Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Efficient parallel boolean matrix based algorithms for computing composite rough set approximations☆

Q1 Junbo Zhang[a,b], Yun Zhu[b], Yi Pan[b], Tianrui Li[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 6 January 2014
Revised 4 April 2015
Accepted 11 September 2015
Available online xxx

Keywords:
Composite rough sets
Boolean matrix
CUDA
Multi-GPU

ABSTRACT

In information systems, there may exist multiple different types of attributes like categorical attributes, numerical attributes, set-valued attributes, interval-valued attributes, missing attributes, *etc.* Such information systems are called as composite information systems. To process such attributes with rough set theory, composite rough set model and corresponding matrix methods were introduced in our previous study. Calculation of rough set approximations of a concept is the key step for rule acquisition and attribute reduction in rough set based methods. To accelerate the computation process of rough set approximations, this paper first presents the boolean matrix representation of the lower and upper approximations in the composite information system, then designs a parallel method for computing approximations based on matrix, and implements it on Multi-GPU. The experiments on data sets from UCI and user-defined data sets show that the proposed method can accelerate the computation process efficiently. The Multi-GPU implementation achieves up to a speedup of 334.9 over the CPU implementation.

1. Introduction

The rough set (RS) theory is a powerful mathematical tool to describe the dependencies among attributes, evaluate the significance of attributes, and derive decision rules [4,14,22,24,31]. It plays an important role in the fields of data mining and machine learning [8,15,17,19,28,34,42]. Different types of attributes can be processed by different rough set models [6,7,10,11,29,30,35]. Guan et al. introduced the tolerance relation which can be used to process set-valued attributes [7]. Hu et al. generalized classical rough set model by the neighborhood relation to process numerical attributes [10]. Grzymała-Busse integrated the tolerance relation [11] and the similarity relation [29], and proposed the characteristic relation [6] to deal with missing attributes in incomplete information systems. In real applications, there always exist multiple different types of attributes in information systems like categorical attributes, numerical attributes, set-valued attributes, and missing attributes. Such information systems are called as composite information systems. However, most of rough set models fail to deal with information systems with more than two types of attributes. To solve this problem, we proposed the composite rough set (CRS) model, defined a composite relation and used composite classes to drive approximations from composite information systems in our previous work [36]. Table 1 shows a comparison of the above rough set models.

☆ This is an extended version of the paper presented at the 8th International Conference on Rough Sets and Knowledge Technology (RSKT'13), Halifax, Canada, 2013.

* Corresponding author. Tel.: +86 28 86466426.
E-mail addresses: jbzhang@my.swjtu.edu.cn, jbzhang@cs.gsu.edu (J. Zhang), yzhu7@student.gsu.edu (Y. Zhu), pan@cs.gsu.edu (Y. Pan), trli@swjtu.edu.cn, trli30@gmail.com (T. Li).

http://dx.doi.org/10.1016/j.ins.2015.09.022
0020-0255/© 2015 Published by Elsevier Inc.

Table 1
Rough set models.

Model	Relation	Data types			
		C	N	S	M
Classical RS	Equivalence [22]	√	×	×	×
Neighborhood RS	Neighborhood [10]	√	√	×	√
Set-valued RS	Tolerance [7]	√	×	√	×
Characteristic RS	Characteristic [5]	√	×	×	√
Composite RS	Composite [36]	√	√	√	√

C: Categorical, N: Numerical, S: Set-valued, M: Missing

Calculation of rough set approximations of a concept is the key step for rule acquisition and attribute reduction in rough set based methods. The efficient computation of rough set approximations can accelerate the process of knowledge discovery effectively. Parallelization of algorithms is a good way to speed up the computational process [25,26,37,40]. In our previous work, we proposed a parallel algorithm for computing rough set approximations [37], and presented a comparison of rough set based knowledge acquisition on different MapReduce runtime systems [39], where MapReduce is a popular computing model for cloud computing platforms [41]. However, this proposed parallel method can only process categorical attributes. In addition, GPUs may accelerate computationally intensive tasks [3] and have recently been utilized in various domains, including high-performance computing [21]. NVIDIA GPUs [3] power millions of desktops, notebooks, workstations and supercomputers around the world, and NVIDIA CUDA [2] is a General Purpose Computation on GPUs (GPGPUs) framework, which uses a C-like programming language and does not require re-mapping algorithms to graphics concepts. These attributes help users develop correct and efficient GPU programs easily. In this paper, we first present a boolean-based method to deal with composite attributes and compute rough set approximations from composite information systems. It means that the computation of rough set approximations can be processed as boolean matrix operations. We design a parallel boolean matrix based method for computing approximations and parallelize it on GPUs.

The remainder of the paper is organized as follows. Section 2 introduces rough set models as well as CRS model. Section 3 proposes the boolean matrix-based method for calculation of CRS approximations. Section 4 presents the algorithm design for computing CRS approximations and complexity analysis. Section 5 illustrates the efficient parallel method and GPU implementations. Section 6 demonstrates the experimental analysis. The paper ends with conclusions and future work in Section 7.

2. Preliminaries

In this section, we first briefly review the concepts of classical rough set model as well as its extensions [7,10,13,22,23,36].

2.1. Rough set models

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 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 $x \in U$, $a \in A$ where $f(x, a)$ denotes the value of sample x in the attribute a.

Definition 1 (Equivalence Relation [22])**.** Given an information system $IS = (U, A, V, f)$, $B \subseteq A$. The equivalence relation ER_B is defined as follows:

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

Classical rough set model is based on the equivalence relation. The elements in an equivalence class satisfy reflexive, symmetric and transitive. It cannot deal with the non-categorical attributes like numerical attributes, set-valued attributes, etc. However, non-categorical attributes appear frequently in real applications [5,9,11,27]. Therefore, many scholars devote to develop methods to deal with non-categorical attributes in information systems. We here review some representational binary relations for processing different non-categorical attributes [5,9,11,12,18,27,33] in rough set models.

Definition 2 (Neighborhood Relation [10])**.** Given a neighborhood information system $NIS = (U, A, V, f)$, $B \subseteq A$. The neighborhood relation NR_B is defined as

$$NR_B = \{(x, y) \in U \times U | \Delta_B(x, y) \le \delta\} \tag{2}$$

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

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