Knowledge-Based Systems 000 (2017) 1–17

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Dynamic variable precision rough set approach for probabilistic set-valued information systems[☆]

Yanyong Huang[a], Tianrui Li[a,*], Chuan Luo[b], Hamido Fujita[c], Shi-jinn Horng[a]

[a] *School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China*
[b] *College of Computer Science, Sichuan University, Chengdu 610065, China*
[c] *Faculty of Software and Information Science, Iwate Prefectural University, 020-0693, Iwate, Japan*

ARTICLE INFO

Article history:
Received 9 October 2016
Revised 22 January 2017
Accepted 1 February 2017
Available online xxx

Keywords:
Set-valued information systems
Incremental learning
Rough sets
Matrix operators

ABSTRACT

Set-valued information systems are important type of data tables in many real applications, where the attribute values are described by sets to characterize uncertain and incomplete information. However, in some real situations, set-values may be depicted by probability distributions, which results in that the traditional tolerance relation based on intersection operation could not reasonably describe the indiscernibility relation of objects. To address this issue, we introduce the concept of probabilistic set-valued information systems (PSvIS), and present the extended variable precision rough set model (VPRS) based on the λ-tolerance relation in terms of Bhattacharyya distance. Considering the features of information systems will evolve over time in a dynamic data environment, it will lead to the change of information granulation and approximation structures. A matrix representation of rough approximation is presented based on two matrix operators and two vector functions in PSvIS. Then incremental mechanisms by the utilization of previously learned approximation results and region relation matrices for updating rough approximations are proposed, and the corresponding algorithms are developed and analyzed. Experimental results show that the proposed algorithms outperform the static algorithms and related incremental algorithms while inserting into or removing from attributes in PSvIS.

1. Introduction

Rough set theory (RST), originated by Pawlak, has become an effective mathematical tool for dealing with uncertain or inconsistent information [1]. Since the equivalence relation in RST is too restrictive to be employed in practical applications, various extended rough set models based on different binary relations have been developed in recent two decades [2–7]. Nowadays, RST has been widely applied in machine learning, data mining, pattern recognition and knowledge discovery [8–12].

Set-valued information systems (SvIS) are important generalized models of single-valued information systems, in which sets are used to characterize the imprecise and missing information. Orlowska et al. studied SvIS based on non-deterministic information and introduced the concept of non-deterministic information system [13]. Yao et al. explicitly introduced SvIS and presented some set-based operations for set-valued data [14]. Guan et al. presented three different relative reducts via the maximal tolerance relation in SvIS [15]. Dai et al. constructed a fuzzy rough set model for set-valued data and presented corresponding method for attribute reduction [16]. Wei et al. presented two different fuzzy rough set models based on the fuzzy similarity class and the fuzzy similar degree, respectively [17]. Considering the queuing problems in the presence of multiple criteria, Qian et al. presented two set-valued ordered information systems in terms of disjunctive and conjunctive semantics and constructed two extended dominance-based rough set models [18]. Zhang et al. combined quantitative rough sets and dominance-based rough sets for dealing with feature selection and approximate reasoning in large-scale set-valued decision systems [19]. Although SvIS have been investigated extensively, no previous study has specifically focused on that set values of SvIS are described by probability distributions. There are many research work on interval-valued data with continuous distribution, which could not be directly applied in the set-valued data with probability distribution due to its distribution is discrete [20–22]. Set-valued data with discrete distribution exist in many real world situations. For example, in a language-test information system, a set value {German, Polish} under the test attribute of spo-

---

[☆] This is an extended version of the paper presented at the 2016 International Joint Conference on Rough Sets (IJCRS 2016), Santiago 2016, Chile.

[*] Corresponding author.

E-mail addresses: yyhswjtu@163.com (Y. Huang), trli@swjtu.edu.cn (T. Li), cluo@scu.edu.cn (C. Luo), HFujita-799@acm.org (H. Fujita), horngsj@yahoo.com.tw (S.-j. Horng).

http://dx.doi.org/10.1016/j.knosys.2017.02.002
0950-7051/© 2017 Published by Elsevier B.V.

Please cite this article as: Y. Huang et al., Dynamic variable precision rough set approach for probabilistic set-valued information systems, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.02.002

ken language indicates that a candidate can speak German and Polish in terms of conjunctive semantic [18]. However, in reality, the candidate may speak German fluently, but a little Polish. In order to describe this phenomenon more accurately, we distinguish the ability of spoken language by characterizing the set value with a discrete probability distribution, such as $\frac{\{\text{German},\text{Polish}\}}{\{0.99,0.01\}}$. The information systems with such probability distribution of data are suggested as Probabilistic Set-valued Information Systems (PSvIS) in our study. Moreover, the traditional tolerance relation based on intersection operation in SvIS could not be applied directly in PSvIS. For instance, two candidates with set values $\frac{\{\text{German},\text{Polish}\}}{\{0.99,0.01\}}$ and $\frac{\{\text{German},\text{Polish}\}}{\{0.01,0.99\}}$ under the attribute "speaking a language" are indiscernible according to the tolerance relation. However, it is not reasonable in terms of the probability distributions, i.e., two people where one speaks well in German and only a little Polish and the other is reverse are in the same tolerance class. Furthermore, considering that the classical rough set model is sensitive to misclassified and noisy data, Ziarko presented a robust model, e.g., variable precision rough set model (VPRS), which allows some degree of partial classification by introducing the majority inclusion degree [23]. VPRS has been widely applied in various fields, such as water demand prediction [23], economic and financial prediction [24,25], medical decision making [26–28], and so on. Motivated by these considerations, this paper presents the λ-tolerance relation based on Bhattacharyya distance and the extended VPRS approach for PSvIS.

Another important issue inspiring this work is that the data will continuously update due to the new data are added and the outdated data are discarded in a dynamic information system. This paper focus on the variation of attributes, which is a common case in a dynamic data environment. For example, patient symptoms in clinical trials [29], texture features in image processing [30], geographic attributes in environmental monitoring [31]. For the emergence of a novel attribute set or the disappearance of an outdated attribute set in an information system, the static learning approach needs to retrain the whole model on the entire updated data, which is time-consuming or even infeasible for real-time decision making. To improve the computational efficiency with the variations of attributes, incremental learning incorporating with RST has been explored by utilizing the accumulated knowledge for analyzing the newly updated data [32–34]. Li et al. presented incremental strategies for computing approximations with respect to the characteristic relation in the incomplete information systems [35]. Luo et al. developed matrix-based incremental approach for updating approximations in set-valued ordered information systems [36]. Yang et al. proposed incremental algorithms for maintaining multigranulation rough approximations under dynamic granulation [37]. Wang et al. investigated incremental mechanisms based on three different information entropies for attribution reduction [38]. Shu et al. introduced incremental approaches of calculating the positive region and tolerance classes for positive region-based attribute reduction [39]. Since calculation of approximations is a necessary step for attribute reduction and knowledge discovery in RST [40,41], this study aims at investigation of the incremental mechanisms to speed up the calculation of rough approximations when evolving features in PSvIS.

Since the matrix form is benefit for intuitive description, simplifying calculation and easy maintainability, it has been widely employed for rough data analysis [42–44]. Zhang et al. presented four cut matrices for constructing rough approximations and developed incremental algorithms to update approximations under attribute generalization in SvIS [45]. Wang et al. introduced Boolean matrices for representing covering approximation operators and developed a corresponding decomposition algorithm [46]. Tan et al. presented matrix-based approaches for calculating set approximations and reducts in a covering decision information system [47]. Luo et al. investigated the matrix representation of probabilistic rough approximations and designed dynamic algorithms for updating approximations when adding or deleting objects [48]. Inspired by these advantages of matrix, this paper presents a matrix-based method for dynamic maintenance approximations in PSvIS. Firstly, we present a matrix characterization of rough approximations based on the relation matrix and two vector functions associated with multiplication and dot divide operators. Considering that attributes will evolve over time in a dynamic PSvIS, which will result in the changes of information granulation and approximation structures. Hence, we develop incremental mechanisms for maintenance of approximations by utilizing previously accumulated approximation results and region relation matrices. Finally, a series of comparative experiments are conducted to demonstrate the effectiveness of proposed incremental methods.

The main contributions include: (i) we proposed an extension of SvIS, viz., PSvIS, which depict set-valued objects with probability distribution. (ii) Bhattacharyya distance is adopted to measure the similarity degree of objects in PSvIS, and VPRS is extended by the introduction of $\lambda-$tolerance relation. (iii) two matrix operators and two vector functions are designed to characterize the matrix representation of rough approximations. (iv) two incremental mechanisms of updating approximations are presented under the variation of attributes in PSvIS. Theoretical and experimental results demonstrated the efficiency of the proposed method compared with the static and existing incremental approaches. The rest of this paper is organized as follows. In Section 2, some preliminary concepts of RST and VPRS are briefly reviewed. Then the concept of PSvIS and the extended VPRS based on λ-tolerance relation is presented. In Section 3, we construct rough approximations based on matrix approaches and discuss the related properties. In Section 4, by introducing the concept of region matrices, incremental mechanisms for maintaining rough approximations are presented with respect to the variation of attributes. Incremental algorithms are presented for updating rough approximations in Section 5. Section 6 reports experimental results, and the conclusions are presented in Section 7.

2. Preliminaries

In this section, we review some basic concepts and notations of classical rough set model and VPRS. More details can be found in [1,23]. Then we introduce the basic concept of PSvIS and present λ-tolerance relation in terms of Bhattacharyya distance for extending VPRS model in PSvIS.

Definition 2.1 [1]**.** Let $S = \{U, AT = C\bigcup D, V, f\}$ be an information system, where U is a non-empty finite set of objects, called the universe; AT is a non-empty finite set of attributes including condition attributes C and decision attributes D; $V = \bigcup_{a\in AT} V_a$ and V_a is a domain of attribute a; f: $U \times AT \rightarrow V$ is an information function such that $f(x, a) \in V_a$ for every $a \in AT$, $x \in U$. $\forall X \subseteq U$ and $B \subseteq A$, the lower and upper approximations of X with respect to the equivalence relation R_B are respectively defined as:

$$\underline{R_B}(X) = \{x | [x]_{R_B} \subseteq X\} \tag{1}$$

$$\overline{R_B}(X) = \{x | [x]_{R_B} \cap X \neq \emptyset\} \tag{2}$$

where $[x]_{R_B} = \{y | (x, y) \in R_B\}$ is the equivalence class determined by the equivalence relation $R_B = \{(x, y) \in U \times U | f(x, b) = f(y, b), \forall b \in B\}$.

According to the lower and upper approximations, the positive, negative and boundary regions of X are easy to obtain as follows.

Please cite this article as: Y. Huang et al., Dynamic variable precision rough set approach for probabilistic set-valued information systems, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.02.002