International Journal of Approximate Reasoning 81 (2017) 28–48

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Incremental fuzzy probabilistic rough sets over two universes ☆

Jie Hu [a], Tianrui Li [a,*], Chuan Luo [b], Hamido Fujita [c], Shaoyong Li [d]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] College of Computer Science, Sichuan University, Chengdu, 610065, China
[c] Fac. of Software and Information Science, Iwate Prefectural University, 020-0693, Iwate, Japan
[d] Naval Aviation Institute, PLA, Huludao, 125001, China

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 2 September 2016
Accepted 3 November 2016
Available online 14 November 2016

Keywords:
Rough sets
Two universes
Fuzzy relation
Approximations
Incremental updating

ABSTRACT

The fuzzy Information System over Two Universes (ISTU) formalizing a data table corresponding to two universes as well as their relations is common in real-world applications, e.g., recommender system and clinical diagnosis system. The fuzzy probabilistic rough sets over two universes (FPRSMTU) can deal with a fuzzy relation and allow a tolerance inaccuracy in the construction of rough approximations in a fuzzy ISTU, which is a generalization of classic rough sets with fuzzy and probabilistic theories. As a necessary step for knowledge discovery based on rough sets, the fuzzy rough approximations of fuzzy ISTU need to be updated efficiently under dynamic data environment. Incremental technique is an efficient approach for dynamic information processing by making full use of previously obtained knowledge. In this paper, incremental approaches for updating approximations of fuzzy ISTU are proposed while some objects adding into or deleting from the two universes, and the corresponding incremental algorithms are designed. Experimental evaluations on real datasets as well as artificial datasets show the effectiveness of the proposed incremental updating method compared with the non-incremental method.

1. Introduction

Along with the information coming from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeping increasing, the era of Big Data has arrived [1,2]. Volume, velocity and volatility, as the three principal challenges of big data mining, have been proposed in [3]. The continuously generated and dynamic changed data needs to be analyzed and processed in a limited time. For example, an unprecedented rate of data creation by sensors or smartphones, even conventional retailers, is driving a growing need for real-time analysis [4].

Rough set theory (RST) proposed by Pawlak is one of powerful information processing tools dealing with uncertainty, incomplete and inconsistent information [5,6]. In order to extend the RST of Pawlak to more widely data processing applications, a large number of generalized rough set models have been proposed. Some of them loosed the equivalence relation to non-equivalence relations, such as tolerance relation [7], non-symmetric similarity relation [8] and dominance relation [9],

☆ This is an extended version of the paper presented at the 2015 IEEE International Conference on Fuzzy System, Istanbul 2015, Turkey.

* Corresponding author.
E-mail addresses: jiehu@swjtu.edu.cn (J. Hu), trli@swjtu.edu.cn (T. Li), cluo@scu.edu.cn (C. Luo), HFujita-799@acm.org (H. Fujita), meterer@163.com (S. Li).

http://dx.doi.org/10.1016/j.ijar.2016.11.002
0888-613X/

while others combined RST with other soft computing theories such as fuzzy theory [10], probability and fuzzy theory [11]. Nowadays, RST has been widely applied in many fields related to data mining and knowledge discovery [12–20].

Data analysis based on RST deals with a data table (formulated as an information system formally) where rows represent objects of the universe, columns represent attributes and entries of the table are attribute values for objects. In classic RST, the objects are all in the same universe, the indiscernibility relation is confined to the objects of the same universe and the target concept and its rough approximations are all located in the same universe. However, there usually exist many situations involving two or more different universes and their relations in real-life applications. For example, in the problem of personalized recommendation [21], there often involve two universes (the set of all users and the set of all possible items that can be recommended, such as movies, books, music, or restaurants) and the relation between the two universes (utility function that measures the usefulness of item to user). Then, for a given user, we want to find the item that maximizes the user's utility. The Information System over Two Universes (ISTU) is a formalization of data table corresponding to two universes and their relation, which is common in real-life applications, e.g., fuzzy multiple criteria decision making [22], disease diagnosis system [23–25], and emergency material demand prediction [26].

To deal with ISTU, many scholars have generalized the RST from single universe to two universes in recent years. Yao et al. firstly proposed a Rough Set Model over Two Universes (RSMTU) [27] under the concept of Shafer's compatibility relation [28], the interpretation of which depends on the available knowledge and the area of applications. For example, in a medical diagnosis system, U may denote a set of symptoms and V a set of diseases. A symptom u is said to be compatible with a disease v if any patient with symptom u may suffer from disease v. Subsequently, many generalizations of Yao's RSMTU [27] have been proposed. Wu et al. presented a generalized fuzzy RSMTU with an arbitrary fuzzy relation, which can be viewed as the extension of Yao [27] to the fuzzy environment [29]. Considering the concept may be fuzzy, Li et al. discussed rough fuzzy approximations on two universes of discourse with a crisp relation [30]. By constructive and axiomatic approaches, Zhang et al. proposed a general model of the interval-valued fuzzy rough sets over two universes by combining the rough set theory with the interval-valued fuzzy set theory [31]. Considering the incompatible relation may exist between two universes, Yan et al. extended RSMTU of Yao [27] to a general relation (incompatible or compatible), which can be used to data processing under the dynamic ISTU where the previous compatible relation may become incompatible relation over time [24]. Sun and Ma presented a fuzzy RSMTU with a fuzzy compatible relation between two different universes [25]. Then Sun et al. further proposed a fuzzy RSMTU in which both the concept and the relation between two universes are fuzzy, and then Sun et al. presented the fuzzy rough set model of emergency material demand prediction over two universes [26]. The work of Sun et al. [26,32] promotes the application of RSMTU in emergency decision-making. Meanwhile, various types of probabilistic rough set models over two universes have been introduced. Shen et al. introduced a variable precision RSMTU [33], which loosed the strict definitions of rough approximations by using the inclusion degree and thus can be used to solve the problems with noise data. Ma et al. studied the properties of the probabilistic RSMTU detailedly and discussed the relationship between probabilistic RST and Bayesian risk decision over two universes [34,35]. Afterwards, Sun et al. introduced a fuzzy rough set on probabilistic approximation space over two universes and applied it to emergency decision-making in unconventional emergency management [36], in which the concept is fuzzy and it can be viewed as the further exploration of work in [26]. Yang et al. proposed Fuzzy Probabilistic Rough Set Model on Two Universes (FPRSMTU) [37] by generalizing the probabilistic rough set model on two universes in [35] with fuzzy relation as well as generalizing the rough set model on two universes in [24] to probabilistic and fuzzy theory. Combining the rough set theory with evidence theory, Xu et al. proposed a knowledge reduction method in a generalized approximation space over two universes, which is the first work about knowledge reduction in the field of RSMTU [38]. Sun et al. presented the generalization of the multigranulation rough set on two universes [39].

In many real-time cases, ISTU is dynamic in the sense that the objects in the two universes are continuously being added, deleted, and the relation values are continually being updated over time. We take a real movie recommender system as example, which involves user profiles, movies profiles and the rating information from users to movies. Over time, some new users may enter into the system and rate some movies, some old users may logout from system, some new movies will be added into the system, some old movies may be deleted from the system due to the copyright expired or outdated, and the rating may also be updated because of the revised opinions from users on movies. Consequently, the variation of dataset will result in the dynamic change of recommend information. Therefore, decision behaviors need to be reinforced continuously under the dynamic data environment. However, from the above survey about RSMTU, we can find that although there have been various attempts to establish different theoretical models of RSMTU to deal with ISTU, there still lacks of methods aiming to speed up the information processing under dynamic ISTU.

Approximations are basic operators in rough set theory [40]. The approximations of a concept (decision classes) partition the universe into three regions, namely, the positive, the negative, and the boundary regions, from which different kinds of rules can be induced, e.g., accept, reject, delay rules or optimal prediction, etc. [41,26,36]. The approximations can also be used to compute the significance of attributes in decision region based attribute reduction (called feature selection in pattern recognition, machine learning, data mining, and so on) or knowledge reduction [38]. By attribute reduction, the acquired decision rules have more generalization ability with the same quality. In rough set based soft clustering analysis, clusters are re-formulated using interval sets, that is, lower approximation and upper approximation [42]. In the clustering attribute selection based on rough cluster analysis, the attribute is selected to split according to the roughness of the attribute or the attribute dependencies, e.g., Min-Min Roughness (MMR) [43], Maximum Dependency of Attribute (MDA) [44], Maximum Attribute Relative (MAR) [45], and Mean of Accuracy of approximation using variable precision of attributes (MA) [46], all