Combination Entropy and Combination Granulation in Incomplete Information System

Yuhua Qian[1] and and Jiye Liang[2]

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
School of Computer and Information Technology, Shanxi University
Taiyuan, 030006, People's Republic of China
[1] jinchengqyh@126.com, [2] ljy@sxu.edu.cn

Abstract. Based on the intuitionistic knowledge content characteristic of information gain, the concepts of combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system are introduced, their some properties are given. Furthermore, the relationship between combination entropy and combination granulation is established. These concepts and properties are all special instances of those in in complete information system. These results will be very helpful for understanding the essence of knowledge content and uncertainty measurement in incomplete information system.

Keywords: Incomplete information system, combination entropy, combination granulation.

1 Introduction

Rough set theory, introduced by Pawlak [1, 2], is a relatively new soft computing tool for the analysis of a vague description of an object. The indiscernibility relation generated constitutes a mathematical basis of the rough set theory; it induces a partition of the universe into blocks of indiscernible objects, called elementary sets, that can be used to build knowledge about a real or abstract world [1-4]. The use of the indiscernibility relation results in information granulation.

The entropy of a system as defined by Shannon gives a measure of uncertainty about its actual structure [5]. It has been a useful mechanism for characterizing the information content in various modes and applications in many diverse fields. Several authors (Düntsch and Gediga, [6]; Beaubouef et al., [7]; Klir and Wierman, [8]; Liang and Xu, [9]; Liang et al. [10]) have used Shannon's concept and its variants to measure uncertainty in rough set theory. But Shannon's entropy is not fuzzy entropy, and cannot measure the fuzziness in rough set theory. A new information entropy is proposed by Liang in [11-13], some important properties of this entropy are also derived. In [14], Combination entropy and combination granulation in complete information system are proposed, their gain function possesses intuitionistic knowledge content characteristic. Combination entropy can be used to measure the uncertainty of knowledge and knowledge content.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 184–190, 2006.

This paper introduces combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system. The gain function considered here possesses intuitionistic knowledge content characteristic, i.e., the whole number of pairs of elements which can be distinguished each other on the universe. Furthermore, the relationship between combination entropy and combination granulation is established. These results will be very helpful for understanding the essence of knowledge content, uncertainty measurement and the significance of an attribute in incomplete information system.

2 Incomplete Information System

An information system is a pair $S = (U, A)$, where,

(1) U is a non-empty finite set of objects;
(2) A is a non-empty finite set of attributes;
(3) for every $a \in A$, there is a mapping a, $a : U \to V_a$, where V_a is called the value set of a.

If V_a contains a null value for at least one attribute $a \in A$, then S is called an incomplete information system, otherwise it is complete. Further on, we will denote the null value by $*$.

Let $S = (U, A)$ be an information system, $P \subseteq A$ an attribute set. We define a binary relation on U as follows

$$SIM(P) = \{(u, v) \in U \times U |\ \forall a \in P, a(u) = a(v) \ \text{ or } \ a(u) = * \ \text{ or } \ a(v) = *\}.$$

In fact, $SIM(P)$ is a tolerance relation on U, the concept of a tolerance relation has a wide variety of applications in classification [15]. It can be easily shown that $SIM(P) = \bigcap_{a \in P} SIM(\{a\})$. Let $S_P(u)$ denote the set $\{v \in U | (u, v) \in SIM(P)\}$. They constitute a covering of U, i.e., $S_P(u) \neq \emptyset$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u) = U$.

Let $S = (U, A)$ be an incomplete information system, $P, Q \subseteq A$. We say that Q is coarser than P (or P is finer than Q), denoted by $P \preceq Q$, if and only if $S_P(u_i) \subseteq S_Q(u_i)$ for $\forall i \in \{1, 2, \cdots, |U|\}$. If $P \preceq Q$ and $P \neq Q$, we say that Q is strictly coarser than P (or P is strictly finer than Q) and denoted by $P \prec Q$. In fact, $P \prec Q \Leftrightarrow$ for $\forall i \in \{1, 2, \cdots, |U|\}$, we have that $S_P(u_i) \subseteq S_Q(u_i)$, and $\exists j \in \{1, 2, \cdots, |U|\}$, such that $S_P(u_j) \subset S_Q(u_j)$.

3 Combination Entropy

In this section, combination entropy in an incomplete information system is introduced. Its some properties are discussed.

Definition 1. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$. The combination entropy of knowledge A is defined by*

$$CE(A) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{C_{|U|}^2 - C_{|S_A(u_i)|}^2}{C_{|U|}^2} = \frac{1}{|U|} \sum_{i=1}^{|U|} (1 - \frac{C_{|S_A(u_i)|}^2}{C_{|U|}^2}), i \leq |U|, \qquad (1)$$

where $\frac{C^2_{|U|}-C^2_{|S_A(u_i)|}}{C^2_{|U|}}$ *denotes the probability of pairs of elements which are probably distinguishable each other within the whole number of pairs of elements on the universe* U.

Obviously, we have that $0 \leq CE(A) \leq 1$.

Proposition 1. *Let* $S = (U, A)$ *be an incomplete information system,* $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$. *Then the combination entropy of knowledge* A *degenerate into*

$$CE(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}(1 - \frac{C^2_{|X_i|}}{C^2_{|U|}}). \tag{2}$$

Proof. Let $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$, $X_i = \{u_{i1}, u_{i2}, \cdots, u_{is_i}\}$ $(i \leq m)$, where $|X_i| = s_i$, and $\sum_{i=1}^{m} |s_i| = |U|$, then the relationships among elements in $U/SIM(A)$ and elements in $U/IND(A)$ are as follows

$$\begin{gathered} X_i = S_A(u_{i1}) = S_A(u_{i2}) = \cdots = S_A(u_{is_i}), \\ |X_i| = |S_A(u_{i1})| = |S_A(u_{i2})| = \cdots = |S_A(u_{is_i})|. \end{gathered}$$

Hence, one have that

$$\begin{aligned} CE(A) &= \sum_{i=1}^{m} \tfrac{|X_i|}{|U|}(1 - \tfrac{C^2_{|X_i|}}{C^2_{|U|}}) \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} |X_i| \times \tfrac{C^2_{|X_i|}}{C^2_{|U|}} \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} \tfrac{|S_A(u_{i1})|+|S_A(u_{i1})|+\cdots+|S_A(u_{is_i})|}{|X_i|} \times \tfrac{C^2_{|X_i|}}{C^2_{|U|}} \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} \tfrac{C^2_{|S_A(u_i)|}}{C^2_{|U|}} \\ &= \tfrac{1}{|U|} \sum_{i=1}^{|U|} (1 - \tfrac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}). \end{aligned}$$

This completes the proof.

Remark. In [14], the combination entropy of a complete information system $S = (U, A)$ with $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$ is defined as $CE(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}(1 - \frac{C^2_{|X_i|}}{C^2_{|U|}})$. Proposition 1 states that the combination entropy in complete information system is a special instance of the combination entropy in incomplete information system.

Proposition 2. *Let* $S = (U, A)$ *be an incomplete information system,* $P, Q \subseteq A$ *two subsets on* A. *If* $P \prec Q$, *then* $CE(P) > CE(Q)$.

Proof. Let $U/SIM(P) = \{S_P(u_1), S_P(u_2), \cdots, S_P(u_{|U|})\}$, $U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), \cdots, S_Q(u_{|U|})\}$. If $P \prec Q$, then for $\forall i \in \{1, 2, \cdots, |U|\}$, one have that $S_P(u_i) \subseteq S_Q(u_i)$ and there exists $j \in \{1, 2, \cdots, |U|\}$ such that $S_P(u_i) \subset S_Q(u_i)$, i.e., $|S_P(u_j)| < |S_Q(u_j)|$.

Hence, one have that

$$\begin{aligned}
&|S_P(u_j)| < |S_Q(u_j)| \\
\Longrightarrow\ & C^2_{|S_P(u_j)|} < C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & \sum_{i=1}^{|U|} C^2_{|S_P(u_j)|} < \sum_{i=1}^{|U|} C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & 1 - \tfrac{1}{|U|} \sum_{i=1}^{|U|} \tfrac{C^2_{|S_Q(u_j)|}}{C^2_{|U|^2}} < 1 - \tfrac{1}{|U|} \sum_{i=1}^{|U|} \tfrac{C^2_{|S_P(u_j)|}}{C^2_{|U|^2}} \\
\Longrightarrow\ & CE(Q) < CE(P).
\end{aligned}$$

This completes the proof.

Proposition 2 states that combination entropy of knowledge increases as tolerance classes become smaller through finer classification.

4 Combination Granulation

In this section, combination granulation in an incomplete information system is introduced. It has some very useful properties. The relationship between combination entropy and combination granulation in incomplete information system is established.

Definition 2. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$. Then combination granulation of A is defined by*

$$CG(A) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}, \tag{3}$$

where $\frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}$ denotes the probability of pairs of elements on tolerance class $S_A(u_i)$ within the whole number of pairs of elements on the universe U.

Clearly, one have that $0 \leq CE(G) \leq 1$.

Proposition 3. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, and $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$. Then knowledge granulation of knowledge A degenerates into*

$$CG(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|} \frac{C^2_{|X_i|}}{C^2_{|U|}}. \tag{4}$$

Proof. Similar to proposition 1, we have that

$$CG(A) = \sum_{i=1}^{m} \tfrac{|X_i|}{|U|} \tfrac{C^2_{|X_i|}}{C^2_{|U|}}$$

$$= \frac{1}{|U|}\sum_{i=1}^{m}\frac{|S_A(u_{i1})|+|S_A(u_{i2})|+\cdots+|S_A(u_{is_i})|}{|X_i|}\frac{C^2_{|X_i|}}{C^2_{|U|}}$$
$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}.$$

This completes the proof.

Remark. In [14], the combination granulation of a complete information system $S = (U, A)$ with $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$ is defined as $CG(A) = \sum_{i=1}^{m}\frac{|X_i|}{|U|}\frac{C^2_{|X_i|}}{C^2_{|U|}}$. Proposition 3 states that the combination granulation in complete information system is a special instance of the combination granulation in incomplete information system.

Proposition 4. *Let $S = (U, A)$ be an incomplete information system, $P, Q \subseteq A$ two subsets on A. If $P \prec Q$, then $CG(P) < CG(Q)$.*

Proof. Let $U/SIM(P) = \{S_P(u_1), S_P(u_2), \cdots, S_P(u_{|U|})\}$ and $U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), \cdots, S_Q(u_{|U|})\}$. If $P \prec Q$, then $S_P(u_i) \subseteq S_Q(u_i)$ ($i \in \{1, 2, \cdots, |U|\}$), and $\exists j \in \{1, 2, \cdots, |U|\}$ such that $S_P(u_i) \subset S_Q(u_j)$, i.e., $|S_P(u_j)| < |S_Q(u_j)|$.

Hence, it follows that

$$\begin{aligned}
&|S_P(u_j)| < |S_Q(u_j)| \\
\Longrightarrow\ & C^2_{|S_P(u_j)|} < C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & \sum_{i=1}^{|U|} C^2_{|S_P(u_i)|} < \sum_{i=1}^{|U|} C^2_{|S_Q(u_i)|} \\
\Longrightarrow\ & CG(P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_P(u_i)|}}{C^2_{|U|}} < \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_Q(u_i)|}}{C^2_{|U|}} = CG(Q).
\end{aligned}$$

This completes the proof.

Proposition 4 states that combination granulation of knowledge decreases as tolerance classes become smaller through finer classification.

Here, we will establish the relationship between combination entropy and combination granulation in incomplete information system as follows.

Proposition 5. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, then the relationship between the combination entropy $CE(R)$ and combination granulation $CG(R)$ is as follows*

$$CE(A) + CG(A) = 1. \tag{5}$$

Proof. It is straightforward.

Remark. Proposition 5 shows the relationship between combination entropy and combination granulation is strict complement relationship, i.e., they possess the same capability on depicting the uncertainty of an incomplete information system.

5 Conclusions

In the present research, the concepts of combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system are introduced, their important properties are obtained, the relationship between them is established. The relationship can be expressed as $CE(A)+CG(A) = 1$. These concepts and properties in complete information system are all special instances of those in in complete information system. These conclusions have a wide variety of applications, such as measuring knowledge content, measuring the significance of an attribute, constructing decision trees and building the heuristic function in a heuristic reduct algorithm in incomplete information system. They will paly a significant role in further researches in incomplete information system.

Acknowledgements. This work was supported by the national natural science foundation of China (No. 70471003, No. 60573074, No. 60275019), the foundation of doctoral program research of the ministry of education of China (No. 20050108004), the natural science foundation of Shanxi, China (No. 20031036) and the top scholar foundation of Shanxi, China, key project of science and technology research of the ministry of education of China.

References

1. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht (1991).
2. Pawlak, Z., Grzymala-Busse, J.W., Slowinski, R. and Ziarko, W.: Rough sets. *Comm. ACM*. 38 (11) (1995) 89-95.
3. Mi, J.S., Wu, W.Z. and Zhang, W.X.: Approaches to knowledge reduction based on variable precision rough set model. *Information Sciences*. 159 (2004) 255-272.
4. Zhang, W.X., Wu, W.Z., Liang, J.Y., Li, D.Y.: *Theory and Method of Rough Sets*. Science Press, Beijing, China (2001).
5. Shannon, C.E.: The mathematical theory of communication. *The Bell System Technical Journal*. 27 (3, 4) (1948) 373-423, 623-656.
6. Düntsch, I. and Gediga, G.: Uncertainty measures of rough set prediction. *Artificial Intelligence*. 106 (1998) 109-137.
7. Beaubouef, T., Perty, F.E. and Arora, G.: Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Information Sciences*. 109 (1998) 185-195.
8. Klir, G.J. and Wierman, M.J.: *Uncertainty Based Information*. Physica-Verlag, New York (1998).
9. Liang, J.Y. and Qu, K.S.: Information mesaures of roughness of knowledge and rough sets in incomplete information systems. *Journal of System Science and System Engineering*. 24 (5) (2001) 544-547.
10. Liang, J.Y., Xu, Z.B.: The algorithm on knowledge reduction in incomplete information systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 24 (1) (2002) 95-103.
11. Liang, J.Y., Chin, K.S., Dang, C.Y. and Richard C.M.Yam.: A new method for measuring uncertainty and fuzziness in rough set theory. *International Journal of General Systems*. 31 (4) (2002) 331-342.

12. Liang, J.Y., Shi, Z.Z., Li, D.Y. and Wierman, M.J.: The information entropy, rough entropy and knowledge granulation in incomplete information system. *International Journal of General Systems*. (to appear)
13. Liang, J.Y., Li, D.Y.: *Uncertainty and Knowledge Acquisition in Information Systems*. Science Press, Beijing, China (2005).
14. Liang, J.Y., Qian, Y.H.: Combination entropy and combination granulation in rough set theory. *Foundamenta Informaticae*. (to appear)