Knowledge-Based Systems 83 (2015) 138–148

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Update of approximations in composite information systems

Shaoyong Li [a,b,c], Tianrui Li [a,c,*], Jie Hu [a,c]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
[b] Naval Aviation Institute, PLA, Huludao 125001, China
[c] Key Laboratory of Cloud Computing and Intelligent Technique, Sichuan Province, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 22 June 2014
Received in revised form 16 March 2015
Accepted 17 March 2015
Available online 25 March 2015

Keywords:
Rough set
Tolerance relation
Dominance relation
Composite information system
Incremental update

ABSTRACT

In this paper, a new composite rough set model is proposed to process incomplete composite information with criteria and regular attributes. Following that, some strategies for incrementally updating approximations of composite rough sets when adding or removing some objects are discussed and the corresponding incremental algorithms are designed. Several numerical examples illustrate the feasibilities of the composite rough set model and the incremental strategies, respectively. Experimental evaluation shows that the incremental algorithms can effectively reduce the running time than their counterparts.

1. Introduction

In our daily life, we often need to make predictions or decisions for the coming tasks. For example, sometimes people decide to travel or not mainly depend on a weather forecast without considering the other factors. It is important that an accurate forecast can provide the opportunities for people to reduce loss before the disaster weather comes. Analogously, an accurate prediction of earthquake can also provide the opportunities for people to reduce damages before the calamity arrives. The estimation of risk-at-value is often necessary for business investment and the other related activities. Practically, these forecasts, predictions and estimations are provided by the experts in corresponding fields through processing a great mass of data in information systems.

In many applications, information systems may contain more than one type of values, e.g., an information system for weather forecast includes values of barometric pressure, visibility, wind speed, wind direction, temperature and cloudiness. Among these, the wind direction which is a regular attribute does not have a preference-ordered domain, while others have their preference-ordered domains and the corresponding attributes can be regarded as criteria. A weather observer sometimes cannot decide the value of wind direction due to weak air pressure, which means that value is uncertain. And the uncertain value may be regarded as a type of missing value. If there may exist a missing value in a regular attribute's domain and there is no missing value in the criteria's domains, then we called this information system for weather forecast as an Incomplete Composite Information System (ICIS) with criteria and regular attributes.

The rough set theory proposed by Pawlak is one of information processing tools, which can process uncertain, fuzzy and inconsistent information [30,31]. In many cases, the notions in Pawlak's rough set model are too strict to satisfy the requirements of real applications. Therefore, many generalized rough set models have been proposed in different domains [12,15,19,33,35–37,40,46]. However, these rough set models are based on a binary relation which can only process a type of values and cannot be employed to process composite information. In recent years, some scholars have attempted to propose new rough set models to process composite information. To the best of our knowledge, Greco et al. firstly introduced the idea of using a global binary relation by the intersection of indiscernibility, similarity and dominance relations to process information [13]. Following their idea, An and Tong redefined the definitions of lower and upper approximations of upward and downward unions of decision classes [2]. Abu-Donia presented some types of rough set' approximations based on multi-knowledge base, which involved reflexive, tolerance, dominance and equivalence relations [1]. Zhang et al. introduced a composite rough set model which can be used to deal with different kinds of relations, e.g., indiscernibility, neighborhood, tolerance and

* Corresponding author at: School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China.

E-mail addresses: meterer@163.com (S. Li), trli@swjtu.edu.cn (T. Li), jiehu@swjtu.edu.cn (J. Hu).

http://dx.doi.org/10.1016/j.knosys.2015.03.016
0950-7051/

characteristic relations, simultaneously [40,44]. Even though these rough set models can be employed to process their corresponding composite information respectively, they cannot be employed directly to mine the useful knowledge hidden in the ICIS. This paper introduces a new Composite Rough Set Model (CRSM) based on a composite binary relation by the intersection of tolerance and dominance relations. The CRSM can be applied to the ICIS directly when all criteria and attributes are considered at the same time.

In many real-time cases, information systems may evolve over time, that is to say, some new information becomes available continuously while some information is no longer useful. In rough set theory, many scholars have paid their attention to process dynamic information and used the incremental techniques to maintain rules or approximations under dynamic data environment [4–11,14,16–18,20–29,32,34,38,39,41–43,45]. These studies have significantly enriched the theory of rough set and guided a way for dynamic data mining, even big data mining. To apply CRSM in dynamic data environment, we also use the incremental techniques in discussion of strategies on how to update approximations when the object set varies with time in this paper. Then we design the corresponding algorithms for incrementally updating approximations when adding or removing some objects. Two numerical examples illustrate the feasibility of these strategies. Experimental evaluation validates that the incremental algorithms can reduce time than their counterparts.

The remainder of this paper is organized as follows. In Section 2, we introduce the CRSM and illustrate it by a numerical example. An algorithm for computing approximations in CRSM is also provided. Section 3 presents strategies for updating approximations in CRSM when adding or removing some objects and the corresponding algorithms. Section 4 shows the experimental evaluation on several public data sets downloaded from UCI [3]. The paper ends with conclusions and further research work in Section 5.

2. Preliminaries

An ICIS can be denoted by a quarter-tuple $S = (U, A, V, f)$, where U is a non-empty finite set of objects, called the universe. $A = C_1 \cup C_2 \cup \{d\}$ is a non-empty finite set of attributes. d is a decision criterion. C_1 is a set of attributes with preference-ordered domains which may be regarded as criteria. C_2 is a set of regular attributes whose domains may contain the missing value "$*$", which means it may exist that $* \in V_a$ for any $a \in C_2$. $V = \cup_{a\in A} V_a$ is regarded as the domain of all attributes, where V_a is the domain of attribute a. $f : U \times A \to V$ is an information function such that $f(x,a) \in V_a, \forall a \in A, \forall x \in U$.

Assume $P_1 \subseteq C_1$ and $P_1 \neq \emptyset$. Since P_1 is a set of criteria, there is a dominance relation with respect to P_1 [12], which is denoted by

$$D(P_1) = \{(x,y) \in U \times U | \forall a \in P_1, f(x,a) \geqslant f(y,a)\}$$

Assume $P_2 \subseteq C_2$ and $P_2 \neq \emptyset$. Since P_2 is a set of regular attributes and "$*$" may belong to $\cup_{a\in P_2} V_a$, there is a tolerance relation with respect to P_2 [19], which is denoted by

$$T(P_2) = \{(x,y) \in U \times U | \forall a \in P_2, f(x,a) = f(y,a) \vee f(x,a) = * \vee f(y,a) = *\}$$

$\forall P^* \subseteq C$ and $P^* \neq \emptyset$, we may use DRSA [12] or tolerance-based rough set approach [19] directly if $P^* = P_1$ or $P^* = P_2$, respectively. But when $P^* = P_1 \cup P_2$, there is no a rough set approach that can be used directly. Therefore, a new rough set model is required for this case. In many cases, a generalized rough set model for processing information describing a specific problem mainly differs from the others on its binary relation and approximations' definitions [12,15,19,33,35,36,40,44,46]. Here, we take each attribute or criterion in P^* into accounted and then construct a new composite binary relation $CR(P^*)$ by the intersection of $D(P_1)$ and $T(P_2)$ following the viewpoints in [1,2,13,40], which is denoted by

$$CR(P^*) = \{(x,y) \in U \times U | (x,y) \in D(P_1) \cap T(P_2)\}$$

$CR(P^*)$ is also written as CR_{P^*}. Given $P^* \neq \emptyset$, if $P_1 = \emptyset$, then CR_{P^*} becomes a tolerance relation; and if $P_2 = \emptyset$, then CR_{P^*} becomes a dominance relation.

With CR_{P^*}, the basic information granules in CRSM are defined as follows:

- $CR^+_{P^*}(x) = \{y \in U | (y,x) \in CR_{P^*}\}$ called as P^*-left neighborhood of x with respect to CR_{P^*};
- $CR^-_{P^*}(x) = \{y \in U | (x,y) \in CR_{P^*}\}$ called as P^*-right neighborhood of x with respect to CR_{P^*}.

Since d is a decision criterion in ICIS, the concepts approximated in CRSM still are the upward and downward unions of decision classes following that of DRSA [12]. Let $Cl = \{Cl_n, n \in T\}, T = \{1, \ldots, m\}$, be a family of decision classes. Assume $\forall r, s \in T$, such that $r > s$, the objects from Cl_r are preferred to the objects from Cl_s. For any decision class Cl_n, its upward and downward unions are defined respectively as follows:

$$Cl_n^{\geqslant} = \bigcup_{s \geqslant n} Cl_s, \quad Cl_n^{\leqslant} = \bigcup_{s \leqslant n} Cl_s, \quad \forall n,\ s \in T$$

$x \in Cl_n^{\geqslant}$ means "x belongs to at least class Cl_n", and $x \in Cl_n^{\leqslant}$ means "x belongs to at most class Cl_n".

In CRSM, P^*-lower approximation and P^*-upper approximation of $Cl_n^{\geqslant}$ are defined respectively as follows:

$$\underline{P^*}(Cl_n^{\geqslant}) = \{x \in U | CR^+_{P^*}(x) \subseteq Cl_n^{\geqslant}\} \quad (1)$$

$$\overline{P^*}(Cl_n^{\geqslant}) = \{x \in U | CR^-_{P^*}(x) \cap Cl_n^{\geqslant} \neq \emptyset\} \quad (2)$$

Analogously, P^*-lower approximation and P^*-upper approximation of $Cl_n^{\leqslant}$ are defined respectively as follows:

$$\underline{P^*}(Cl_n^{\leqslant}) = \{x \in U | CR^-_{P^*}(x) \subseteq Cl_n^{\leqslant}\} \quad (3)$$

$$\overline{P^*}(Cl_n^{\leqslant}) = \{x \in U | CR^+_{P^*}(x) \cap Cl_n^{\leqslant} \neq \emptyset\} \quad (4)$$

Then P^*-boundaries of $Cl_n^{\geqslant}$ and $Cl_n^{\leqslant}$ are defined respectively as follows:

$$Bn_{P^*}(Cl_n^{\geqslant}) = \overline{P^*}(Cl_n^{\geqslant}) - \underline{P^*}(Cl_n^{\geqslant})$$

$$Bn_{P^*}(Cl_n^{\leqslant}) = \overline{P^*}(Cl_n^{\leqslant}) - \underline{P^*}(Cl_n^{\leqslant})$$

Proposition 1. *The following items hold.*

1. $\underline{P^*}(Cl_n^{\geqslant}) = U - \overline{P^*}(Cl_{n-1}^{\leqslant})$ and $\overline{P^*}(Cl_n^{\geqslant}) = U - \underline{P^*}(Cl_{n-1}^{\leqslant}), n = 2, \ldots, m.$
2. $\underline{P^*}(Cl_n^{\leqslant}) = U - \overline{P^*}(Cl_{n+1}^{\geqslant})$ and $\overline{P^*}(Cl_n^{\leqslant}) = U - \underline{P^*}(Cl_{n+1}^{\geqslant}), n = 1, \ldots, m-1.$

Proof 1.

1. $\because \underline{P^*}(Cl_n^{\geqslant}) = \{x \in U | CR^+_{P^*}(x) \subseteq Cl_n^{\geqslant}\} = \{x \in U | CR^+_{P^*}(x) \subseteq U - Cl_{n-1}^{\leqslant}\} = \{x \in U | CR^+_{P^*}(x) \cap Cl_{n-1}^{\leqslant} = \emptyset\} = U - \{x \in U | CR^+_{P^*}(x) \cap Cl_{n-1}^{\leqslant} \neq \emptyset\} = U - \overline{P^*}(Cl_{n-1}^{\leqslant})$, $\therefore \underline{P^*}(Cl_n^{\geqslant}) = U - \overline{P^*}(Cl_{n-1}^{\leqslant})$ holds for $n = 2, \ldots, m$. Similarly, $\overline{P^*}(Cl_n^{\geqslant}) = U - \underline{P^*}(Cl_{n-1}^{\leqslant})$ holds for $n = 2, \ldots, m$.
2. Its proof is similar to that of item 1. □

It is obvious that $\underline{P^*}(Cl_1^{\geqslant}) = \overline{P^*}(Cl_1^{\geqslant}) = \underline{P^*}(Cl_m^{\leqslant}) = \overline{P^*}(Cl_m^{\leqslant}) = U$.