基于广义粗集覆盖约简的粗糙熵*

黄 兵+, 何 新, 周献中

(南京理工大学 自动化系,江苏 南京 210094)

Rough Entropy Based on Generalized Rough Sets Covering Reduction

HUANG Bing+, HE Xin, ZHOU Xian-Zhong

(Department of Automation, Nanjing University of Science and Technology, Nanjing 210094, China)

+ Corresponding author: Phn: +86-25-84315460, Fax: +86-25-84315460, E-mail: hbhuangbing@sohu.com

Received 2002-05-10; Accepted 2002-09-20

Huang B, He X, Zhou XZ. Rough entropy based on generalized rough set covering reduction. *Journal of Software*, 2004,15(2):215~220.

http://www.jos.org.cn/1000-9825/15/215.htm

Abstract: In generalized rough set covering reduction theory, it is necessary to find a new measure to knowledge and rough set because the upper and lower approximations of rough sets are determined by their covering reduction. In this paper, information entropy is introduced to discuss the rough entropy of knowledge and the roughness of rough set, based on generalized rough set covering reduction. A new kind of measurement about the roughness of knowledge and rough set is presented. The conclusion that the rough entropy of knowledge and rough set decreases monotonously as the generalized rough set covering reduction becomes finer is obtained. This paper presents some useful exploration about the incomplete information system from information views.

Key words: rough set; information entropy; reduction; covering

摘 要: 在广义粗集覆盖约简理论中,由于集合的上下近似是由其覆盖约简来确定的,因此有必要寻求一种新的度量来刻画知识和粗集的粗糙性.通过引入信息熵以刻画广义粗集覆盖约简的知识粗糙性以及粗集粗糙性,提出了一种新的知识粗糙性和粗集粗糙性度量.得到知识粗糙熵和粗糙集的粗糙熵都随广义覆盖约简的变细而单调减少的结论,从信息论观点出发,对不完备信息系统粗集理论进行了探讨.

关键词: 粗集;信息熵;约简;覆盖

中图法分类号: TP18 **文献标识码**: A

* Supported the Core Teachers Foundation of Ministry of Education of China (教育部骨干教师基金)

HUANG Bing was born in 1972. He is a Ph.D. candidate at the Department of Automation, Nanjing University of Science & Technology. His current research interest includes rough set theory and its applications. **HE Xin** was born in 1979. He is a Ph.D. candidate at the Department of Automation, Nanjing University of Science & Technology. His current research interest is speech signal processing. **ZHOU Xian-Zhong** was born in 1962. He is a professor and doctoral supervisor at the Department of Automation, Nanjing University of Science & Technology. His research areas are intelligent information processing, information system theory and technology.

1 Introduction

Rough set theory, developed in recent years, has made a great progress in knowledge acquisition. The theory is used in various fields and arouses the attention of scholars all over the world. The classical rough set theory, however, has its own limitation: a) sensitivity to noises; b) uneasy to be understood in its algebra views; c) lack of the disposal methods for an incomplete information system.

The variable precision rough set model is applied[1] to handle the first problem. To the second defect, the relation between knowledge and information is founded and the information representation of concepts and operations is presented in rough set theory[2]. A heuristic reduction algorithm based on mutual information is presented[3]. The algebra and information views in rough set theory are discussed systemically[4], and the conclusion is obtained that the reduction from the algebra and information views are equivalent in consistent decision tables, while the reduction based on the algebra is included in the information view in inconsistent decision tables. The knowledge reduction algorithm based on rough set and conditional information entropy is presented[5]. As far as the third defect, the classical rough sets based on the hypothesis of equivalence relation are not applicable and then the equivalence relation is broadened to a toleration relation[6] or a similarity relation[7]. On this condition, the toleration classes determined by the toleration relation or the similarity classes by similarity relation do not form one partition of the universe but one covering of the universe. Consequently, the classical rough set theory is extended to the generalized rough set theory. On this basis, the covering theory of the generalized rough set is studied deeply[8]. A sufficient and necessary condition of the same generalized rough set covering generated by two coverings in one universe is obtained[9]. Information entropy is introduced into incomplete information systems[10], and a kind of new rough entropy is defined to describe the incomplete information systems and the roughness of rough set.

On these bases, a kind of information entropy is defined in this paper, which is based on the generalized rough set covering reduction and is used to represent the roughness of knowledge and that of rough set for the incomplete information systems. The conclusion is obtained that rough entropy decreases monotonously as the covering reduction becomes finer. Thereby a new kind of measure is presented to the roughness of rough set and knowledge for the incomplete information system. A rough set method for incomplete information systems is developed. The bases are also established for knowledge acquisition in an incomplete information system.

The rest of this paper is arranged as follows. In Section 2, the covering theory of the generalized rough sets to be used in this paper is introduced. The information measure for the roughness of rough set and knowledge based on the toleration relation is introduced in Section 3, and an example is given to illuminate the defects of this method for the generalized covering reduction. The main results of this paper are given in Section 4. Section 5 summarizes this paper and discusses the applications of rough entropy presented.

2 Basic Concepts and Properties of Generalized Rough Set Covering and Its Reduction[8,9]

Definition 2.1. Let U be a universe, and C be its subsets family. If all subsets are not empty and $\bigcup C = U$, then C is called a covering of U.

Definition 2.2. Let U be a non-empty set, C be a covering of U, then the ordered pair (U,C) is called a covering approximation space.

Definition 2.3. Let (U,C) be a covering approximation space, $x \in U$, then $Md(x)=\{K \in C | x \in K \wedge (\forall S \in C \wedge x \in S \wedge S \subseteq K \Rightarrow K=S)\}$ is called the minimal description of x.

Definition 2.4. If (U,C) is a covering approximation space, $X \subseteq U$, then the set family $C_*(X)=\{K \in C | K \subseteq X\}$ is called the family of sets bottom approximation sets of the set X as to C. The set $X_* = \bigcup C_*(X)$ is called the covering lower approximation sets of the set X. The set $X_*^* = X - X_*$ is called the boundary sets of the set X. The

family of sets $Bn(X)=\{Md(x)|x\in X_*^*\}$ is called the family of set approximation boundary of the set X, and the family of sets $C^*(X)=C_*(X)\bigcup Bn(X)$ is called the family of set top approximation of the set X. The set $X^*=\bigcup C^*(X)$ is called the covering upper approximation sets of the set X. If $C^*(X)=C_*(X)$, then X is called relatively exact to C. Otherwise X is called relatively inexact to C.

Definition 2.5. Let $\langle U,C\rangle$ be an approximation space, and $K\in C$. If K is denoted by conjunction of some sets in $C-\{K\}$, then K is called a redundant set of C, otherwise is called a non-redundant set.

Obviously, a covering is still a covering after the redundant sets are taken out.

Definition 2.6. If $\langle U,C\rangle$ is an approximation space, then the covering, in which each set is non-redundant, is called a reduction of C and denoted by $red(C)$.

Theorem 2.1. Let $\langle U,C\rangle$ be an approximation space, the same covering upper and lower approximations are generated respectively by C and $red(C)$.

Theorem 2.2. Let C_1 and C_2 be two coverings of U, then the covering lower and upper approximations are generated respectively by C_1 and C_2, if and only if $red(C_1)=red(C_2)$.

3 Knowledge Roughness Measure in Incomplete Information Systems

Definition 3.1[10]. $S=(U,A)$ is an incomplete information system, $U=\{x_1,x_2,\ldots,x_{|U|}\}$, where $|U|$ denotes the cardinality of the set U. $P\subseteq A$, $S_p(x_i)$ denotes the toleration class or similarity class of x_i. The rough entropy of the knowledge P is defined as follows:

$$E_1(P)=\sum_{i=1}^{|U|}\frac{|S_P(x_i)|}{|U|}\log|S_P(x_i)|.$$

Theorem 3.1[10]. Let $S_1=(U,P)$ and $S_2=(U,Q)$ be two incomplete information systems. If $U/SIM(Q)\subset U/SIM(P)$, namely the classes determined by the knowledge Q are included in the classes determined by the knowledge P, then $E_1(Q)<E_1(P)$.

We can conclude from Theorem 3.1 that the rough entropy of knowledge monotonously decreases when the information granularities become smaller.

Example 1. $S=(U,A)$ is an incomplete information system, $U=\{x_1,x_2,x_3,x_4\}$, $P,Q\subseteq A$, and $C_P=\{\{x_1\},\{x_2,x_3\},\{x_4\},\{x_1,x_2,x_3,x_4\}\}$, $C_Q=\{\{x_1\},\{x_2,x_3\},\{x_2,x_3\},\{x_4\}\}$, then $E_1(P)=5/2$, $E_1(Q)=1/2$, $\text{red}(C_P)=\text{red}(C_Q)=\{\{x_1\},\{x_2,x_3\},\{x_4\}\}$, where C_P and C_Q denote respectively the coverings of the universe determined by the toleration or similarity classes of P and Q.

It can be known from Theorem 2.1 and Theorem 2.2 that the generalized covering upper and lower approximations are determined by the covering reduction, but not their covering itself. In Example 1, the covering reductions generated by knowledge P and Q are the same ones, but their rough entropies are not equal to each other. Therefore it is necessary to seek for a new kind of knowledge roughness measure in the generalized rough set covering theory.

4 Rough Entropy of Knowledge and Rough Set Based on Generalized Rough Sets Covering Reduction

4.1 Rough entropy of knowledge

Definition 4.1.1. $S=(U,A)$ is an incomplete information system, $U=\{x_1,x_2,\ldots,x_{|U|}\}$, $P\subseteq A$. If the covering of the universe U determined by the knowledge P is denoted by $C_P=\{C_1,C_2,\ldots,C_n\}$ and $red(C_P)=\{C_{01},C_{02},\ldots,C_{0m}\}$, then the rough entropy of knowledge P in the generalized rough set covering reduction is defined as follows:

$$E(P)=\sum_{k=1}^{m}\frac{|C_{0k}|}{m}\log_2|C_{0k}|.$$

Example 2. In Example 1, it can be calculated by the new definition. The result is

$$E(P)=E(Q)=1/2.$$

Property 4.1.1. $S=(U,A)$ is an incomplete information system, $P\subseteq A$, then $E(P)$ can obtain its maximum $|U|\log_2|U|$ if and only if $red(C_P)=U$, and it can also obtain its minimum 0, if and only if $red(C_P)=\{\{x_1\},\{x_2\},\ldots,\{x_{|U|}\}\}$.

From Property 4.1.1, it can be concluded that information quantity provided by knowledge *P* is zero when its rough entropy reaches maximum, and it cannot distinguish any two objects in *U*, when the covering of the universe is no meaning. When the rough entropy of knowledge *P* obtains its minimum, the information quantity is the most and every objects can be discriminated by *P* in the universe.

Definition 4.1.2. Let C_1 and C_2 be two coverings of the universe $U=\{x_1,x_2,...,x_n\}$ and $red(C_1)=\{C_{11},C_{12},...,C_{1p}\}$, $red(C_2)=\{C_{21},C_{22},...,C_{2q}\}$ are respectively the reductions of the covering C_1 and C_2. If for every $x_i\in U, x_i\in C_{1l}(1\le l\le p)$, and $x_i\in C_{2l'}(1\le l'\le q)$, $C_{1l}\subseteq C_{2l'}$ holds, then $red(C_1)$ is finer than $red(C_2)$, denoted by $red(C_1)\subseteq red(C_2)$. If $red(C_1)\subseteq red(C_2)$, and there exists $x_{i_0}\in U$, $C_{1l_0}\in red(C_1)$, and $C_{2l_0'}\in red(C_2)$, which makes $x_{i_0}\in C_{1l_0}\subset C_{2l_0'}$, then $red(C_1)$ is called strictly finer than $red(C_2)$, denoted by $red(C_1)\subset red(C_2)$.

Property 4.1.2. Let $S=(U,A)$ be an incomplete information system, and $P,Q\subseteq A$. If $red(C_P)\subseteq red(C_Q)$, then $E(P)\le E(Q)$. Especially, if $red(C_P)\subset red(C_Q)$, then $E(P)<E(Q)$.

Proof. Let $U=\{x_1,x_2,...,x_n\}$, $red(C_P)=\{C_{11}, C_{12}, C_{12},...,C_{1p}\}$ and $red(C_Q)=\{C_{21},C_{22},...,C_{2q}\}$. Since $red(C_P)\subseteq red(C_Q)$, and then suppose $C_{1i_1},C_{1i_2},...,C_{1i_{p'}}\in red(C_P), C_{2j_1},C_{2j_2},...,C_{2j_{q'}}\in red(C_Q)$, such that $x_i\in U$ and $x_i\in C_{1i_{i'}}\subseteq C_{2j_{j'}}(1\le i'\le p',\ 1\le j'\le q')$. Obviously:

$$|C_{1i_{i'}}|\log_2|C_{1i_{i'}}|\le|C_{2j_{j'}}|\log_2|C_{2j_{j'}}|.$$

When $1\le i\le n$, objects of $red(C_P)$ and $red(C_Q)$ are respectively chosen by $C_{1i_{i'}}$ and $C_{2j_{j'}}(1\le i'\le p',\ 1\le j'\le q')$. Sets are marked once if they appear repeatedly in $red(C_P)$ and $red(C_Q)$, then

$$\frac{1}{p}\sum_{k=1}^{p}|C_{1k}|\log_2|C_{1k}|\le\frac{1}{q}\sum_{k=1}^{q}|C_{2k}|\log_2|C_{2k}|.$$

So $E(P)\le E(Q)$.

Similarly, it is easy to prove that $E(P)<E(Q)$ when $red(C_P)\subset red(C_Q)$.

It can be concluded from Property 4.1.2 that the rough entropy of knowledge monotonously decreases as the covering reduction becomes finer.

Example 3. Let $red(C_P)=\{\{x_1\},\{x_2,x_3\},\{x_1,x_2,x_3,x_4\}\}$ and $red(C_Q)=\{\{x_1,x_2\},\{x_1,x_2,x_3,x_4\}\}$, then $E(P)=10/3<E(Q)=5$, while $red(C_P)\subset red(C_Q)$ does not hold.

From Example 3, it can be concluded that the converse proposition of Property 4.1.2 does not hold.

4.2 Rough entropy of rough set

The roughness of a rough set can be measured by its rough degree.

Definition 4.2.1. Let $S=(U,A)$ be an incomplete information system, $P\subseteq A$, the rough degree of $X\subseteq U$ about knowledge *P* is defined as follows:

$$\rho_P(X)=1-\frac{|P(X_*)|}{|P(X^*)|},$$

where $P(X_*)$ and $P(X^*)$ denote respectively the covering lower and upper approximation sets of X about knowledge P.

Example 4. Let $S=(U,A)$ be an incomplete information system, $P,Q\subseteq A$, if $red(C_P)=\{\{x_1,x_2\},\{x_2,x_3,x_4\}\}$, $red(C_Q)=\{\{x_1\},\{x_2\},\{x_3,x_4\}\}$, $X=\{x_1,x_2,x_3\}$, then $P(X_*)=Q(X_*)=\{x_1,x_2\}$, $P(X_*^*)=Q(X_*^*)=\{X_3\}$, $P(Bn(X))=\{x_2,x_3,x_4\}$, $Q(Bn(X))=\{x_3,x_4\}$, $P(X^*)=Q(X^*)=\{x_1,x_2,x_3,x_4\}$, $\rho_P(X)=\rho_Q(X)=1/2$.

The uncertainty of knowledge P is larger than that of Q in Example 4, but X has the same rough degree. Therefore, it is necessary to find a new and more accurate uncertainty measure.

Definition 4.2.2. Let $S=(U,A)$ be an incomplete information system, $P\subseteq A$, the rough entropy of $X\subseteq U$ about knowledge P is defined as follows:

$$E_P(X)=\rho_P(X)E(P).$$

From Definition 4.2.2, the rough entropy of rough set is related not only to its own rough degree, but also to the uncertainty of knowledge covering reduction in the universe (rough entropy of knowledge).

Example 5. The rough entropy of X in Example 4 is calculated under knowledge P and Q.

$$E_P(X)=\rho_P(X)E(P)=(2+3\log_2 3)/4,\ E_Q(X)=\rho_Q(X)E(Q)=1/3,\ E_P(X)>E_Q(X).$$

Obviously, the rough entropy of rough set is more accurate than the rough degree to measure the roughness of rough set.

Property 4.2.1. Let $S=(U,A)$ be an incomplete information system, $P\subseteq A$, $X\subseteq U$ is a rough set under Q, $red(C_P)\subset red(C_Q)$, then

$$E_P(X)<E_Q(X).$$

Proof. $\forall x\in Q(X_*),\exists C_{2j_0}\in red(C_Q)$, such that $x\in C_{2j_0}\subseteq X$, for $red(C_P)\subset red(C_Q)$, $\exists C_{1i_0}\in red(C_P)$, $x\in C_{1i_0}\subseteq C_{2j_0}\subseteq X$, so $x\in P(X_*)$, thus $P(X_*)\supseteq Q(X_*)$. $\forall x\in P(X^*)=P(X_*)\bigcup P(Bn(X_*^*))$, for $x\in P(X_*)$, $\exists C_{1i_0}\in red(C_P)$, such that $x\in C_{1i_0}\subseteq X$, as $red(C_P)\subset red(C_Q)$, $\exists C_{2j_0}\in red(C_Q)$, then $x\in C_{1i_0}\subseteq C_{2j_0}\bigcap X\neq\phi$, thus $x\in Q(X^*)$.When $x\in P(Bn(X))$, $\exists C_{1i_0}\in red(C_P)$, $x'\in X-P(X_*),x,x'\in C_{1i_0}$, from $P(X_*)\supseteq Q(X_*)$, it can be concluded $x'\in X-Q(X_*)=Q(Bn(X))$. By $red(C_P)\subset red(C_Q)$, $\exists C_{2j_0}$ (including the minimal description of x,x'), leading to $x,x'\in C_{1i_0}\subseteq C_{2j_0}$, then $C_{2j_0}\in Q(Bn(X))$, hence $x\in Q(X^*)$. Accordingly, it is concluded $P(X^*)\subseteq Q(X^*)$.

Then the following inequalities hold:

$$\frac{|P(X_*)|}{|P(X^*)|}\geq\frac{|Q(X_*)|}{|Q(X^*)|},\quad 1-\frac{|P(X_*)|}{|P(X^*)|}\leq 1-\frac{|Q(X_*)|}{|Q(X^*)|}.$$

Namely, $\rho_P(X)\leq\rho_Q(X)$ and $X\subseteq U$ is rough sets about Q, so $\rho_Q(X)\neq 0$. It can be known from Property 4.2 that $E(P)<E(Q)$, hence $\rho_P(X)E(P)<\rho_Q(X)E(Q)$. That is to say, $E_P(X)<E_Q(X)$.

It can be deduced from Property 4.2.1 that the rough entropy of a rough set monotonously decreases as the covering reduction becomes finer.

5 Conclusions and Discussions

Rough set theory is a new mathematical tool to deal with vagueness and uncertainty. Development of a rough computational method is one of the most important research tasks. While in reality, incomplete information confines the applications of classical rough set theory. In this paper, a measure to knowledge and its important properties in incomplete information systems are established by introducing information entropy to the covering reduction theory of a generalized rough set. The conclusion that the rough entropy of knowledge and rough set monotonously decrease as the covering reduction becomes finer is obtained. It is also clarified that the rough entropy of a rough set is more accurate than its rough degree.

Rough entropy based on the generalized rough set covering reduction more accurately represents the roughness of knowledge and rough set than Ref.[10]. Algorithm based rough entropy presented in this paper can be designed in knowledge acquisition under incomplete information systems. This is our future work.

At the same time, retrieval for video information is a new research field. In general, video information systems are incomplete. How to utilize rough set theory based generalized rough sets covering reduction in video information retrieval is also our important task.

References:

[1] Chen XH, Zhu SJ, Ji YD. Entropy based uncertainty measures for classification rules with inconsistency tolerance. In: Proc. Of the 2000 IEEE Int'l Conf. on Systems, Man, and Cybernetics, Vol.4. 2000. 2816 ~2821.

[2] Miao DQ, Wand J. An information representation of the concepts and operations in rough set theory. Journal of Software, 1999, 10(2):113~116 (in Chinese with English abstract).

[3] Miao DQ, Hu GR. A heuristic algorithm for reduction of knowledge. Computer Research & Development, 1999,36(6):681~684 (in Chinese with English abstract).

[4] Wang GY. Algebra view and information view of rough sets theory. In: Data Mining and Knowledge Discovery: Theory, Tools, and Technology Ⅲ, Proc. of the SPIE. 2001. 200~207.

[5] Yu H, Wang GY, Yang DC ,Wu ZF. Knowledge reduction algorithms based on rough set and conditional information entropy. In: Data Mining and knowledge Discovery: Theory, Tools, and TechnologyⅣ, Proc. of the SPIE. 2002. 422~431.

[6] Kryszkiewicz M, Rough set approach to incomplete information systems. Information Sciences, 1998,112:39~49.

[7] Slowinski R, Vsnderpooten D. A generalized definition of rough approximations based on similarity. IEEE Trans. on Data and Knowledge Engineering, 2000,12(2):331~336.

[8] Bonikowski Z, Bryniarski E, Wybranied-Skardowska U. Extensions and intentions in the rough set theory. Information Sciences, 1998,107:149~167.

[9] Zhu F, Wang FY. Some results on covering generalized rough sets. Pattern Recognition and Artificial Intelligence, 2002,15:6~13.

[10] Liang JY, Xu ZB. Uncertainty measures of roughness of knowledge and rough sets in incomplete information systems. In: Proc. of the 3rd World Congress on Intelligent Control and Automation, Vol.4. 2000. 2526~2529

附中文参考文献：

[2] 苗夺谦,王珏.粗集理论中概念与运算的信息表示.软件学报,1999,10(2):113~116.

[3] 苗夺谦,胡桂荣.知识约简的一种启发式算法.计算机研究与发展,1999,36(6):681~684.

[9] 祝峰,王飞跃.关于覆盖广义粗集的一些基本结果.模式识别与人工智能,2002,15(1):6~13.