Uncertainty and Feature Selection in Rough Set Theory

Jiye Liang

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China
ljy@sxu.edu.cn

Abstract. In rough set theory, the uncertainty of granulation and efficient feature selection algorithms have attracted much attention in recent years. We focus on the review of several common uncertainty measures and the relationships among them. An efficient accelerator is developed to accelerate a heuristic process of feature selection.

Keywords: Rough set, information entropy, information granulation, granular space distance, feature selection.

1 Introduction

Rough set theory, as one of important models of granular computing, has become a popular mathematical framework for pattern recognition, image processing, data mining and knowledge discovery from various types of data[1, 4, 11–13, 22, 24]. In rough set theory, the uncertainty of granulation and efficient feature selection algorithms have aroused wide concern and research in recent years. Therefore, we focus on the review of several common measures of uncertainty such as information entropy, information granulation, axiom approach of information granulation and granular space distance [6–9, 15–17]. Furthermore, based on the existing heuristic feature selection algorithms[2, 3, 20], an efficient accelerated feature selection algorithm is developed to save computational cost[14]. Finally, some further research topics in rough set theory are presented. The main works are as follows.

2 Uncertainty in Rough Set Theory

In practices, the expression of information is usually uncertain, which comes from the disorder, incidental, vague and approximation of information. In this section, we focus on the review of several common measures of uncertainty in rough set theory, such as information entropy and information granulation. And an axiom approach of information granulation was developed, and it has been proved each of the existing definitions of information granulation is a special instance of this axiom definition. In addition, a notion of granular space distance, which can be also conceived as a new uncertainty measure, is introduced to measure the divergence between granular spaces.

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 8–15, 2011.

2.1 Information Entropy

Many scholars have used Shannon's entropy and its variants to measure uncertainty in rough set theory[19, 20].

Let $S = (U, A)$ be an information system and $U/A = \{R_1, R_2, \cdots, R_m\}$. Shannon's entropy is defined as

$$H(A) = -\sum_{i=1}^{m} \frac{|R_i|}{|U|} log_2 \frac{|R_i|}{|U|}. \tag{1}$$

Let $P, Q \subseteq A$, $U/P = \{P_1, P_2, \cdots, P_m\}$ and $U/Q = \{Q_1, Q_2, \cdots, Q_n\}$. Shannon's conditional entropy is defined as

$$H(Q|P) = -\sum_{i=1}^{m} \frac{|P_i|}{|U|} \sum_{j=1}^{n} \frac{|P_i \cap Q_j|}{|P_i|} log_2 \frac{|P_i \cap Q_j|}{|P_i|}. \tag{2}$$

However, Shannon's entropy is not a fuzzy entropy, which cannot measure the fuzziness in rough set theory. To address this issue, complementary entropy is introduced, which can measure both randomness and fuzziness of an information system [6].

For the information system $S = (U, A)$, complementary entropy is defined as

$$E(A) = \sum_{i=1}^{m} \frac{|R_i|}{|U|} \left(1 - \frac{|R_i|}{|U|}\right). \tag{3}$$

Complementary conditional entropy is defined as

$$E(Q|P) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|P_i \cap Q_j|}{|U|} \frac{|P_i^c \cap Q_j^c|}{|U|}. \tag{4}$$

From the view of knowledge-content nature, we introduced a new information entropy (combination entropy) to measure uncertainty of an information system [15].

For the information system $S = (U, A)$, combination entropy is defined as

$$CE(A) = \sum_{i=1}^{m} \frac{|R_i|}{|U|} \left(1 - \frac{C_{|R_i|}^2}{C_{|U|}^2}\right). \tag{5}$$

Conditional combination entropy is defined as

$$CE(Q|P) = \sum_{i=1}^{m} \left(\frac{|P_i|}{|U|} \frac{C_{|P_i|}^2}{C_{|U|}^2} - \sum_{j=1}^{n} \frac{|P_i \cap Q_j|}{|U|} \frac{C_{|P_i \cap Q_j|}^2}{C_{|U|}^2} \right). \tag{6}$$

In rough set theory, the uncertainty of a target concept results from its boundary region, which is called Roughness. In [12], Pawlak proposed rough degree to measure the roughness of approximated concepts. However, for the different approximation spaces, the rough degrees of a target concept may be identical. To address this issue, we introduced the notion of rough entropy.

For the information system $S = (U, A)$ and $X \subseteq U$. Rough entropy of X is defined as

$$E_A(X) = -\rho_A(X)\left(\sum_{i=1}^{m} \frac{|R_i|}{|U|} log_2 \frac{|R_i|}{|U|}\right), \tag{7}$$

where, $\rho_A(X)$ is the rough degree of X. Compared with the rough degree, the proposed rough entropy is a more accurate measurement of the roughness.

For the information system $S = (U, A)$, rough entropy of A is defined as

$$E_r(A) = -\sum_{i=1}^{m} \frac{|R_i|}{|U|} log_2 \frac{1}{|R_i|}. \tag{8}$$

Further investigations indicate that the relationship between rough entropy and Shannon's entropy is

$$E_r(A) + H(A) = log_2|U|. \tag{9}$$

2.2 Information Granulation

Information granulation is mainly used to study the uncertainty in rough set theory. To characterize the average measure of information granules, we presented the definition of knowledge granulation as follows.

$$GK(A) = \frac{1}{|U|^2} \sum_{i=1}^{m} |R_i|^2. \tag{10}$$

And the relationships between the complementary entropy and the knowledge granulation is also established in [7], which is expressed as

$$E(A) + GK(A) = 1. \tag{11}$$

Another information granulation called combination granulation is proposed in [15], defined as

$$CG(A) = \sum_{i=1}^{m} \frac{|R_i|}{|U|} \frac{C_{|R_i|}^2}{C_{|U|}^2}. \tag{12}$$

And the relationships between the combination entropy and combination granulation is expressed as

$$CE(A) + CG(A) = 1. \tag{13}$$

Based on the above characterizations and representation of information granules, we presented an axiom approach of information granulation in [9].

Firstly, two new binary relations are introduced, which will be used in the following.

For the information system $S = (U, A)$ and $P, Q \subseteq A$. $S_P(x_i)$ denotes the tolerance class of x with respect to P, $K(P) = \{S_P(x_i) | x_i \in U\}$ and $K(Q) = \{S_Q(x_i) | x_i \in U\}$. f' and f'' are bijective functions.

Then, a binary relation $\approx$ is defined as

$K(P) \approx K(Q) \Leftrightarrow \exists f' : K(P) \to K(Q)$ such that $|S_P(x_i)| = |f'(S_P(x_i))|$, $x_i \in U$, just $P \approx Q$.

A binary relation $\preceq'$ is defined as

$K(P) \preceq' K(Q) \Leftrightarrow \exists f'' : K(P) \to K(Q)$ such that $|S_P(x_i)| \leq |f''(S_P(x_i))|$, $x_i \in U$, just $P \preceq' Q$.

If $K(P) \preceq' K(Q)$ and $K(P) \not\approx K(Q)$, then $K(P) \prec' K(Q)$, just $P \prec' Q$.

Based on above expressions, we introduce the axiom approach as follows.

For the information system $S = (U, A)$, G be a mapping from the power set of A to the set of real numbers. We say that G is an information granulation in an information system if G satisfies the following conditions:

1) $G(P) \geq 0$; (Non-negative)

2) $\forall P, Q \subseteq A$, if $P \approx Q$, then $G(P) = G(Q)$; (Invariability)

3) $\forall P, Q \subseteq A$, if $P \prec' Q$, then $G(P) < G(Q)$. (Monotonicity)

Note that, if $K(P) = \omega$ ($\omega = \{\{x_i\} | x_i \in U\}$), $G(P)$ achieves its minimum value; if $K(P) = \delta$ ($\delta = \{S_P(x_i) | S_P(x_i) = U\}$), $G(P)$ achieves its maximum value.

As mentioned above, we have proved that some of above definitions are various special forms of information granulation, which are as follows:

1) $GK(A)$ is an information granulation, $\frac{1}{|U|} \leq GK(A) \leq 1$.

2) $CG(A)$ is an information granulation, $0 \leq CG(A) \leq 1$.

3) $E_r(A)$ is an information granulation, $0 \leq E_r(A) \leq log_2|U|$.

Through using this axiom definition of information granulation, one can construct new knowledge granulations according to various opinions.

2.3 Granular Space Distance

As mentioned above, in rough set theory, information entropy and information granulation are two main approaches to measuring the uncertainty of a granular space. However, information entropy and information granulation cannot characterize the difference between any two granular space structures in an approximation space. To address this issue, we introduce in this section a notion of granular space distance to differentiate two given granular structures, which aims to reveal the geometrical structure underlying the granular spaces[16].

For the information system $S = (U, A)$ and $P, Q \subseteq A$. Granular space distance between $K(P)$ and $K(Q)$ is defined as

$$D(K(P), K(Q)) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{|S_P(x_i) \cup S_Q(x_i)| - |S_P(x_i) \cap S_Q(x_i)|}{|U|}. \tag{14}$$

Note that the granular space distance satisfies the three properties of a distance space on all granular structures induced by a given universe.

Let $P, Q, R \subseteq A$. $K(P)$, $K(Q)$ and $K(R)$ are granular structures on S. One can get the following properties:

1) $0 \leq D(K(P), K(Q)) \leq 1 - \frac{1}{|U|}$;

2) $D(\omega, \delta) = 1 - \frac{1}{|U|}$, and $D(K(P), \omega) + D(K(P), \delta) = 1 - \frac{1}{|U|}$;

3) $D(K(P), K(Q)) = D(\neg K(P), \neg K(Q))$;

4) If $K(P) \preceq K(Q) \preceq K(R)$, then $D(K(P), K(R)) = D(K(P), K(Q)) + D(K(Q), K(R))$;

5) If $K(P) \preceq K(Q)$, then $D(K(P), \omega) \leq D(K(Q), \omega)$ and $D(K(P), \delta) \geq D(K(Q), \delta)$.

Remark. Based on the analyses in above three subsections, we know that there exists a complement relationship between information entropy and information granulation. In other words, the bigger the entropy is, the smaller the information granulation; the smaller the entropy is, the bigger the information granulation. The granular space distance can be used to distinguish the divergence between two granular structures with the same information granulation (or information entropy), which characterizes the essence of uncertainty of granular structures.

The above measures of uncertainty have a wide variety of applications, such as measuring the significance of attributes, constructing decision trees and evaluating uncertainty of rules, designing heuristic feature selection algorithms, etc. They will play a significant role in further researches in granular computing.

3 Accelerator of Feature Selection

In rough set theory, feature selection (also called attribute reduction) aims to retain the discriminatory power of original features. It plays an important role in many areas including pattern recognition, machine learning and data mining. In the last two decades, many techniques of attribute reduction have been developed. Skowron proposed a discernibility matrix approach to obtain all attribute reducts of an information system [18]. Kryszkiewicz proposed an approach to computing the minimal set of attributes that functionally determine a decision attribute[5]. In addition, based on the generalized rough set models[10, 23, 25], several reduction algorithms were introduced, such as $\beta-$reduct, $\alpha-$reduct, $\alpha-$relative reduct, reduction based on decision-theoretic, etc. To improve the time efficiency, many heuristic attribute reduction algorithms have been developed[2, 3, 6, 7, 20]. However, quite often, the above algorithms are computationally time-consuming for large-scale data sets. To overcome this shortcoming, we designed a accelerated mechanism, which can be used to accelerate a heuristic process of feature selection[14].

Theorem 1. *Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \cdots, R_n\}$ be a family of attribute sets with $R_1 \preceq R_2 \preceq \cdots \preceq R_n$ ($R_i \in 2^C$). Given $P_i = \{R_1, R_2, \cdots, R_i\}$, we have*

$$POS^U_{P_{i+1}}(D) = POS^U_{P_i}(D) \cup POS^{U_{i+1}}_{P_{i+1}}(D), \tag{15}$$

where $U_1 = U$ and $U_{i+1} = U - POS^U_{P_i}(D)$.

According to this theorem, we know that the target decision D can be positively approximated by using granulation orders P on the gradually reduced universe. This mechanism implies the idea of the accelerator for improving the computing performance of a heuristic attribute reduction algorithm.

We focus on the four representative attribute reduction algorithms, which are reduction algorithms based on positive region, Shannon's entropy, complementary entropy and combination entropy, respectively. The significance measures of attributes in those four reduction algorithms are as follows. For convenience, we introduced a uniform notation $EF(D|B)$ to denote those measures. For example, if one adopts Shannon's conditional entropy to define the attribute significance, then $EF(D|B) = H(D|B)$.

For the decision table $S = (U, C \cup D)$ and $B \subseteq C$. $\forall a \in B$, the significance measure of a in B is defined as

$$Sig^{inner}(a, B, D, U) = EF(D|B - \{a\}) - EF(D|B). \tag{16}$$

$\forall a \in C - B$, the significance measure of a in B is defined as

$$Sig^{outer}(a, B, D, U) = EF(D|B) - EF(D|B \cup \{a\}). \tag{17}$$

Based on Theorem 1, we concentrate on the rank preservation of the significance measures of attributes, which can be studied in the following theorem.

Theorem 2. *Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS_B^U(D)$. For $\forall a, b \in C - B$, if $Sig^{outer}(a, B, D, U) \geq Sig^{outer}(b, B, D, U)$, then $Sig^{outer}(a, B, D, U') \geq Sig^{outer}(b, B, D, U')$.*

Based on the rank preservation proved in Theorem 2, a general accelerated attribute reduction algorithm is designed. In this general accelerated algorithm framework, we denote the evaluation function (stop criterion) by $EF^U(B, D) = EF^U(C, D)$.

Algorithm 1. A general accelerated feature selection algorithm (FSPA)

Input: Decision table $S = (U, C \cup D)$;
Output: One reduct *red*.

$Step$ 1: $red \leftarrow \emptyset$;//*red* is the pool to conserve the selected attributes
$Step$ 2: Compute $Sig^{inner}(a_k, C, D, U)$, $k \leq |C|$;
$Step$ 3: Put a_k into *red*, where $Sig^{inner}(a_k, C, D, U) > 0$;// These attributes form the core of the given decision table
$Step$ 4: $i \leftarrow 1$, $R_1 = red$, $P_1 = \{R_1\}$ and $U_1 \leftarrow U$;
$Step$ 5: While $EF^{U_i}(red, D) \neq EF^{U_i}(C, D)$ Do
{Compute the positive region $POS_{P_i}^U(D)$,
$U_i = U - POS_{P_i}^U(D)$,
$i \leftarrow i + 1$,
$red \leftarrow red \cup \{a_0\}$, where $Sig^{outer}(a_0, red, D, U_i) = max\{Sig^{outer}(a_k, red, D, U_i), a_k \in C - red\}$,
$R_i \leftarrow R_{i-1} \cup \{a_0\}$,
$P_i \leftarrow \{R_1, R_2, \cdots, R_i\}$ };
$Step$ 6: return *red* and end.

The proposed accelerator provides an efficient accelerated strategy for heuristic feature selection in rough set theory. Note that each of the modified algorithms can choose the same attribute reduct as its original version, which possesses the same classification accuracy. Experiments carried out on nine UCI data sets show that these accelerated algorithms outperform their original counterparts, especially for large-scale data sets.

4 Conclusions and Further Work

In rough set theory, uncertainty of granulation and efficient feature selection algorithms have aroused wide concern and study. We focus on the review and analysis of several common uncertainty measures, an axiom approach of information granulation and granular space distance. Furthermore, an efficient accelerator is developed to accelerate a heuristic process of feature selection.

Based on the above results, some further researches are as follows.

- Uncertainty measures for generalized rough set models.
- Feature selection for the large-scale data sets by separating and fusing data sets.
- Efficient accelerated feature selection mechanism for hybrid data sets.
- Incremental feature selection algorithms for dynamic data sets.

It is our wish that this study provides new views and thinking on dealing with large-scale and complicated data sets in applications.

Acknowledgments. This work was supported by National Natural Science Fund of China (Nos. 71031006, 60903110), National Key Basic Research and Development Program of China(973)(No. 2007CB311002), The Research Fund for the Doctoral Program of Higher Education (20101401110002).

References

1. Bazan, J., Peters, J.F., Skowron, A., Nguyen, H.S., Szczuka, M.: Rough Set Approach to Pattern Extraction from Classifiers. J. Electron. Notes Theoretica. Comput. Sci. 82, 20–29 (2003)
2. Hu, X.H., Cercone, N.: Learning in relational databases: a rough set approach. Int. J. Comput. Intell. 11, 323–338 (1995)
3. Hu, Q.H., Xie, Z.X., Yu, D.R.: Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recognit. 40, 3509–3521 (2007)
4. Grzymala-Busse, J.W.: An algorithm for computing a single covering. In: Grzymala-Busse, J.W. (ed.) Managing Uncertainty in Expert Systems, p. 66. Kluwer Academic Publishers, Netherlands (1991)
5. Kryszkiewicz, M., Lasek, P.: FUN: fast discovery of minimal sets of attributes functionally determining a decision attribute. Trans. Rough Sets 9, 76–95 (2008)
6. Liang, J.Y., Dang, C.Y., Chin, K.S., Yam Richard, C.M.: A new method for measuring uncertainty and fuzziness in rough set theory. Int. J. General Syst. 31, 331–342 (2002)
7. Liang, J.Y., Shi, Z.Z.: The information entropy, rough entropy and knowledge granulation in rough set theory. Int. J. Uncertain., Fuzziness Knowl.-Based Syst. 12, 37–46 (2004)

8. Liang, J.Y., Shi, Z.Z., Li, D.Y., Wierman, M.J.: The information entropy, rough entropy and knowledge granulation in incomplete information system. Int. J. General Syst. 35, 641–654 (2006)
9. Liang, J.Y., Qian, Y.H.: Information granules and entropy theory in information systems. Sci. China., Ser. F. 51, 1427–1444 (2008)
10. Nguyen, H.S., Ślęzak, D.: Approximate reducts and association rules correspondence and complexity results. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 137–145. Springer, Heidelberg (1999)
11. Pedrycz, W., Vukovich, G.: Feature analysis through information granulation and fuzzy sets. Pattern Recognit. 35, 825–834 (2002)
12. Pawlak, Z.: Rough Sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Boston (1991)
13. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. 177, 3–27 (2007)
14. Qian, Y.H., Liang, J.Y., Pedrycz, W., Dang, C.Y.: Positive approximation: an accelerator for attribute reduction in rough set theory. Artifi. Intell. 174, 597–618 (2010)
15. Qian, Y.H., Liang, J.Y.: Combination entropy and combination granulation in rough set theory. Int. J. Uncertain., Fuzziness Knowl.-Based Syst. 16, 179–193 (2008)
16. Qian, Y.H., Liang, J.Y., Dang, C.Y.: Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. Int. J. Approx. Reasoning 50, 174–188 (2009)
17. Shannon, C.E.: A mathematical theory of communication. The Bell System Technology Journal 21, 372–423, 623–656 (1948)
18. Skowron, A.: Extracting laws from decision tables: a rough set approach. Comput. Intell. 11, 371–388 (1995)
19. Slezak, D.: Approximate entropy reducts. Fundam. Inform. 53, 365–390 (2002)
20. Wang, G.Y., Yu, H., Yang, D.C.: Decision table reduction based on conditional information entropy. Chin. J. Comput. 25, 759–766 (2002)
21. Wierman, M.J.: Measuring uncertainty in rough set theory. Int. J. General Syst. 28, 283–297 (1999)
22. Yao, J.T., Yao, Y.Y.: Information granulation for Web based information retrieval support systems. In: Data Mining and Knowledge Discovery: Theory, Tools and Technology. SPIE, vol. 5098, pp. 138–146 (2003)
23. Yao, Y.Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Inf. Sci. 178, 3356–3373 (2008)
24. Yao, Y.Y.: Neighborhood systems and approximate retrieval. Inf. Sci. 176, 3431–3452 (2006)
25. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. 46, 39–59 (1993)