Information Sciences 373 (2016) 351–368

Contents lists available at ScienceDirect

Information Sciences

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

Parallel attribute reduction in dominance-based neighborhood rough set

Hongmei Chen[a,*], Tianrui Li[a], Yong Cai[a], Chuan Luo[b], Hamido Fujita[c]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] College of Computer Science, Sichuan University, Chengdu 610065, China
[c] School of Intelligent Software Systems, Iwate Prefectural University, 152-52 Sugo,Takizawa-shi 020-0693, Japan

ARTICLE INFO

Article history:
Received 27 January 2016
Revised 26 July 2016
Accepted 5 September 2016
Available online 6 September 2016

Keywords:
Parallel algorithm
Rough sets
Big data
Attribute reduction

ABSTRACT

The amount of data collected from different real-world applications is increasing rapidly. When the volume of data is too large to be loaded to memory, it may be impossible to analyze it using a single computer. Although efforts have been taken to manage big data by using a single computer, the problem may not be solved in an acceptable time frame, making parallel computing an indispensable way to handle big data. In this paper, we investigate approaches to attribute reduction in parallel using dominance-based neighborhood rough sets (DNRS), which take into consideration the partial orders among numerical and categorical attribute values, and can be utilized in a multicriteria decision-making method. We first present some properties of attribute reduction in DNRS, and then investigate principles of parallel attribute reduction in DNRS. Parallelization on different components of attribute reduction are explored in detail. Furthermore, parallel attribute reduction algorithms in DNRS are proposed. Experimental results on UCI data and big data show that the proposed parallel algorithm is both effective and efficient.

1. Introduction

In all kinds of applications in real life, data is constantly being generated, stored and processed as a result of the progress in communication technology and computer science. It is work of data mining to excavate valuable information from this data. However, the volume of the data may be large and the scope of the data may increase, and the uncertainty in data may rise. On the one hand, more valuable information can be mined from large amounts of data; on the other hand, it is challenging to tackle these big data sets. There are two types of important research work. The first one is to process big data in an allowable time [16,27]. The second one is to omit the redundant and unnecessary features within the data that may impede classification ability and increase the consumption of computation and memory storage [1,18]. Running parallel algorithms on a cluster is one common way to handle a large amount of data when it cannot be processed on a single computer. Parallel algorithms have been studied in different domains of machine learning. Parallel methods for computing approximations in rough set theory (RST) were addressed by Zhang et al. in framework of MapReduce [49] and multi-core [21]. Lee et al. applied parallel algorithm in the simulation [20]. Diaz-Morales et al. presented parallel kernel methods [34]. Tian et al. proposed a divide and combine method in NPSVM [46]. Zhu et al. designed a parallel algorithm in

* Corresponding author.
E-mail addresses: hmchen@swjtu.edu.cn (H. Chen), trli@swjtu.edu.cn (T. Li), yongcai@my.swjtu.edu.cn (Y. Cai), cluo@scu.edu.cn (C. Luo), issam@iwate-pu.ac.jp (H. Fujita).

http://dx.doi.org/10.1016/j.ins.2016.09.012
0020-0255/

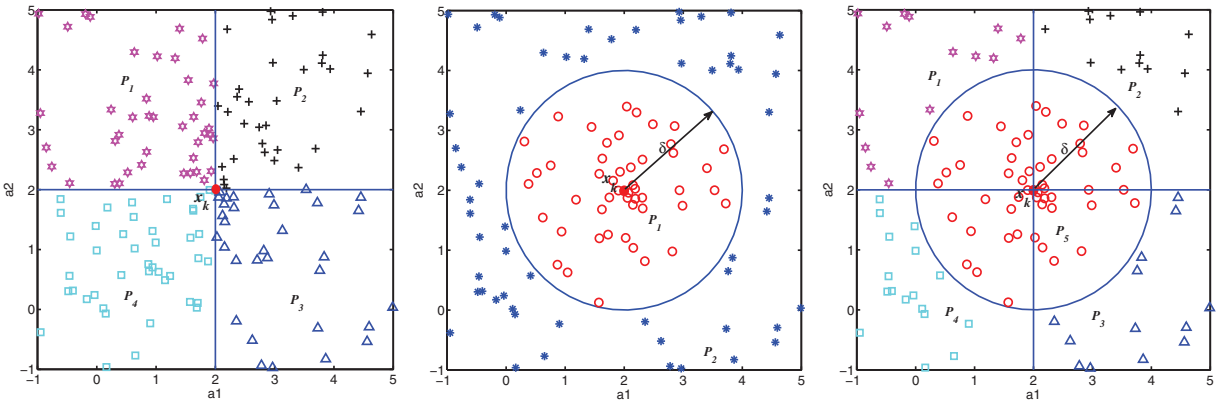

Fig. 1. The relationship among objects in different rough set models.

the multi objective problem [52]. Attribute reduction (also known as feature selection) is one of important applications of RST. Dominance-based neighborhood rough sets (DNRS) is an extension of RST when applying to deal with hybrid data in the context of dominance relation. This paper investigates an approach for attribute reduction in parallel under DNRS.

RST, which emerged in the early 1980s, is one of the most significant mathematical theories employed in processing data with fuzziness and uncertainty [35,36]. The dominance-based rough set approach (DRSA) is an extension of RST that is employed to handle multicriteria decision-making (MCDM) by using a dominance relation to replace the equivalence relation [7,10,12,13,17,28,42]. DRSA has been used in analyzing IT business values [37], formulating airline service strategies [26], developing a decision support system (DSS) for speed management [2], solving multicriteria classification problems in groups [4], assessing rural sustainable development potentialities [3], studying European crisis and welfare patterns using fuzzy integral fusion [29], and investigating robust ordinal regression [14] and so on. DNRS is a kind of DRSA [5] that deals with partially ordered numerical and categorical data in the attribute value domains. The degree of preference between numerical and categorical data is taken into consideration.

Attribute reduction is one of the most important applications of RST, and it has been studied extensively [9,22,24,31,33,40,43,45,47,51,53,59]. There are many novel attribute reduction methods which have considered decision regions [30,50], dynamic properties of the information system [23,54,55], combining with other methods [8,41], dealing with hybrid data [19,56], and giving a final unification of interclass reductions, intraclass reductions and constructs across CRSA and DRSA [44]. Parallel reduction in RST has attracted the interest of scholars. Zhang et al. investigated parallel algorithms for heuristic attribute reduction based on four representative significance measures of attributes [48]. Qian et al. proposed parallel algorithms for calculating equivalence classes and attribute significance in [39], and computing different levels of granularity induced by hierarchical attribute values in [38]. Ma et al. proposed heuristic parallel algorithms for attribute reduction and attribute value reduction in [32]. Then they applied the methods to the fault analysis of large scale data accumulated in power systems. Parallel computation of attribute reduction in DRSA has not been studied. At present, parallel algorithms in RST for attribute reduction are based on the heuristic algorithm which may result in the high computation complexity. In this paper, we investigate the discernibility matrix based parallel algorithm for attribute reduction in DNRS. In a parallel algorithm, the computational tasks are assigned to different computers. The dataset is divided and computed in parallel by different computers to obtain intermediate results. Then they are combined to obtain final results, which need to be the same as the one obtained by a serial algorithm.

The rest of the paper is organized as follows: The motivation for this study is addressed in Section 2. Section 3 introduces the fundamental definitions and concepts in DNRS. Section 4 reviews attribute reduction in DNRS and presents some related properties. In Section 5, the principles of parallel attribute reduction are investigated, and algorithms for attribute reduction are proposed. The experimental results are reported in Section 6, and this paper is concluded with a brief summary in Section 7.

2. Motivation

In RST, elementary classes are formed by objects in different relations to one another. Approximations are defined on the basis of the relation of a concept with these elementary classes. The relations among objects are depicted in Fig. 1 with respect to DRSA, NRS and DNRS. In Fig. 1, $x-$coordinate and $y-$coordinate pertain to attributes $a1$ and $a2$, respectively. Different points denote different objects in the universe. Suppose object x_k is an arbitrary object in the universe. The attribute values of x_k on $a1$ and $a2$ are both 2.

In Fig. 1(a), the objects in the universe are partitioned into four groups: P_1, P_2, P_3, and P_4. In DRSA, objects in different groups have different relations with object x_k. Objects in P_2 are the objects dominating x_k. Objects in P_4 are the objects