第38卷第10期
2008年10月

中国科学技术大学学报
JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Vol. 38，No. 10
Oct. 2008

Article ID：0253-2778(2008)10-1202-09

Feature reduction based on boundary conditional entropy and its application in qualitative simulation

CHENG Yu-sheng[1,2]，ZHANG You-sheng[2]， HU Xue-gang[2]

(1. *School of Computer Science*，*Anqing Teachers College*，*Anqing* 246011，*China*；
2. *School of Computer Science*，*Hefei University of Technology*，*Hefei* 230009，*China*)

Abstract：Some new definitions of knowledge rough entropy and boundary conditional entropy were given from the aspect of Pawlak topology. These definitions accurately reflect an idea that set uncertainty can be described by boundary region，which will measure knowledge and rough set uncertainty more accurately. Meanwhile，an important conclusion was that boundary conditional entropy of knowledge monotonously reduced with the diminishing of information granularity. Through an example of spring qualitative simulation reasoning technology combined with knowledge information entropy based on rough sets theory，a heuristic algorithm for feature reduction was proposed that could be used to eliminate the redundancy in the qualitative description and the qualitative differential equations were obtained from the spring physical system. Experimental results show that the rough sets theory is of good reliability and prospect in qualitative reasoning and simulation and that our algorithm is an effective method.

Key words：qualitative simulation；rough sets theory；feature reduction；boundary conditional entropy

CLC number：TP18　　**Document code**：A

边界条件熵的属性约简及在定性仿真中的应用

程玉胜[1,2]，张佑生[2]，胡学钢[2]

（1.安庆师范学院计算机与信息学院，安徽安庆 246011；2.合肥工业大学计算机与信息学院，安徽合肥 230009）

摘要：从Pawlak拓扑的角度，给出了一种知识边界粗糙熵和边界条件熵的新定义，并反映出集合的不确定性可以通过边界域来描述的思想，证明了边界条件熵随着信息粒度的变小而单调减少的重要结论.弹簧定性仿真实例，结合定性推理技术，以边界条件熵为基础构造属性约简的启发式算法，消去定性描述中的冗余，获得了弹簧系统定性微分方程式.实验结果表明，粗集理论在定性推理与定性仿真技术中的重要应用价值，基于边界条件熵的属性约简是有效的.

关键词：定性仿真；粗集理论；属性约简；边界条件熵

Received：2006-10-23；**Revised**：2007-02-01

Foundation item：Suported by Nature Science Fourdation of Anhui Province (070412061).

Biography：CHENG Yu-sheng (corresponding author)，male，born in 1969，Ph. D/assistant professor. Research field：Rough set theory data mining. E-mail：chengysaq@163. com

0 Introduction

Rough sets theory (RST), as a new mathematical tool to deal with inexact, uncertain knowledge, has been successfully employed in machine learning, data mining and other fields since it was put forward by Pawlak[1]. It is established on the basis of classification mechanism, which takes classification according to equivalence relation[2]. On the other hand, RST holds that knowledge has granularity. The smaller the granularity, the more concepts are precisely expressed. That is, the finer the classification, the smaller the granularity, and the more precise the knowledge will be. Therefore, knowledge and granulation are associated with equivalence relation.

Meanwhile, uncertainty and its measure have always been important issues in the study of RST[1~3]. Wierman[4] introduces the definition of granularity measure, connecting Shannon entropy[5] with uncertainty measure. Besides, MIAO[6] discusses the relation between knowledge roughness and information entropy, proving the monotony of knowledge rough entropy; WANG[7,8] defines the equivalence of feature reduction from the aspect of informational view and algebraic view of RST and provides a reduction algorithm of decision table based on conditional information entropy[8]. Liang [9] defines a new information entropy that can be better used for measuring rough sets and rough classification.

In the above study, knowledge rough entropy failed to show accurately the reason for conceptual uncertainty—the existence of boundary region[1~3]. The present paper defines a new knowledge rough entropy and conditional entropy based on boundary region, which can express set uncertainty more accurately, and is an attempt at solving measure uncertainty from the angle of set topology (Pawlak topology[2,3]). It provides a feature reduction algorithm of decision table based on boundary conditional entropy that will be used in qualitative reasoning and simulation.

Qualitative reasoning technology began in 1980s. The publication of de • Kleer's vision based on qualitative behavior [10], Forbus's qualitative process theory (QPT)[11] and Kuipers' qualitative simulation process (QSIM)[12] in Artificial Intelligence in 1984 symbolize the maturity of qualitative reasoning[13]. This technology holds that as long as the system's behavior is in accordance with physical rules, it is applicable to describe all possible states of behavior by using non-numerical values. Qualitative reasoning is to ignore the details and collect specific values of the system's variables at different times to simulate the system's behavior. But this method has a relatively bigger knowledge redundancy. Thus, it is advisable to delete the problem of knowledge redundancy by using feature reduction method in RST. The qualitative simulation of the spring physical system uses attribute significance as a heuristic algorithm for feature reduction together with the technology of qualitative reasoning and simulation. The result is in accordance with that of the qualitative differential equation with qualitative reasoning technology[13], which further explains the important practical value of RST in qualitative reasoning and qualitative simulation technology.

1 General meaning of conditional entropy of knowledge

An information system is usually denoted by a triplet $S=(U,C\cup D,f)$, called a decision table, where U is the universe which consists of a finite set of objects, C is the set of condition attributes and D is the set of decision attributes. With every attribute $a\in C\cup D$, a set of its values V_a is associated. Each attribute a determines an information function $f:U\rightarrow V_a$ such that for any $a\in C\cup D$, and $x\in U$, $f(x)\in V_a$. Each non-empty subset $B\subseteq C$ determines an indiscernible relation

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

R_B is called an equivalence relation and partitions U into a family of a disjoint subsets

U/R_B called a quotient set of U

$$U/R_B = \{[x]_B : x \in U\}$$

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

The sets $B_*(X) = \{x \in U | [x]_B \subseteq x\}$, $B^*(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}$ are called B-lower approximation and B-upper approximation respectively, where $X \subseteq U$. The boundary region of knowledge B is defined as $BN_B(X) = B^*(X) - B_*(X)$. Given a decision system $S = (U, C \cup D, f)$, $B \subseteq C$, partition of condition attributes $U/R_C = \{X_1, X_2, \cdots, X_n\}$, $U/R_B = \{Y_1, Y_2, \cdots, Y_k\}$ and partition of decision attributes $U/R_D = \{D_1, D_2, \cdots, D_m\}$. The set $B_*(D_1) \cup B_*(D_2) \cup \cdots \cup B_*(D_m)$ is called the B-positive region of classification induced by D and is denoted by $POS_B(D)$. The set $BN_B(D) = U - POS_B(D)$ is called the B-boundary of classification induced by D.

Definition 1. 1[6~9] The information entropy of knowledge B is defined as

$$H(B) = -\sum_{i=1}^{k} p(Y_i) \log_2 p(Y_i) \tag{1}$$

Definition 1. 2[8] Conditional information entropy of knowledge C with respect to D is defined as

$$H(D \mid C) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(D_j \mid X_i) \log_2 p(D_j \mid X_i) \tag{2}$$

where

$$p(X_i) = \frac{|X_i|}{|U|},\ p(D_j \mid X_i) = \frac{|X_i \cap D_j|}{|X_i|}$$

From formula (2), we get $H(D \mid C) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(X_i \cap D_j)[\log_2 p(D_j \cap X_i) - \log_2 p(X_i)]$, thus when $X_i \in POS_C(D)$, we have $\log_2 p(D_j \cap X_i) - \log_2 p(X_i) = 0$. Therefore, the positive region of the decision system has no effect on $H(D|C)$.

2 Conditional information entropy based on boundary region

According to the definition of set topology[2], set uncertainty is mainly caused by the existence of boundary region. If it is empty, then the set is accurate; otherwise, it is rough [1~3]. Therefore, it is quite reasonable to describe knowledge uncertainty by boundary region.

2.1 Boundary rough entropy and conditional entropy of knowledge

Decision system $S = (U, C \cup D, f)$, $P, Q \subseteq C$, defines partial order relation $\lesssim$: $P \lesssim Q \Leftrightarrow U/R_p \subseteq U/R_Q$, then P is more refined than Q (or: Q is rougher than P). If $P \lesssim Q$, and $P \neq Q$, then P is strictly more refined than Q (or: Q is strictly rougher than P), shown as $P \prec Q$.

Definition 2. 1 Decision system $S = (U, C \cup D, f)$, P, $Q \subseteq C$, the partition of condition attributes P is $U/R_P = \{X_1, X_2, \cdots, X_m\}$ and P's boundary region against knowledge Q is $BN_Q(P)$, the corresponding classification is $BN_Q(P)/Q = \{G_1, G_2, \cdots, G_t\}$, then P's boundary entropy against Q and P's boundary conditional entropy against Q are defined as follows respectively

$$E_{BN}(Q) = \sum_{i=1}^{t} p(G_i) \log_2 |G_i|$$

$$E_{BN}(P \mid Q) = -\sum_{i=1}^{t} p(G_i) \sum_{j=1}^{m} p(X_j \mid G_i) \log_2 p(X_j \mid G_i)$$

Lemma 2. 1 Decision system $S = (U, C \cup D, f)$, $P, Q \subseteq C, X \subseteq U$. If $P \lesssim Q$, then $BN_P(X)/P \lesssim BN_Q(X)/Q$.

Proof Order $U/P = \{X_1, X_2, \cdots, X_m\}$, $U/Q = \{Z_1, Z_2, \cdots, Z_n\}$. It is easily known that $BN_P(X)/P \subseteq U/P$, $BN_Q(X)/Q \subseteq U/Q$. Now to prove: $BN_P(X)/P \subseteq BN_Q(X)/Q$.

Order $\forall X_i \subseteq BN_P(X)/P$. From $BN_P(X)/P \subseteq U/P$, we know $X_i \subseteq U/P$. Because of $P \lesssim Q$, so $Z_j \in U/Q$ and $X_i \subseteq Z_j$.

On the other hand, because $X_i \subseteq BN_P(X)/P$, $X_i \cap X \neq \emptyset$ and $X_i \not\subset X$. From $X_i \subseteq Z_j$, we have $Z_j \cap X \neq \emptyset$, $Z_j \not\subset X$, thus $Z_j \subseteq BN_Q(X)/Q$, so $X_i \subseteq BN_Q(X)/Q$.

That is: $BN_P(X)/P \subseteq BN_Q(X)/Q$ is proved, so $BN_P(X)/P \lesssim BN_Q(X)/Q$.

Proposition 2. 1 Decision system $S = (U, C \cup$

$D,f)$, $P,Q\subseteq C$, $X\subseteq U$. If $P\lesssim Q$, then $E_{BN}(P)\leqslant E_{BN}(Q)$.

Proof Because $P\lesssim Q$, $BN_P(X)/P\lesssim BN_Q(X)/Q$. Order $BN_P(X)/P=\{Y_1,Y_2,\cdots,Y_k\}$, $BN_Q(X)/Q=\{G_1,G_2,\cdots,G_t\}$. Because $BN_P(X)/P\lesssim BN_Q(X)/Q$, for $\forall Y_i$, we get $Y_i\subseteq G_j$ and $k\geqslant t$. Order $x=|Y_i|$, $y=|G_j|$, so $1\leqslant x\leqslant y$.

Besides

$$E_{BN}(Q)=\sum_{j=1}^{t}p(G_j)\log_2|G_j|=$$
$$\frac{1}{|U|}\sum_{j=1}^{t}|G_j|\log_2|G_j|=$$
$$\frac{1}{|U|}\sum_{j=1}^{k}|G_j|\log_2|G_j|=$$
$$\frac{1}{|U|}\sum_{j=1}^{k}y\log_2 y$$

Order $f_j(x)=x\log_2 x$, because $f_j'(x)=1/\ln 2+\log_2 x>0$, $f_j(x)$ is an increasing function, therefore, $E_{BN}(Q)=\frac{1}{|U|}\sum_{j=1}^{k}y\log_2 y\geqslant\frac{1}{|U|}\sum_{i=1}^{k}x\log_2 x=E_{BN}(P)$. That's $E_{BN}(P)\leqslant E_{BN}(Q)$.

From proposition 2.1, we know that if knowledge Q is rougher than P, then Q's boundary region is not smaller than P's.

Proposition 2.2 Decision system $S=(U,C\cup D,f)$, $P,Q\subseteq C$, $U/R_P=\{X_1,X_2,\cdots,X_m\}$, $BN_Q(P)/Q=\{G_1,G_2,\cdots,G_t\}$, then $E_{BN}(Q\cup P)=E_{BN}(P|Q)-E_{BN}(Q)$.

Proof

$$E_{BN}(Q\cup P)=$$
$$-\sum_{i=1}^{t}\sum_{j=1}^{m}p(X_j\cap G_i)\log_2 p(X_j\cap G_i);$$
$$E_{BN}(P\mid Q)=$$
$$-\sum_{i=1}^{t}p(G_i)\sum_{j=1}^{m}p(X_j\mid G_i)\log_2 p(X_j\mid G_i)=$$
$$-\sum_{i=1}^{t}\sum_{j=1}^{m}p(X_j\cap G_i)\cdot$$
$$[\log_2 p(X_j\cap G_i)-\log_2 p(G_i)]=$$
$$-\sum_{i=1}^{t}\sum_{j=1}^{m}p(X_j\cap G_i)\log_2 p(X_j\cap G_i)+$$
$$\sum_{i=1}^{t}\sum_{j=1}^{m}p(X_j\cap G_i)\log_2 p(G_i).$$

Additionally because $\sum_{j=1}^{m}p(X_j\cap G_i)=p(G_i)$, so $E_{BN}(P|Q)=E_{BN}(Q\cup P)+\sum_{i=1}^{t}p(G_i)\log_2 p(G_i)=E_{BN}(Q\cup P)+E_{BN}(Q)$. That's $E_{BN}(Q\cup P)=E_{BN}(P|Q)-E_{BN}(Q)$.

Proposition 2.3 Decision system $S=(U,C\cup D,f)$, $A,B\subseteq C$, If $A\lesssim B$, then $E_{BN}(D|A)\leqslant E_{BN}(D|B)$.

Proof Order $U/R_D=\{D_1,D_2,\cdots,D_m\}$, $BN_A(D)=\{G_1,G_2,\cdots,G_t\}$.

Because $A\lesssim B$, then $BN_A(D)\subseteq BN_B(D)$. According to lemma 2.1, we have $BN_A(D)/A\lesssim BN_B(D)/B$. Suppose $BN_B(D)/B=\{G_1,G_2,\cdots,G_{p-1},G_{p+1},\cdots,G_{q-1},G_{q+1},\cdots,G_t,G_p\cup G_q\}$. According to proposition 2.2

$$E_{BN}(D\mid A)=$$
$$-\sum_{i=1}^{m}\sum_{j=1}^{t}p(D_i\cap G_j)\log_2 p(D_i\cap G_j)+$$
$$\sum_{j=1}^{t}p(G_j)\log_2 p(G_j)$$
$$E_{BN}(D\mid B)=$$
$$E_{BN}(D\mid A)-\sum_{i=1}^{m}p[(G_p\cup G_q)\cap D_i]\cdot$$
$$\log_2 p[(G_p\cup G_q)\cap D_i]+$$
$$p(G_p\cup G_q)\log_2 p(G_p\cup G_q)+$$
$$\sum_{i=1}^{m}p(G_p\cap D_i)\log_2 p(G_p\cap D_i)-$$
$$p(G_p)\log_2 p(G_p)+$$
$$\sum_{i=1}^{m}p(G_q\cap D_i)\log_2 p(G_q\cap D_i)-$$
$$p(G_q)\log_2 p(G_q)$$

So

$$\Delta E=E_{BN}(D\mid B)-E_{BN}(D\mid A)=$$
$$-\sum_{i=1}^{m}p[(G_p\cup G_q)\cap D_i]\cdot$$
$$\log_2 p[(G_p\cup G_q)\cap D_i]+$$
$$p(G_p\cup G_q)\log_2 p(G_p\cup G_q)+$$
$$\sum_{i=1}^{m}p(G_p\cap D_i)\log_2 p(G_p\cap D_i)-$$
$$p(G_p)\log_2 p(G_p)+$$
$$\sum_{i=1}^{m}p(G_q\cap D_i)\log_2 p(G_q\cap D_i)-$$

$$p(G_q)\log_2 p(G_q)$$

Additionally because $\sum_{i=1}^{m} p(D_i \cap G_p) = p(G_p)$, $\sum_{i=1}^{m} p(D_i \cap G_q) = p(G_q)$. Thus

$$\begin{aligned}\Delta E = &-\sum_{i=1}^{m} p[(G_p \cup G_q) \cap D_i]\cdot \\ &\log_2 p[(G_p \cup G_q) \cap D_i] + \\ &\sum_{i=1}^{m} p[(G_p \cup G_q) \cap D_i]\log_2 p(G_p \cup G_q) + \\ &\sum_{i=1}^{m} p(G_p \cap D_i)\log_2 p(G_p \cap D_i) - \\ &\sum_{i=1}^{m} p(G_p \cap D_i)\log_2 p(G_p) + \\ &\sum_{i=1}^{m} p(G_q \cap D_i)\log_2 p(G_q \cap D_i) - \\ &\sum_{i=1}^{m} p(G_q \cap D_i)\log_2 p(G_q) = \\ &\sum_{i=1}^{m} p(G_p \cap D_i)\{\log_2 p(G_p \cap D_i) + \\ &\log_2 p(G_p \cap G_q) - \\ &\log_2 p(G_p) - \log_2 p[(G_p \cup G_q) \cap D_i]\} + \\ &\sum_{i=1}^{m} p(G_q \cap D_i)\{\log_2 p(G_q \cap D_i) + \\ &\log_2 p(G_p \cap G_q) - \\ &\log_2 p(G_q) - \log_2 p[(G_p \cup G_q) \cap D_i]\} = \\ &\frac{1}{|U|}\sum_{i=1}^{m}\Big\{|G_p \cap D_i|\cdot \\ &\log_2 \frac{|G_p \cap D_i||G_p \cup G_q|}{|G_p|(|G_p \cap D_i|+|G_q \cap D_i|)} + \\ &|G_q \cap D_i|\cdot \\ &\log_2 \frac{|G_q \cap D_i||G_p \cup G_q|}{|G_q|(|G_p \cap D_i|+|G_q \cap D_i|)}\Big\}.\end{aligned}$$

Order $|G_p|=x$, $|G_q|=y$, $|G_p \cap D_i|=ax$, $|G_q \cap D_i|=by$, obviously get $x>0$, $y>0$, $0\leqslant a\leqslant 1$, $0\leqslant b\leqslant 1$, then $\Delta E = \frac{1}{|U|}\sum_{i=1}^{m}\Big\{ax\log_2 \frac{ax+ay}{ax+by} + by\log_2 \frac{bx+by}{ax+by}\Big\} = \frac{1}{|U|}\sum_{i=1}^{m} f_i$.

If $a\times b=0$, get $f_i\geqslant 0$. $0<a\leqslant 1$, $0<b\leqslant 1$ shall be only considered in the following.

Order $ax=\lambda$, $by=\beta$, $\frac{a}{b}=\theta$, obviously get $\lambda>0$, $\beta>0$, $\theta>0$ and $f_i=\lambda\log_2 \frac{\lambda+\theta\beta}{\lambda+\beta}+\beta\log_2 \frac{\beta+\theta^{-1}\lambda}{\lambda+\beta}$, then $\frac{\mathrm{d}(f_i)}{\mathrm{d}(\theta)}=\frac{\lambda\beta(\theta-1)}{\theta(\lambda+\theta\beta)}$. So, $\frac{\mathrm{d}(f_i)}{\mathrm{d}(\theta)}<0$, $0<\theta<1$; $\frac{\mathrm{d}(f_i)}{\mathrm{d}(\theta)}=0$, $\theta=1$; $\frac{\mathrm{d}(f_i)}{\mathrm{d}(\theta)}>0$, $\theta>1$. When $\theta=\frac{a}{b}=1$, function f_i gets the minimal $f_i|_{\theta=1}=0$.

The above shows, when $\Delta E\geqslant 0$, $E_{BN}(D|A)\leqslant E_{BN}(D|B)$ is proved. The proposition shows that boundary conditional entropy of knowledge monotonously reduces with the diminishing of information granularity.

2.2 Several propositions of knowledge boundary conditional entropy

Some conclusions shall be given based on boundary conditional entropy.

Proposition 2.4 $S=(U,C,D)$ is a consistent decision information system iff $E_{BN}(D|C)=0$.

Proof (⇒) consistency of decision system S doesn't have a boundary region, thus $E_{BN}(D|C)=0$;

(⇐) order $U/R_D=\{D_1, D_2, \cdots, D_m\}$, $BN_C(D)/C=\{G_1,G_2,\cdots,G_t\}$. From $E_{BN}(D|C)=0$ surely there is $p(G_i)=0$ or $P(D_j|G_i)=1$. If $P(D_j|G_i)=1$, then $G_i\subseteq D_j$, that is, $G_i\subseteq POS_C(D)$, which is in contradiction with the formula $G_i\subseteq BN_C(D)$. Therefore, $p(G_i)=0$, i. e., $G_i=\varnothing$ means that the system doesn't have a boundary region. So, when $E_{BN}(D|C)=0$, $S=(U,C,D)$ is a consistent decision information system.

Proposition 2.5 Decision system $S=(U,C,D)$, $P\subseteq C$, $r\in P$. If $S'=(U,P,D)$ is consistent, then

$$\begin{aligned}POS_P(D) = POS_{P\backslash\{r\}}(D) \Leftrightarrow \\ E_{BN}(D\mid P) = E_{BN}(D\mid P\backslash\{r\}).\end{aligned}$$

Proof Because $P\subseteq C$, $S'=(U,P,D)$ is consistent, i. e., $POS_P(D)=U$, according to proposition 2.4 we have $E_{BN}(D|P)=0$. Order $BN_{P\backslash\{r\}}(D)/P\backslash\{r\}=\{G_1,G_2,\cdots,G_t\}$.

(⇒) because $POS_P(D)=POS_{P\backslash\{r\}}(D)=U$, then $BN_{P\backslash\{r\}}(D)=\varnothing$, surely there is $p(G_i)=0$, so $E_{BN}(D|P\backslash\{r\})=0$, then $E_{BN}(D|P)=E_{BN}(D|P\backslash\{r\})$.

(⇐) because $E_{BN}(D|P)=0$, $E_{BN}(D|P\backslash\{r\})=0$, according to proposition 2.4, we have

$BN_{P\setminus\{r\}}(D)=\emptyset$, that's $POS_{P\setminus\{r\}}(D)=U$, so $POS_P(D)=POS_{P\setminus\{r\}}(D)$.

Proposition 2.6 Decision system $S=(U,C,D)$, $P\subseteq C$, $r\in P$. If $S'=(U,P,D)$ is consistent, then

$$POS_P(D)\neq POS_{P\setminus\{r\}}(D)\Leftrightarrow E_{BN}(D\mid P)\neq E_{BN}(D\mid P\setminus\{r\})$$

Proof Because $P\subseteq C$, $S'=(U,P,D)$ is consistent, i. e., $POS_P(D)=U$, according to proposition 2.4 we have $E_{BN}(D|P)=0$, $POS_P(D)\neq POS_{P\setminus\{r\}}(D)\Leftrightarrow POS_{P\setminus\{r\}}(D)\neq U\Leftrightarrow BN_{P\setminus\{r\}}(D)\neq\emptyset\Leftrightarrow E_{BN}(D|P\setminus\{r\})\neq 0\Leftrightarrow E_{BN}(D|P)\neq E_{BN}(D|P\setminus\{r\})$.

3 Qualitative simulation and reasoning with feature reduction based on boundary conditional entropy of knowledge

Knowledge entropy based on boundary region definition and conditional information entropy are considered by observing the basic cause of knowledge uncertainty, so it can more accurately reflect the essence of the issue. In addition, the informational view of boundary conditional entropy includes the conclusion in the algebraic view. Thus, we can use boundary conditional entropy as heuristic knowledge for reduction algorithm.

3.1 Heuristic algorithm for feature reduction based on boundary conditional entropy

Definition 3.1 Decision system $S=(U,C\cup D,f)$, $B\subseteq C$, the significance of b in B with respect to D is defined as

$$\mathrm{Sig}_{B\setminus\{b\}}(D\mid\{b\})=E_{BN}(D\mid B\setminus\{b\})-E_{BN}(D\mid B)$$

We know such important conclusions as information entropy is monotonously decrease with the diminishing of information granularity. Because B is more refined than $B\setminus\{b\}$, therefore, $\mathrm{Sig}_{B\setminus\{b\}}(D\mid\{b\})\geqslant 0$. Definition 3.1 (attribute significance) shows that b is important in B, which can be measured based on the increment of boundary conditional entropy. Especially, when system $S=(U,C\cup D,f)$ is not a decision system, i. e., $D=\emptyset$, then we can consider $E_{BN}(\emptyset|B)=H(B)$.

Proposition 3.1 b is necessary in B when $\mathrm{Sig}_{B\setminus\{b\}}(D|\{b\})>0$.

Definition 3.2 Decision system $S=(U,C\cup D,f)$, $B\subset C$, $a\in C\setminus B$, the significance of a relative to B with respect to D is defined as

$$\mathrm{Sig}_B(D\mid\{a\})=E_{BN}(D\mid B)-E_{BN}(D\mid B\cup\{a\})$$

The above definition (attribute relative significance) shows that the greater change of boundary conditional entropy of knowledge caused by adding an attribute, the more relatively important this attribute is. Thus, it is possible to use attribute significance and attribute relative significance as heuristic knowledge for feature reduction.

Algorithm 3.1 KIEBAFR (knowledge information entropy-based algorithm for feature reduction).

Input: Decision system $S=(U,C,D)$;

Output: A feature reduction Red of decision system $S=(U,C,D)$.

Step 1 Calculate the boundary conditional information entropy $E_{BN}(D|C)$;

Step 2 For any $c\in C$, calculate the significance of c in C: $\mathrm{Sig}_{C\setminus\{c\}}(\{c\})$ and then obtain $\mathrm{Red}=\{c|\mathrm{Sig}_{C\setminus\{c\}}(\{c\})>0\}$;

Step 3 Repeat:

(Ⅰ) Calculate boundary conditional information entropy $E_{BN}(D|\mathrm{Red})$. If $E_{BN}(D|C)=E_{BN}(D|\mathrm{Red})$, output a reduction set Red and stop. Otherwise, continue (Ⅱ).

(Ⅱ) For each attribute $a\in C\setminus\mathrm{Red}$, calculate $\mathrm{Sig}_{\mathrm{Red}}(\{a\})$; select attribute a_0 to make $\mathrm{Sig}_{\mathrm{Red}}(\{a\})$ the maximal, and compute $\mathrm{Red}=\mathrm{Red}\cup\{a_0\}$, go to (Ⅰ).

Next to analyze the time complexity of the algorithm:

Step 1 For each $x\in U$, calculation of the equivalence is needed, which involves, comparison of a number of $|U|-1$ objects on $|C|$ attributes, thus time complexity is $O(|C||U|^2)$;

Step 2 Time complexity is $O(|C||U|^2)$ when $\mathrm{Sig}_{C\setminus\{c\}}(\{c\})$ is computed for $\{c\}$. In the worst condition, $|C|$ time need to be circulated, so time complexity is $O(|C|^2|U|^2)$.

Time complexity of (Ⅰ) is $O(|C|^2|U|^2)$, in the worst condition, $|C|$ time need to be circulated, so time complexity is $O(|C|^3|U|^2)$. So, the time complexity of the algorithm is $O(|C|^3|U|^2)$.

3.2 Heuristic algorithm for feature reduction based on conditional entropy [8]

Definition 1.2 (conditional entropy) shows that B is important in C, which can be measured based on the increment of conditional entropy $H(D|B)$. Therefore, WANG[8] presents two heuristic knowledge reduction algorithms based on conditional information entropy, that is, a conditional entropy based algorithm for reduction of knowledge with computing core (CEBARKCC) and a conditional entropy based algorithm for reduction of knowledge without computing core (CEBARKNC). CEBARKNC is given as follows.

Algorithm 3.2 CEBARKNC

Input: Decision system $S=(U,C,D)$;

Output: A feature reduction Red of decision system $S=(U,C,D)$.

Step 1 Calculate the conditional information entropy $H(D|C)$;

Step 2 For any $c\in C$, calculate the significance of $\{c\}$ in C: and then sort descending according to the conditional entropy of $H(D|\{c\})$;

Step 3 Order $B=C$, for any $c\in C$, select attribute $\{c\}$ in descending order of importance of $H(D|\{c\})$, repeat:

(Ⅰ) Calculate conditional information entropy $H(D|B\setminus\{c\})$ when attribute $\{c\}$ is removed from B;

(Ⅱ) If $H(D|C)=H(D|B\setminus\{c\})$ then $\{c\}$ is a redundant attribute, so $B=B\setminus\{c\}$; otherwise, $\{c\}$ is necessary in B, so $B=B$.

The time complexity of CEBARKNC is $O(|U|^3)$ shown in Ref. [8]. So the time complexity of KIEBAFR algorithm is less than CEBARKNC algorithm.

3.3 Example analysis

For example, as shown in Tab. 1, a reduction set $\{a, e\}$ is obtained by CEBARKNC and CEBARKCC[8]:

Tab. 1 A decision system

U	a	b	c	e	d
1	1	0	1	1	0
2	0	1	0	1	1
3	0	0	0	0	0
4	0	0	0	1	1
5	0	0	0	1	1
6	0	0	0	1	1
7	0	0	1	1	1
8	0	0	1	0	0
9	0	0	1	1	1

The decision classes of objects are: $D_1=\{1,3,8\}$, $D_2=\{2,4,5,6,7,9\}$. The condition classes of objects are: $X_1=\{1\}$, $X_2=\{2\}$, $X_3=\{3\}$, $X_4=\{4,5,6\}$, $X_5=\{7,9\}$, $X_6=\{8\}$.

According to CEBARKNC, compute $H(D|C)$ first

$$
\begin{aligned}
H(D\mid C)=&-\Big\{\frac{1}{9}\Big(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\Big)-\\
&\frac{1}{9}\Big(\frac{0}{1}\log_2\frac{0}{1}+\frac{1}{1}\log_2\frac{1}{1}\Big)-\\
&\frac{1}{9}\Big(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\Big)-\\
&\frac{3}{9}\Big(\frac{0}{3}\log_2\frac{0}{3}+\frac{3}{3}\log_2\frac{3}{3}\Big)-\\
&\frac{2}{9}\Big(\frac{0}{2}\log_2\frac{0}{2}+\frac{2}{2}\log_2\frac{2}{2}\Big)\Big\}-\\
&\frac{1}{9}\Big(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\Big)=0
\end{aligned}
$$

Then compute $H(D|\{a\})$

$$
\begin{aligned}
H(D\mid\{a\})=&-\Big\{\frac{1}{9}\Big(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\Big)-\\
&\frac{8}{9}\Big(\frac{2}{8}\log_2\frac{2}{8}+\frac{6}{8}\log_2\frac{6}{8}\Big)\Big\}=0.217
\end{aligned}
$$

Similarly, we can get $H(D|\{b\})=0.217$, $H(D|\{c\})=0.255$, $H(D|\{e\})=0.518$, then $H(D|C\setminus\{e\})=0.198$, $H(D|C\setminus\{c\})=0$, $H(D|C\setminus\{b\})=0$, $H(D|C\setminus\{a\})=0.09$, so we can get $B=\{a,e\}$.

From definition 1.2, if $X_i\in POS_C(D)$, then the positive region of the decision system has no effect on $H(D|C)$. Therefore, many equations are computed in conditional entropy just like $H(D|C)$. Our method, that is, boundary conditional entropy doesn't need to compute the classes X_i if $X_i\in POS_C(D)$.

According to our method KIEBAFR, because

$BN_C(D)/C=\emptyset$, then easily get $E_{BN}(D|C)=0$. Next, we calculate the significance of $\{e\}$ in attributes C. Because $BN_{C\setminus\{e\}}(D)/C\setminus\{e\}=\{\{3,4,5,6\},\{7,8,9\}\}$, therefore

$$\mathrm{Sig}_{C\setminus\{e\}}(D\mid\{e\})=-\left\{\frac{4}{9}\left(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4}\right)+\frac{3}{9}\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)\right\}=0.198$$

Similarly, we can get $\mathrm{Sig}_{C\setminus\{c\}}(D\mid\{c\})=0$; $\mathrm{Sig}_{C\setminus\{b\}}(D|\{b\})=0$; $\mathrm{Sig}_{C\setminus\{a\}}(D\mid\{a\})=0.09$, so $\mathrm{Red}=\{a,e\}$. Compute $E_{BN}(D|\mathrm{Red})=0$, because $E_{BN}(D|C)=E_{BN}(D|\mathrm{Red})$, output a result: $\{a,e\}$, stop.

3.4 Application of KIEBAFR in qualitative simulation and reasoning

Take the qualitative simulation of physical system of spring by example[13], four variables can be described as follows: (Ⅰ) x, means the position of the object; (Ⅱ) v, means the velocity of the object: $v=\mathrm{d}x/\mathrm{d}t$; (Ⅲ) a, means the acceleration of the object: $a=\mathrm{d}v/\mathrm{d}t$; (Ⅳ) f, means the strength exerted by the pulling object. The qualitative analysis obtains knowledge expression system for qualitative description of spring physical system as shown in Tab. 2[13], i. e., $S=(U,C=\{[x],[f],[a],[v]\}$.

Tab. 2　A qualitative descriptive knowledge system

U	$[x]$	$[f]$	$[a]$	$[v]$
S1	+	−	−	+
S2	+	−	−	0
S3	+	−	−	−
S4	0	0	0	+
S5	0	0	0	0
S6	0	0	0	−
S7	−	+	+	+
S8	−	+	+	0
S9	−	+	+	−

The qualitative differential equations of the spring physical system can be obtained by KIEBAR algorithm. i. e., $[f]=[a]$, $[f]=[x]$, $[a]=[x]$.

The explanation is as follows: In the qualitative expression information system, $\{[v],[x]\}$ is the reduction of the original qualitative expression system (Fig. 1), which shows that it makes no difference to the classification ability of the original knowledge expression system whether to delete $[a]$'s or $[f]$'s attribute, so $[a]$ and $[f]$ have a consistent effect on the information system, marked as $[f]=[a]$, and the first qualitative differential equation is obtained, so $[f]=[x]$, $[a]=[x]$. The result is in accordance with that of the qualitative differential equation after the qualitative calculation of $f=ma$ (m is the mass of the object) and $f=-kx$ (k is the modulus of spring flexibility)[13]. In other words, as long as the knowledge expression system accords with physical rules, we can surely obtain the qualitative differential equation by analyzing the state of the physical system, constructing the qualitative expression system as well as using the method of RST, even if the qualitative equation is unknown; Vice versa, the qualitative differential equation can also be a guidance to estimate the qualitative equation of the system (when it is unknown). To summarize, RST is a powerful method in data mining.

4 Experiment result

The experiments are performed on several different real-life data sets obtained from UCI. Notice that some of the data were discretized by

Tab. 3　Experiment result of CEBARKNC and KIEBAFR algorithm

UCI	condition attribute numbers	condition attribute numbers after reduction	record numbers	CEBARKNC/s	KIEBAFR/s
iris	4	3	150	5.391	4.031
liver-disorders	6	3	345	38.423	31.922
mushroom	21	7	100	1.953	1.484
zoo	16	14	101	2.172	1.875
balloons	4	3	20	0.140	0.094
letter	16	13	800	1 021.997	881.859
vehicle	18	3	846	832.563	291.547

【Note】 Experimental environment: Windows XP, P1.6GHz, 512M, Matlab 6.5

Rosetta, a rough set toolkit. Simulation result (see Tab. 3) shows that KIEBARF algorithm is more efficient than CEBARKNC.

5 Conclusion

The existence of the boundary region is the major cause of set uncertainty. The information entropy and rough set entropy in general meaning can't explain it clearly. Based on this, the present paper puts forward the definition of knowledge boundary rough entropy and boundary conditional entropy, and describes some algebraic views in RST by using the method of boundary conditional entropy, establishes the connection with the algebraic view of RST. These important conclusions also guarantee the feature reduction algorithm based on boundary conditional entropy. Qualitative simulation of the spring physical system shows that RST is a powerful method for data mining and of good reliability and prospect in qualitative reasoning and qualitative simulation.

References

[1] Pawlak Z. Rough Sets: Theoretical Aspects of reasoning about data[M]. Boston: Kluwer Academic Publishers, 1991.

[2] 张文修,吴伟志,等. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[3] 李德毅,杜鹢. 不确定性人工智能[M]. 北京:国防工业出版社,2005.

[4] Wierman M J. Measuring uncertainty in rough set theory[J]. International Journal of General Systems, 1999, 28(4):283-297.

[5] Shannon C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948,27(7): 373-423, 623-656.

[6] MIAO D Q, WANG J. An information representation of the concepts and operations in rough set theory[J]. Journal of Software,1999,10(2):113-116.
苗夺谦,王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报,1999,10(2):113-116.

[7] WANG G Y. Algebra view and information view of rough set theory[C]//Proceeding of SPIE,2001, 4384: 200-207.

[8] WANG G Y, YU H, YANG D C. Decision table reduction based on conditional information entropy[J]. Chinese journal of Computers, 2002, 25(7):759-766.
王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7): 759-766.

[9] Liang J Y, Dang C Y, Chin K S, et al. A new method for measuring of rough sets and rough relational databases[J]. Information Sciences, 2002, 31(4): 331-342.

[10] de Kleer J, Brown J S. A qualitative physics based on confluences[J]. Artificial Intelligence, 1984, 24(1-3): 7-83.

[11] Forbus K D. Qualitative process theory[J]. Artificial Intelligence,1984, 24(1-3): 85-168.

[12] Kuipers B J. Qualitative simulation [J]. Artificial Intelligence,1986, 29(3): 289-338.

[13] 石纯一,廖士中. 定性推理方法[M]. 北京:清华大学出版社,2002.