Knowledge-Based Systems 94 (2016) 13–22

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Rough sets in distributed decision information systems

Jun Hu [a,b,*], Witold Pedrycz [b,c,d], Guoyin Wang [a], Kai Wang [a]

[a] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts & Telecommunications, Chongqing 400065, China
[b] Department of Electrical & Computing Engineering, University of Alberta, Edmonton T6R 2V4, Canada
[c] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[d] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Received 18 April 2015
Revised 29 September 2015
Accepted 28 October 2015
Available online 9 December 2015

Keywords:
Rough sets
Knowledge reduction
Distributed data
Distributed decision information systems

ABSTRACT

In "traditional" rough set methodologies, data are assumed to be stored in a single data repository. However, this assumption is not always true in many real-world problems, where data may be distributed across multiple locations, which is especially pertinent with the development of the Internet. To cope with this phenomenon, in this paper we extend the methodology of rough sets to distributed decision information systems. We first present a definition of rough sets in distributed decision information systems. Then we study the reducibility of distributed decision information systems at two different levels of granularity. The conditions for a decision information table or an attribute in distributed decision information systems to be reducible are presented, and an approach to compute reducts of a distributed decision information system is developed. The experimental results show that the proposed approach can be used to simplify distributed decision information systems, while retain their classification abilities.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

As a useful tool in dealing with uncertainty, incompleteness, and imprecision, rough set theory has received significant attention both in theory and practice since it was introduced in 1982 [1]. To meet different requirements of real-world problems, many extensions have been introduced by substituting indiscernibility relations by tolerance relations [2], similarity relations [3], dominance relations [4], general binary relations [5], coverings [6], and neighborhoods [7]. By combining with other theories, some hybrid models, such as decision-theoretic rough sets [8], game-theoretic rough sets [9], fuzzy rough sets and rough fuzzy sets [10], have been developed. These theoretical models have been successfully applied to different areas, such as mining knowledge from databases, feature selection, decision making, fault diagnosis, medical diagnosis, outlier detection, credit rating, etc [11–22].

Generally, the data to be processed are static, limited and centralized. However, this assumption is no longer true with the development of the Internet where rough set theory meets three new challenges: dynamic data, large volumes of data, and distributed data. To process dynamic data, many incremental approaches have been formulated from different perspectives. Under the variation of the object set, Blaszczynski and Slowinski proposed an algorithm for induction of a satisfactory set of decision rules [23]; Zheng and Wang developed a rough set and rule tree based incremental knowledge acquisition algorithm [24]. Liang et al. studied incremental attribute reduction [25]; Chen et al. and Luo et al. reported research on incremental updating approximations [26,27]; Liu et al. introduced three matrices for inducing knowledge dynamically from incomplete information systems [28]. Under the variation of the attribute set, Li et al. presented an approach to incrementally update approximations of concept [29]; Cheng, Li et al. and Liu et al. extended this work to rough fuzzy sets, dominance based rough sets and probabilistic rough sets respectively [30–32]; Wang et al. and Zeng et al. studied the incremental strategy for reduct computation [33,34]. Under the variation of attribute values, Liu et al. discussed rule induction [35,36]; Chen et al. proposed incremental algorithms for updating the approximations of a concept in complete and incomplete information systems [37,38]; Luo et al. studied the same problem in set-valued decision systems [39].

To cope with large data, many highly efficient methods were proposed in the last decades. Xu et al. developed a quick attribute reduction algorithm, for which the time complexity was substantially reduced [40]. Qian et al. proposed the concept of the positive approximation, and used it to accelerate algorithms of heuristic attribute reduction [41,42]. Lu et al. adopted a boundary region-based significance measure regarded as an evaluation criterion, and devised a fast

* Corresponding author at: Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts & Telecommunications, Chongqing 400065, China. Tel.: +1 7808626174.

E-mail addresses: jun.hu77@gmail.com, hujun@cqupt.edu.cn (J. Hu), wpedrycz@ualberta.ca (W. Pedrycz), wanggy@cqupt.edu.cn (G. Wang), 1575929189@qq.com (K. Wang).

http://dx.doi.org/10.1016/j.knosys.2015.10.025
0950-7051/© 2015 Elsevier B.V. All rights reserved.

feature selection approach [43]. Li et al. derived a fast attribute reduction algorithms for the assignment reduct, the distribution reduct, and the maximum distribution reduct in inconsistent decision tables. Another solution to processing large data is parallel computing. Through dividing a large-scale table into small ones, Liang et al. developed an algorithm to build an approximate reduct by fusing the feature selection results of small tables [44]. Zhang et al. introduced several parallel methods for computing rough set approximation in complete and incomplete information systems [45–47]. Li et al. studied the parallel computing of approximations with dominance-based rough sets approach [48]. Qian et al. proposed a parallel attribute reduction algorithm using MapReduce [49,50]. The basic idea applied in these literatures is dividing the whole data into multiple subsets, then processing them with parallel computer framework such as Hadoop, Phoenix and Twister.

Distributed data are not stored at a single data center, but at multiple sites. A simple approach for the processing of distributed data is to centralize all data together, then the existing methods of rough sets theory could be applied. However, it is not acceptable when the volume of data is very large, because the transmission bandwidth could be relatively limited [51]. In addition, it is sometimes impossible to gather all data in a data center because of privacy and security concern [52]. Many works have generalized classical methods to a distributed framework [53]. Cheung et al. proposed a fast distributed mining algorithm for association rule mining [54]. Kubota et al. parallelized decision tree algorithm to process distributed systems [55]. Kargupta et al. introduced the collective data mining framework for distributed data mining from heterogeneous sites [56]. Basak and Kothari presented a classification algorithm for distributed data [57]. Yang and Wright studied the privacy-preserving computation of bayesian networks on vertically partitioned data [58]. Mangasarian proposed a privacy-preserving support vector machine classifier for distributed data [59]. Kokiopoulou and Frossard studied the problem of classification of multiple observation in the scenario where the observations are collected distributively [60]. Tekin and Schaar developed an online learning algorithm for decentralized big data classification [61]. To our best knowledge, few works have reported the knowledge reduction of distributed data. Qian et al. studied the parallel attribute reduction by dividing the whole data into several small parts in order to accelerate the processing of large data, see [49,50]. However, the proposed methods realize, in essence, centralized computing, and an exchange of large volume of data between computing nodes is required during the processing. This motivates the need for an investigation of knowledge reduction of distributed decision information systems.

Our contributions in this paper are as follows:

(1) The centralized method for processing distributed data is time consuming, and sometimes impossible. To overcome this limitation, a definition of rough sets in distributed decision information systems is presented.
(2) From two different views of information granularity, the reducibility of distributed decision information systems is studied. The conditions for a decision information table or an attribute in distributed decision information systems to be reducible are formulated.
(3) A discernibility function is applied to find the reducts of distributed information systems. Using this method, all the reducts of a distributed decision information system can be established.

The study is structured as follows: In Section 2, some basic concepts of rough sets are reviewed. In Section 3, rough sets in distributed decision information systems are defined, and some important properties of this definition are studied. In Section 4, we discuss how to simplify a distributed decision information system while not reducing its classification abilities. In Section 5, three group of experimental results are presented. In Section 6, we offer some conclusions.

2. Preliminaries

In this section, we briefly review basic notions of rough sets, and the definition of rough sets in decision information tables [1,62].

2.1. Pawlak rough sets

Let U be a finite non-empty set called the universe, and R be an equivalence relation on U. The pair (U, R) is called an approximation space. The equivalence relation R partitions the universe U into disjoint subsets. Such a partition of the universe is denoted by U/R. If two elements x and y in U belong to the same equivalence class, we say that x and y are indistinguishable by R. Otherwise we say that x and y are distinguishable by R.

The equivalence relation and the induced equivalence classes may be regarded as some available knowledge about the objects under consideration.

Definition 1. Given an arbitrary set $X \subseteq U$, it can be characterized by a pair of sets called lower and upper approximations:

$$\underline{R}(X) = \{x \in U | [x]_R \subseteq X\} \tag{1}$$

$$\overline{R}(X) = \{x \in U | [x]_R \cap X \neq \emptyset\} \tag{2}$$

where $[x]_R = \{y \in U | (x, y) \in R\}$ is the equivalence class containing x. In rough set theory, a pair of lower and upper approximations is used to describe a set when the knowledge is incomplete, imprecise, or vague. It also indicates the uncertainty of a set in a given approximation space. When the lower approximation is equal to the upper approximation, the set is precise (definable, certain) in this approximation space; otherwise the set is imprecise (undefinable, uncertain).

2.2. Rough sets in decision information tables

A decision information table (DIT) can be defined as $S = (U, At = C \cup D, \{V_a | a \in At\}, \{f_a | a \in At\})$, where U is a nonempty set of objects, At is a nonempty set of attributes, C is a set of condition attributes, D is a decision attribute, V_a is a nonempty set of values for each attribute $a \in At$, and f_a: $U \to V_a$ is an information function for each attribute $a \in At$.

For a subset of attributes $A \subseteq At$, we define an equivalence relation R_A (or A for short) as follows:

$$R_A = \{(x, y) \in U \times U | \forall_{a \in A}(f_a(x) = f_a(y))\} \tag{3}$$

According to the above definition, two objects in U satisfy R_A if and only if they have the same values on all attributes of A. Consequently, we can form two equivalence relations by C and D respectively.

Definition 2. Let $U/D = \{d_1, d_2, \ldots, d_m\}$ be the partition of the universe U defined by the decision attribute D. Then the positive region and boundary region of D with respect to C in S are as follows:

$$POS_C(D) = \bigcup_{i=1}^{m} \underline{C}(d_i) \tag{4}$$

$$BND_C(D) = U - POS_C(D) \tag{5}$$

where $\underline{C}(d_i) = \{x \in U | [x]_C \subseteq d_i\}$ is the lower approximation of d_i in the approximation space produced by C.

A decision information table is consistent if each equivalence class defined by C leads to a unique decision. In other words, there is $d_i \in U/D$ for any $[x]_C$ such that $[x]_C \subseteq d_i$. In this case, we have $POS_C(D) = U$, and $BND_C(D) = \emptyset$. Otherwise the table is inconsistent.

Let B be a subset of C, an attribute $a \in B$ is dispensable in B if $POS_{B-\{a\}}(D) = POS_B(D)$; otherwise a is indispensable in B. The collection of all the indispensable attributes in C is called the core of S

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات