DATA MINING ON INFORMATION SYSTEM USING FUZZY ROUGH SET THEORY

Phung Thi Thu Hien
University of Economic and Technical Industries, Hanoi

ABSTRACT

Today, thanks to the strong development of applications of information technology and Internet in many fields, a huge of database has been created. The number of records and the size of each record collected very quickly make it difficult to store and process information. Exploiting information sources from large databases effectively is an urgent issue and plays an important role in solving practical problems. In addition to traditional exploiting information methods, researchers have developed attribute reduction methods to reduce the size of the data space and eliminate irrelevant attributes. Our attribute reduction is based on the dependence between attributes in traditional rough set theory and in fuzzy rough set. The author built the tool which is inclusion degree and tolerance-based contingency table to solve the problem of finding the approximation set on set-valued information systems.

Keywords: *rough set; fuzzy rough set; set-valued information system; contingency table; reduct.*

Received: 14/11/2019; Revised: 26/12/2019; Published: 14/02/2020

KHAI PHÁ DỮ LIỆU SỬ DỤNG LÝ THUYẾT TẬP THÔ MỜ

Phùng Thị Thu Hiền
Trường Đại học Kinh tế Kỹ thuật Công nghiệp, Hà Nội

TÓM TẮT

Ngày nay với sự phát triển mạnh mẽ các ứng dụng công nghệ thông tin và Internet vào nhiều lĩnh vực, đã tạo ra nhiều cơ sở dữ liệu khổng lồ. Số lượng các bản ghi cũng như kích thước từng bản ghi được thu thập rất nhanh và lớn gây khó khăn trong việc lưu trữ và xử lý thông tin. Để khai thác hiệu quả nguồn thông tin từ các cơ sở dữ liệu lớn ngày càng trở thành vấn đề cấp thiết và đóng vai trò chủ đạo trong việc giải quyết các bài toán thực tế. Bên cạnh các phương pháp khai thác thông tin truyền thống, các nhà nghiên cứu đã phát triển các phương pháp rút gọn thuộc tính nhằm giảm kích cỡ của không gian dữ liệu, loại bỏ những thuộc tính không liên quan. Trong bài báo này, chúng tôi giới thiệu một số phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ, nghĩa là lý thuyết tập thô kết hợp với lý thuyết tập mờ. Đồng thời, tác giả xây dựng công cụ độ đo và bảng ngẫu nhiên tổng quát hóa để tìm tập xấp xỉ trong hệ thông tin đa trị.

Từ khóa: *Tập thô; tập mờ; tập thô mờ; hệ thông tin đa trị; bảng ngẫu nhiên; rút gọn.*

Ngày nhận bài: 14/11/2019; Ngày hoàn thiện: 26/12/2019; Ngày đăng: 14/02/2020

Email: Thuhiencn1@gmail.com

https://doi.org/10.34238/tnu-jst.2020.02.2330

1. Introduction

Attribute reduction is an important issue in data preprocessing steps which aims at eliminating redundant attributes to enhance the effectiveness of data mining techniques. Rough set theory by Pawlak [1] is an effective tool to solve feature selection problems with discrete attribute value domain.

Attribute reduction methods of rough set theory are performed on decision tables with numerical attribute value domain [2].

In fact, the domain attribute value of the decision table usually contains real-valued or symbolic values. In order to solve this problem, the rough set theory uses discrete methods of data before the implementation of attribute reduction methods. However, the degree of dependence of discrete values is not considered. For example, the two initial attribute values are converted into the same "Positive" value. However, we do not know which value is more positive, which means that discrete methods do not solve the problem of data semantics conservation. To solve this problem, Dubois D and his assistants proposed fuzzy rough set theory [3] which is a combination of rough set theory [4] and fuzzy set theory [5].

The fuzzy set theory assumes the preservation of the semantics of the data, and the rough set theory preserves the indiscernible of the data.

Similar to the traditional rough set model, fuzzy rough set uses fuzzy similarity relation to approximate fuzzy sets into upper approximation set and lower approximation set [6]. So far, many works have published the axiomatic systems, properties of operators in the fuzzy set of models. The work [7] studies attribute reduction method based on the fuzzy set theory approach based on dependency between attributes.

The article structure is as follows. Part II presents some basic concepts and attribute reduction method use of dependencies between attributes in traditional rough set theory. Part III presents some basic concepts in fuzzy rough set and attribute reduct based on fuzzy rough set. Part IV, the author built an algorithm for finding approximations set in in set-valued information systems. Finally, the conclusion and direction of the next development are given.

2. Basic definitions

This section presents some basic concepts in rough set theory and attribute reduction method uses dependencies between attributes [8].

An information system is a pair $IS = (U, A)$, where U is a finite nonempty set of objects and A is a finite nonempty set of attributes such that each $a \in A$ determines a map $a : U \to V_a$, where V_a is the value set of a.

Information system is a tuple $IS = (U, A)$; each sub-set $P \subseteq A$ determines one equivalence relation:

$$IND(P) = \{(u, v) \in U \times U \mid \forall a \in P, a(u) = a(v)\}$$

Partition of U generated by a relation $IND(P)$ is denoted as U / P, while

$U / P = \otimes \{a \in P : U / \mathrm{IND}(\{a\})\}$ where

$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \varnothing\}$.

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P.

Partition of U generated by a relation $IND(P)$ is denoted as U / P and is denoted as $[u]_P$, while $[u]_P = \{v \in U \mid (u, v) \in IND(P)\}$.

Considering information system $IS = (U, A)$, $B \subseteq A$ and $X \subseteq U$, $\underline{B}X = \{u \in U \mid [u]_B \subseteq X\}$ and $\overline{B}X = \{u \in U \mid [u]_B \cap X \neq \varnothing\}$ are called lower approximation and upper approximation of X respect to B respectively.

Considering information system $IS = (U, A)$, $P, Q \subseteq A$, then the positive region can be defined as $POS_P(Q) = \bigcup_{X \in U/Q} (\underline{P}X)$

The positive region contains all objects of U that can be classified to classes of U/Q using the knowledge in attributes P.

For $P, Q \subseteq A$, the quantity $k = \gamma_P(Q)$ represents the dependence of Q on P, denoted $P \Rightarrow_k Q$, can be defined as

$$k = \gamma_P(Q) = \frac{\text{POS}_P(Q)}{|U|} \quad (1)$$

with $|S|$ as the force of S.

If k = 1, Q depends totally on P, if 0 <k < 1, Q depends partially (in a degree k) on P.

For $P \subseteq A$, $X \subseteq U$, member function of object $x \in U$ is defined:

$$\mu_X^P : U \to [0,1] \text{ and } \mu_X^P = \frac{|[x]_P \cap X|}{|[x]_P|} \quad (2)$$

Membership is characteristic of the inclusion of $[x]_P$ in the object set X. From the definition of the membership function, the formula (1) calculates the dependence of the attribute as follows:

$$k = \gamma_P(Q) = \frac{\sum_{x \in U} \mu^P_{\text{POS}_P(Q)}(x)}{|U|} \quad (3)$$

A decision system (or decision table) is an information system (U, A), where A includes two separate subsets: condition attribute subset C and decision attribute subset D. So that, a decision system *(DS)* could be written as $DS = (U, C \cup D)$ where $C \cap D = \varnothing$.

Decision table $DS = (U, C \cup D)$ is called consistent if and only if $POS_C(D) = U$. Opposite *DS* is inconsistent.

Attribute reduction in decision system is a process of selecting the minimal sub-set of conditional attribute set, preserving classified information of the decision systems. In traditional rough set, Pawlak [9] introduced the concept of reduction based on the positive region and developed a heuristic algorithm for finding the best reductions of the decision table based on the criterion of importance of the attribute.

Definition 1. Let $DS = (U, C \cup D)$ be a decision table, $R \subseteq C$, if

1) $POS_R(D) = POS_C(D)$

2) $\forall r \in R,\ POS_{R-\{r\}}(D) \neq POS_C(D)$

then R is a reduct set of C based on the positive region.

Definition 1 combines the definition of dependency between attributes in formula (1), attribute set $R \subseteq C$ is a reduct set of C based on the positive region if $\gamma_R(D) = \gamma_C(D)$ and $\forall r \in R,\ \gamma_{R-\{r\}}(D) \neq \gamma_C(D)$.

Definition 2. Let $DS = (U, C \cup D)$ be a decision table, $B \subset C$ and $b \in B - C$. The importance of attribute b for attribute set B is defined as:

$$IMP_B(b) = \gamma_{B \cup \{b\}}(D) - \gamma_B(D) = \frac{\left|\text{POS}_{B \cup \{b\}}(D)\right| - \left|\text{POS}_B(D)\right|}{|U|} \quad (4)$$

With the assumption $|\text{POS}_\varnothing(D)| = 0$.

We see that $\left|\text{POS}_{B \cup \{b\}}(D)\right| \geq \left|\text{POS}_B(D)\right|$, so $IMP_B(b) \geq 0$.

$IMP_B(b)$ calculated by changing number of dependence of D on B when adding attribute set b into B and $IMP_B(b)$ is larger the greater amount of changing, or attribute set b is more important and reversing.

The importance of this attribute is the criterion for selecting attributes in the heuristic algorithm for the find the reduct set of decision tables.

The ideas of the algorithm initials with empty attribute set $R = \varnothing$, repeat adding

the most important attribute set into set R until finding reduct.

Algorithm 1. Algorithm finds the best reduction set using dependencies between attributes [10].

Input: Decision table $DS = (U, C \cup D)$

Output: a reduct R.

1. $R \leftarrow \varnothing$;
2. While $\gamma_R(D) \neq \gamma_C(D)$ do
3. Begin
4. For $c \in C - R$
   Calculated
   $IMP_R(c) = \gamma_{R\cup\{c\}}(D) - \gamma_R(D)$;
5. Select $c_m \in C - R$ in order to
   $IMP_R(c_m) = \underset{c\in C-R}{Max}\{IMP_R(c)\}$;
6. $R \leftarrow R \cup \{c_m\}$;
7. End;
8. Return R;

In the next section, we present algorithm attribute reduct based on decision tables by fuzzy rough set

3. Attribute reduct based on fuzzy rough set

The fuzzy rough set is based on a combination of rough set theory and fuzzy set theory to approximate fuzzy sets using fuzzy similarity relations [11].

A relation R defined on U is called fuzzy equivalence relation if it satisfies the following conditions:

1) Reflectivity: $(\mu_S(x,x) = 1)$

2) Symmetry: $(\mu_S(x,y) = \mu_S(y,x))$

3) Transitivity:
$(\mu_S(x,z) \geq \mu_S(x,y) \wedge \mu_S(y,z))$

Similar in traditional rough-set theory, based on fuzzy similarity relation, each attribute set $P \subseteq A$ defines a fuzzy partition as follows:

$$U/P = \otimes\{a \in P : U/\mathrm{IND}(\{a\})\} \quad (5)$$

for $A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \varnothing\}$

Each element of U/P is a fuzzy equivalence class $[x]_P$ with $(\mu_{[x]_P}(y) = \mu_P(x,y))$.

Membership function of objects in fuzzy equivalence class is defined based on fuzzy-rough set theory:

$$\mu_{F_1\cap\ldots\cap F_n}(x) = \min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (6)$$

Based on the fuzzy equivalence classes, the concept of the lower and upper approximations is expanded fuzzy lower approximation set and fuzzy upper approximation.

With attribute set $P \subseteq A$, the membership function of objects in the subset of fuzzy sets and the set of fuzzy approximations is defined:

$$\mu_{\underline{P}X}(y) = \sup_{F\in U/P} \min\left(\mu_F(x), \inf_{y\in U} \max\{1-\mu_F(y), \mu_X(y)\}\right) \quad (7)$$

$$\mu_{\overline{P}X}(x) = \sup_{F\in U/P} \min\left(\mu_F(x), \sup_{y\in U} \min\{\mu_F(y), \mu_X(y)\}\right) \quad (8)$$

The symbols *inf X*, *sup X*, respectively are the lower and upper of the set *X*. *F* is fuzzy equivalence class of the fuzzy partition *U/P*. Then $\langle PX, \overline{P}X \rangle$ is called a fuzzy rough set.

In traditional rough set theory, concept of positive region is defined as the intersection of all subsets of the approximation. With $P, Q \subseteq A$, the membership function of the fuzzy positive in the fuzzy rough set is defined:

$$\mu_{\mathrm{POS}_P(Q)}(x) = \sup_{X\in U/Q} \mu_{\underline{P}X}(x) \quad (9)$$

Based on the fuzzy positive region concept, the fuzzy function represents the dependence between the attributes defined as follows:

$$\lambda_P(Q) = \frac{|\mu_{\mathrm{POS}_P(Q)}(x)|}{|U|} = \frac{|\sum_{x\in U} \mu_{\mathrm{POS}_P(Q)}(x)|}{|U|} \quad (10)$$

The importance of the attribute using the fuzzy function in formula (10) is described as follows:

$$IM_B(b) = \lambda_{B\cup(b)}(\mathrm{D}) - \lambda_B(\mathrm{D}) \quad (11)$$

Attribute reduction algorithm in the decision table using in formula (10) is described as follows:

Algorithm 2. Algorithm finds the best reduction set

Input: Decision table $DS = (U, C \cup D)$

Output: a reduct R.

1. $R \leftarrow \varnothing$;
2. $\lambda_{\varnothing}(D) = 0$;
3. While $\lambda_R(D) \neq \lambda_C(D)$ do
4. Begin
5. For $c \in C - R$
   calculated $IM_R(c) = \lambda_{R\cup\{c\}}(D) - \lambda_R(D)$;
6. Select $c_m \in C - R$ in order to
   $IM_R(c_m) = \underset{c\in C-R}{Max}\{IM_R(c)\}$;
7. $R \leftarrow R \cup \{c_m\}$;
8. End;
9. Return R;

4. Building tools to find approximation in set-valued information system

4.1. Set-valued information system [12]

An information system is a quadruple $ISS = (U, A, V, f)$, where U is a non-empty finite set of objects; A is a non-empty finite set of attributes; V is the set of attributes values, f is a mapping from $U \times A$ to V, where $f : U \times A \to 2^V$ is a set-valued mapping. In the convention the abbreviation $ISS = (U, A, V, f)$ is $ISS = (U, A)$.

In the set-valued information system $ISS = (U, A)$, for $B \subseteq A$, the tolerance relation T_B is defined as:

$$T_B = \{(u, v) \in U \times U \mid \forall b \in B, u(b) \cap v(b) \neq \varnothing\}$$

Put $[u]_{T_B} = \{v \in U \mid (u, v) \in T_B\}$, $[u]_{T_B}$ is called a tolerance class corresponding to T_B. The notation $U/T_B = \{[u]_{T_B} \mid u \in U\}$ represents the set of all tolerance classes corresponding to the relation T_B, then U/T_B formed a cover of U because the tolerance classes in U/T_B can intersect and $\bigcup_{u\in U}[u]_{T_B} = U$. Oviously, if $C \subseteq B$ then $[u]_{T_B} \subset [u]_{T_C}$ or all $u \in U$.

Let $ISS = (U, A)$ be a set-valued information system. For any $B \subseteq A$ we denote by $U/T_B = \{[u]_{T_B} : u \in U\}$ the tolerance class related to object $u \in U$. We denote $U/T_B = \{[u]_{T_B} : u \in U\}$ the family of all tolerance classes of T_B.

Set-valued decision information system is a quadruple $DSS = (U, C \cup \{d\}, V, f)$, where U is a non-empty finite set of objects; C is a finite set of condition attributes, d is a decision attribute with $C \cap \{d\} = \varnothing$; $V = V_C \cup V_d$, where V_C is the set of condition attribute values, V_d is the set of decision attribute values; f is a mapping from $(U \times (C \cup \{d\})$ to V such that $f : U \times C \to 2^{V_C}$ is a set-valued mapping, and $f : U \times d \to V_d$ is a single-valued mapping. The set-valued decision information system can always be expressed as a table, called set-valued decision table.

Given a set-valued information system $ISS = (U, A)$, $B \subseteq A$. The lower and upper approximations of $X \subseteq U$ in terms of tolerance relation T_B are defined as:

$$\underline{T_B}(X) = \{x \in U : [x]_{T_B} \subseteq X\};$$

$$\overline{T_B}(X) = \{x \in U : [x]_{T_B} \cap X \neq \varnothing\};$$

4.2. Building tools

Definition 3. (Contingency Table)

Let $DSS = (U, C \cup \{d\})$ is set-valued decision information system, V_d be the set of decision values in decision table,

and let $U/IND(B)=\left\{[u_1]_B,[u_2]_B,....,[u_{n_s}]_B\right\}$ be partition of U defined by indiscernibility relation $IND(B)$ for $B\subseteq C$. Contingency table CT_B related to B is a two dimensional table $CT_B=[CT_B[i,j]]_{i\in\{1,...,n_B\}}^{j\in\{1,...,|V_d|\}}$ where:

$CT_B[i,j]=|\{x\in U: x\in[x_i]_B\wedge d(x)=j\}|$.

Using this structure quickly determines the frequency of occurrences of attributes in the matrix, without having to check the appearance of attributes in every cell in the decision table.

Definition 4. (Tolerance-Based Contingency Table)

Let $DSS=(U,C\cup\{d\})$ is set-valued decision information system, V_d be the set of decision values in decision table, let T_B be a tolerance relation for $B\subseteq C$.

The tolerance based contingency table is a two-dimensional table $TCT_B=\left[TCT[i,j]\right]_{i\in\{1,...n_B\}}^{j\in\{1,..|V_d|\}}$, which is defined as follows:

$$TCT_B[i,j]=\left|\left\{u\in U\mid u\in[u_i]_B \text{ và } d(u)=j\right\}\right|$$

Tolerance-Based Contingency Table is a table that shows the distinction of the tolerance classes relative to the decision attribute.

4.3. Algorithm for finding approximations on set-valued information systems

Algorithm 3. Finding upper and lower approximation of X

Input: Set-valued information table

$ISS=(U,A)$, $X\subseteq U$, $B\subseteq A$,

Tolerance relation T_B,

$U/IND(B)=\{1,2,...,n_B\}$.

Output: Upper and lower approximation of *X*.

1. Create the decision table $DT=(U,C\cup\{d_X\})$
2. Generate CT_B;
3. Generate TCT_B from CT_B;
4. *for* $i\in\{1,2,.....,n_B\}$ *do*
5. Compute a inclusion degree

$$\nu_i=\frac{TCT[i,1]}{TCT[i,1]+TCT[i,0]}$$

6. if ($\nu_i=1$) then
7. LowerAppr← {i}
8. else
9. if ($v_i > 0$) then
10. Upper Appr ← {i}
11. end if
12. end if
13. end for

5. Conclusion

Fuzzy rough set model proposed by D. Dubois is a combination of rough set theory and fuzzy set theory. The rough set theory preserves indiscernible of data, fuzzy set theory preserves the semantics of the data. So that, fuzzy rough set tool is considered to be more efficient than the rough set tool in property reduction and filtering on information systems with domain of continuous attribute value or semantic values, fuzzy values.

In this paper, based on the attribute reduction using the dependence between attributes in traditional rough set theory and the fuzzy rough set, we demonstrate that the fuzzy rough set of approaches on the original data would have been a minimized set of reductions than the set of reductions of the traditional rough set if we use the membership function of the fuzzy set to discrete the data.

At the same time, the article builds on the new data structure as inclusion degree and tolerance-based contingency table in the set-valued information system. This is a powerful tool for constructing the algorithm computing upper and lower approxmation on set-valued information systems. Our future research direction is to build an algorithm for finding reduct set in the case of updating objects on set-valued information systems.

REFERENCES

[1]. M. M. Deza and E. Deza, *Encyclopedia of Distances*, Springer, 2009.

[2]. D. Dubois and H. Prade, *Putting rough sets and fuzzy sets together, Intelligent Decision Support*, Kluwer Academic Publishers. Dordrecht, 1992.

[3]. D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, 17, pp. 191-209, 1990.

[4]. L. A. Zadeh, "Fuzzy sets," *Information and Control*, 8, p. 338353, 1965.

[5]. Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, 11(5), pp. 341-356, 1982.

[6]. Z. Pawlak, *Rough sets: Theoretical Aspects of Reasoning About Data*, Kluwcr Aca-demic Publishers, 1991.

[7]. R. Jensen and Q. Shen., "Fuzzy-Rough Sets for Descriptive Dimensionality Reduction," Proceedings of the 11th International Conference on Fuzzy Systems, pp. 29-34, 2002.

[8]. Y. Y. Yao, "On combining rough and fuzzy sets," Proceedings of the CSC'95 Workshop on Rough Sets and Database Mining, Lin, T.Y. (Ed.), San Jose State University, 1995, 9 pages.

[9]. Yao Y. Y., "A Comparative Study of Fuzzy Sets and Rough Sets," *Information Sciences*, vol.109, p. 2147, 1998.

[10]. Y. Y. Guan, and H. K. Wang, "Set-valued information systems," *Information Sciences*, 176(17), pp. 2507-2525, 2006.

[11]. Y. Qian, C. Dang, J. Liang, and D. Tang, "Set-valued ordered information systems," *Information Sciences*, 179 (16), pp. 2809–2832, 2009.

[12]. C. R. Wang and F. F. Ou, "An Attribute Reduction Algorithm in Rough Set Theory Based on Information Entropy", International Symposium on Computational Intelligence and Design, IEEE ISCID, pp. 3-6, 2008.