Information Sciences 257 (2014) 81–100

Contents lists available at ScienceDirect

Information Sciences

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

Composite rough sets for dynamic data mining ☆

Junbo Zhang [a,b], Tianrui Li [a,*], Hongmei Chen [a]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
[b] Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form 5 May 2013
Accepted 2 August 2013
Available online 21 August 2013

Keywords:
Rough sets
Data mining
Incremental learning
Matrix
Composite information systems

ABSTRACT

As a soft computing tool, rough set theory has become a popular mathematical framework for pattern recognition, data mining and knowledge discovery. It can only deal with attributes of a specific type in the information system by using a specific binary relation. However, there may be attributes of multiple different types in information systems in real-life applications. Such information systems are called as composite information systems in this paper. A composite relation is proposed to process attributes of multiple different types simultaneously in composite information systems. Then, an extended rough set model, called as composite rough sets, is presented. We also redefine lower and upper approximations, positive, boundary and negative regions in composite rough sets. Through introducing the concepts of the relation matrix, the decision matrix and the basic matrix, we propose matrix-based methods for computing the approximations, positive, boundary and negative regions in composite information systems, which is crucial for feature selection and knowledge discovery. Moreover, combined with the incremental learning technique, a novel matrix-based method for fast updating approximations is proposed in dynamic composite information systems. Extensive experiments on different data sets from UCI and user-defined data sets show that the proposed incremental method can process large data sets efficiently.

1. Introduction

Rough set theory, proposed by Pawlak [23–26], is a powerful mathematical tool for analyzing various types of data. It can be used in an attribute value representation model to describe the dependencies among attributes, evaluate the significance of attributes and derive decision rules [11,18,27,31].

Since the classical rough set model can only be used to deal with categorical attributes, many extended rough set models have been developed for attributes of multiple different types, such as numerical ones, set-valued ones, interval-valued ones and missing ones [6,8,9,11,13,15,32,33]. For example, Hu et al. generalized classical rough set model with a neighborhood relation to deal with numerical attributes [9,11]. Guan et al. defined a tolerance relation and used the maximal tolerance classes to derive optimal decision rules from set-valued information systems [8]. Qian et al. used a binary dominance relation to process set-valued data in set-valued ordered information systems [28]. Leung et al. defined α-tolerance relations and employed the α-misclassification rate for rule acquisition from interval-valued information systems [15]. In incomplete information systems, the toleration and similarity relations as well as the limited tolerance relation were proposed respectively to

☆ This is an extended version of the paper presented at the 2012 International Conference on Artificial Intelligence and Computational Intelligence (AICI'12), Chengdu, 2012, China.

[*] Corresponding author. Tel.: +86 28 86466426.
E-mail addresses: JunboZhang86@163.com, jbzhang@cs.gsu.edu (J. Zhang), trli@swjtu.edu.cn (T. Li), hmchen@swjtu.edu.cn (H. Chen).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.08.016

deal with missing data in [13,32,34]. Grzymała-Busse combined the toleration and similarity relations and presented characteristic relations for missing data in incomplete information systems [6].

In real-life applications, there are attributes of multiple different types in information systems, e.g., categorical ones, numerical ones, set-valued ones, interval-valued ones and missing ones. Such information systems are called as composite information systems. Most of the rough set based methods fail to deal with more than attributes of two different types. To solve this problem, Abu-Donia proposed multi knowledge based rough approximations using a family of finite number of relations [1]. In our previous work, we introduced the composite rough set model and proposed the basic idea to deal with attributes of multiple different types [38]. Here, we continue with this work and improve the composite rough set model. In addition, a matrix-based method is introduced into our work. It helps compute the approximations, positive, boundary and negative regions intuitively from the composite information system and composite decision table. To adapt to the dynamic changes of the composite information system, we employ an incremental technique and propose a matrix-based incremental method for fast updating the approximations from dynamic composite information systems. Extensive experiments on different data sets from UCI and user-defined data sets show that the proposed matrix-based incremental method can process large data sets efficiently.

The remainder of this paper is organized as follows: Section 2 introduces basic concepts of rough sets and its extended models. Section 3 proposes the composite rough set model to deal with attributes of multiple different types. Section 4 gives matrix-based methods for computing the approximations, positive, boundary and negative regions in composite information systems. Section 5 presents matrix-based incremental methods for updating the approximations in dynamic composite information systems. Section 6 designs and develops the static and incremental algorithm based on matrix for computing and updating the approximations in composite information systems. In the Section 7, the performances of static and incremental methods are evaluated on UCI and user-defined data. Section 8 discusses about some related work on rough sets using matrix-based and incremental techniques. The paper ends with conclusions and further research work in Section 9.

2. Rough set models

In this section, we first briefly review the concepts of rough set model as well as its extensions [7–9,11,13,23,32].

2.1. Classical rough set model

Given a pair $K = (U,R)$, where U is a finite and non-empty set called the universe, and $R \subseteq U \times U$ is an indiscernibility relation on U. The pair $K = (U,R)$ is called as an approximation space. $K = (U,R)$ is characterized by an information system $IS = (U,A,V,f)$, where U is a non-empty finite set of objects; A is a non-empty finite set of attributes; $V = \bigcup_{a\in A} V_a$ and V_a is a domain of attribute a; $f{:}U \times A \to V$ is an information function such that $f(x,a) \in V_a$ for every $x \in U$, $a \in A$. Let $B \subseteq A$, in the classical rough set model, a binary indiscernibility relation R_B is defined as follows:

$$R_B = \{(x,y) \in U \times U | f(x,a) = f(y,a), \quad \forall a \in B\} \tag{1}$$

R_B is an equivalence relation, and $[x]_{R_B}$ denotes an equivalence class of an element $x \in U$ under R_B, where $[x]_{R_B} = \{y \in U | xR_By\}$.

Classical rough set model is based on the equivalence relation. The elements in an equivalence class satisfy reflexive, symmetric and transitive. It does not allow the non-categorical data (e.g., numerical data, set-valued data, and interval-valued data) and requires the information table should be complete. However, non-categorical data appears frequently in real-life applications [6,10,13,28]. Therefore, it is necessary to investigate the situation of non-categorical data in information systems. In what follows, we introduce several representative rough set models [8,11], which will be used in our examples. More rough set models for dealing with non-categorical data are available in the literatures [7,10,13,15,21,28,37].

2.2. Neighborhood rough set model

To deal with numerical data in neighborhood information systems, Hu et al. first employed a neighborhood relation and proposed neighborhood rough sets [9,11].

Definition 1 (*[9,11]*). Let $B \subseteq C$ be a subset of attributes, $x \in U$. The neighborhood $\delta_B(x)$ of x in B is defined as

$$\delta_B(x) = \{y \in U | \Delta_B(x,y) \leqslant \delta\} \tag{2}$$

where Δ is a distance function. $\forall$ x, y, $z \in U$, it satisfies:

(I) $\Delta(x,y) \geqslant 0, \Delta(x,y) = 0$ if and only if $x = y$;
(II) $\Delta(x,y) = \Delta(y,x)$;
(III) $\Delta(x,z) \leqslant \Delta(x,y) + \Delta(y,z)$.