang and Y. Qian, "A new measure of uncertainty based on knowledge granulation for rough sets," *Information Sciences*, **179(4)**, 458–470 (2009).
22. Y. Lin, T. Wu, S. Huang, Y. Meng and W. Liang, "Rough sets as a knowledge discovery and classification tool for the diagnosis of students with learning disabilities," *International Journal of Computational Intelligence Systems*, **4(1)**, 29–43 (2011).
23. H. Liu, J. Sun, L. Liu and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, **42(7)**, 1330-1339 (2009).
24. D. Liu, Y.Y. Yao and T.R. Li, "Three-way investment decisions with decision-theoretic rough sets," *International Journal of Computational Intelligence Systems*, **4(1)**, 66–74 (2011).
25. S. Nanda and S. Majumdar, "Fuzzy rough sets," *Fuzzy Sets and Systems*, **45(2)**, 157–160 (1992).
26. L. Nanni and A. Lumini, "Ensemble generation and feature selection for the identification of students with learning disabilities," *Expert Systems with Applications*, **36(2)**, 3896–3900 (2009).
27. D.J. Newman, S. Hettich, C.L. Blake and C.J. Merz, "UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, 1998," <http://www.ics.uci.edu/ mlearn/mlrepository.html>.
28. N.M. Parthalain and Q. Shen, "Exploring the boundary region of tolerance rough sets for feature selection," *Pattern Recognition*, **42(5)**, 655–667 (2009).
29. Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, **11(5)**, 341–356 (1982).
30. Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers,

Dordrecht, Boston, London (1991).
31. Z. Pawlak, "Some remarks on conflict analysis," *European Journal of Operational Research*, **166(3)**, 649–954 (2005).
32. A. Petrosino and A. Ferone, "Rough fuzzy set-based image compression," *Fuzzy Sets and Systems*, **160(10)**, 1485–1506 (2009).
33. L. Polkowski and P. Artiemjew, "On knowledge granulation and applications to classifier induction in the framework of rough mereology," *International Journal of Computational Intelligence Systems*, **2(4)**, 315–331 (2009).
34. Y. Qian, C. Dang, J. Liang, H. Zhang and J. Ma, "On the evaluation of the decision performance of an incomplete decision table," *Data and Knowledge Engineering*, **65**, 373–400 (2008).
35. Y. Qian, J. Liang, D. Li, H. Zhang and C. Dang, "Measures for evaluating the decision performance of a decision table in rough set theory," *Information Sciences*, **178(1)**, 181–202 (2008).
36. D. Ruan, G. Chen, E. Kerre and G. Wets, *Intelligent Data Mining: Technique and Applications*, Springer-Verlag, Heidelberg, 2005.
37. L. Sanchez, M. Suarez, J. Villar and I. Couso, "Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data," *International Journal of Approximate Reasoning*, **49(3)**, 607–622 (2008).
38. Q. Shen and R. Jensen, "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring," *Pattern Recognition*, **39(7)**, 1351–1363 (2004).
39. A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," *R. Slowinski (Ed.), Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, Kluwer, Dordrecht*, (1992).
40. D. Slézak, "Degrees of conditional (in)dependence: A framework for approximate bayesian networks and examples related to the rough set-based feature selection," *Information Sciences*, **179(3)**, 197–209 (2009).
41. R.W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, **24(6)**, 833–849 (2003).
42. Y.C. Tsai, C.H. Cheng and J.R. Chang, "Entropy-based fuzzy rough classification approach for extracting classification rules," *Expert Systems with Applications*, **31(2)**, 436–443 (2005).
43. S. Upadhyaya, A. Arora and R. Jain, "Reduct driven pattern extraction from clusters," *International Journal of Computational Intelligence Systems*, **2(1)**, 10–16 (2009).
44. C. Wang, C. Wu, D. Chen, Q. Hu and C. Wu, "Communicating between information," *Information Sciences*, **178**, 3228–3239 (2008).
45. J. Wang and J. Wang, "Reduction algorithms based on discernibility matrix: the ordered attributes method," *Journal of Computer Science and Technology*, **16**, 489–504 (2001).
46. S.K.M. Wong and W. Ziarko, "On optimal decision rules in decision tables," *Bulletin of Polish Academy of Sciences*, **33**, 693–696 (1985).
47. Wei-Zhi Wu, "Attribute reduction based on evidence theory in incomplete decision systems," *Information Sciences*, **178(5)**, 1355–1371 (2008).
48. F.F. Xu, D.Q. Miao and L. Wei, "Fuzzy-rough attribute reduction via mutual information with an application to cancer classification," *Computers and Mathematics with Applications*, **57**, 1010–1017 (2009).
49. Y.Y. Yao and Y. Zhao, "Discernibility matrix simplification for constructing attribute reducts," *Information Sciences*, **179**, 867–882 (2009).
50. L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, **5**, 1205–1224 (2004).
51. L.A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, **90**, 111–127 (1997).
52. W. Zakowski, "Approximations in the space $(U, \Pi)$," *Demonstratio Mathematica*, **16**, 761–769 (1983).
53. D. Zhang, Y. Wang, and H. Huang, "Rough neural network modeling based on fuzzy rough model and its application to texture classification," *Neurocomputing*, **72(10–12)**, 2433–2443 (2008).
54. W. Ziarko, "Variable precision rough set model," *Journal of Computer and System Sciences*, **46**, 39–59 (1993).