International Journal of Approximate Reasoning 55 (2014) 1764–1786

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A rough set-based incremental approach for learning knowledge in dynamic incomplete information systems

Dun Liu [a,*], Tianrui Li [b], Junbo Zhang [b]

[a] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China
[b] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 5 January 2013
Received in revised form 28 February 2014
Accepted 27 May 2014
Available online 2 June 2014

Keywords:
Rough set theory
Incomplete information systems
Incremental learning
Knowledge discovery

ABSTRACT

With the rapid growth of data sets nowadays, the object sets in an information system may evolve in time when new information arrives. In order to deal with the missing data and incomplete information in real decision problems, this paper presents a matrix based incremental approach in dynamic incomplete information systems. Three matrices (support matrix, accuracy matrix and coverage matrix) under four different extended relations (tolerance relation, similarity relation, limited tolerance relation and characteristic relation), are introduced to incomplete information systems for inducing knowledge dynamically. An illustration shows the procedure of the proposed method for knowledge updating. Extensive experimental evaluations on nine UCI datasets and a big dataset with millions of records validate the feasibility of our proposed approach.

1. Introduction

Data mining (also known as knowledge discovery in databases) is the process of extracting patterns from large data sets to analyze data from different perspectives and summarize them into useful information. Many data mining approaches, such as association rule mining [27], sequential pattern mining [25], text mining [32] and temporal data mining [40], are utilized to discover potentially valuable patterns, associations, trends, sequences and dependencies [26]. However, with the fast growth of data sets in real-life applications, it brings a big challenge to quickly acquire the useful information with dynamic data mining techniques.

As an efficient data analysis' technique, the rough set based incremental approaches have become one of the hot topics on extraction of knowledge from the changing data sets in recent decades. With respect to the different angles to recognize the dynamics in rough sets, there exist two main viewpoints. The first one is based on the view of information table. Since an information table consists of data objects (items, records or instances), data attributes (features) and data attribute values [50] recent researches mainly focus on the three types of its variations, namely, variation of objects [1,8,23,36,42–44,52,68,71,80,82,83], variation of attributes [6,9,37–39,47,53,81,85] and variation of attributes' values [7,45,46,48]. The second one is based on the view of pre-topology [57]. The classification of dynamics in rough sets is divided into two aspects: synchronic dynamics (knowledge evolves in time) and diachronic dynamics (changes from one point of view to another) [10]. Furthermore, Ciucci [10,13] listed four main streamlines to investigate dynamics in rough sets, namely, lower and

[*] Corresponding author.
E-mail addresses: newton83@163.com (D. Liu), trli@swjtu.cn (T. Li), JunboZhang86@163.com (J. Zhang).

http://dx.doi.org/10.1016/j.ijar.2014.05.009
0888-613X/© 2014 Elsevier Inc. All rights reserved.

upper approximations [4,11,63,64,80–82], reducts and rules [12,24,65,77], quality indexes [20,21,42–44] and formal logic [29,56,66]. To sum up, both viewpoints provide a basic and clear framework on dynamic studies of rough sets.

However, many databases with thousands of items in real-life applications lead to lots of challenges. The variation of objects significantly affects the knowledge updating. Then, a series of effective algorithms on the incremental learning of approximations, reducts and rules were proposed for knowledge updating to improve the computational efficiency. Shan and Ziarko presented a discernibility-matrix based incremental methodology to find all maximally generalized rules [68]. Bang and Bien proposed another incremental inductive learning algorithm to find a minimal set of rules for a decision table without recalculating all the set of instances when another instance is added into the universe [1]. Tong and An developed an algorithm based on the ∂-decision matrix for incremental learning rules. They listed seven cases that would happen when a new sample enters the system [71]. Zheng and Wang developed a rough set and rule tree based incremental knowledge acquisition algorithm, RRIA, to update knowledge more quickly when new objects are added or removed from a given dataset [85]. Hu et al. constructed a novel incremental attribute reduction algorithm when new objects are added into a decision information system [24]. Blaszczynski and Slowinski discussed the incremental induction of decision rules from dominance-based rough approximations to select the most interesting representatives in the final set of rules [3]. Fan et al. proposed an approach of incremental rule induction based on rough sets [14]. In addition, Liu et al. proposed an incremental approach as well as its algorithm for inducing interesting knowledge when objects change over time [42,43]. Then, Liu et al. further introduced the incremental matrix and presented a new optimization approach for knowledge discovery [44]. Followed by Liu's work, Li et al. proposed an incremental approach for updating approximations in dominance-based rough sets [36]. Zhang et al. proposed a method for dynamic data mining based on neighborhood rough sets [80], and they further presented a parallel method for computing rough set approximations [82,84]. In our studies, we mainly consider the missing data in real databases, and the rough set-based incremental approach for inducing knowledge in incomplete information systems (IIS) is carefully investigated.

This paper is to focus on the dynamic knowledge discovery based on rough set theory with the variation of the object set in IIS. The rest of the paper is organized as follows. We provide the basic concepts of rough sets under IIS in Section 2. We introduce the support matrix, the accuracy matrix and the coverage matrix to illuminate the incremental approach, the related updating strategies and algorithms for learning knowledge in IIS when objects change is given in Section 3. An example is employed to illustrate the proposed model and experimental results on nine UCI datasets and a big dataset with millions of records are presented in Section 4. The paper ends with conclusions and further research topics in Section 5.

2. Preliminaries

In this section, we first briefly review the concepts of rough sets as well as their extensions in IIS from [17–19,30,31,34,35,41,49,51,58,59,61–64,69,70,75]. Then, we introduce some necessary concepts of knowledge discovery from [5,15,42–44].

2.1. Pawlak rough set model

For an approximation space $\mathcal{K} = (U, R)$, let U be a finite and non-empty set called the universe and $R \subseteq U \times U$ a binary relation on U. $\mathcal{K} = (U, R)$ is characterized by an information system $S = (U, A, V, f)$. S is called a decision table with $A = C \cup D$ and $C \cap D = \emptyset$, where C denotes the condition attribute set and D denotes the decision attribute set. $V = \bigcup_{a \in A} V_a$, V_a is a domain of the attribute a. $f : U \times A \rightarrow V$ is an information function such that $f(x, a) \in V_a$ for every $x \in U$, $a \in A$.

Each non-empty subset $B \subseteq C$ determines a binary indiscernibility relation R_B as follows.

$$R_B = \left\{(x, y) \in U \times U \,\middle|\, f(x, a) = f(y, a),\ \forall a \in B\right\}.$$

R_B is an equivalence relation. It constitutes a partition of U, denoted as U/B. The equivalence relation R_B divides U into several disjoint subsets named equivalence classes given by: $U/R_B = \{[x]_B \mid x \in U\}$. Suppose there are two elements $x, y \in U (x \neq y)$ indistinguishable under R_B, we say x and y belong to the same equivalence class. The equivalence class including x is denoted as $[x]_{R_B}$, where $[x]_{R_B} = \{y \in U \mid (x, y) \in R_B\}$. For simplicity, U/R_B is denoted as U/B.

Pawlak rough set model is based on the equivalence relation. The elements in an equivalence class satisfy reflexive, symmetric and transitive properties. It also does not allow the missing data and requires the information table should be complete. However, missing data appears frequently in real-life applications, *e.g.*, in the survey sampling, it may arise out of poorly designed questionnaires, non-response by an interviewer, or errors made by the interviewer [19]. Lipski discussed the semantic issues connected with incomplete information databases [41]. Ibrahim et al. did a comparative review on four missing-data methods for generalized linear models [28]. Leung et al. dealt with knowledge acquisition in incomplete information systems using rough set theory [33]. Saar-Tsechansky et al. investigated three methods for applying classification trees to instances with missing values [67]. Therefore, many extended models of Pawlak rough sets were presented to deal with missing data in IIS.