Expert Systems with Applications 41 (2014) 655–662

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An effective parallel approach for genetic-fuzzy data mining

Tzung-Pei Hong [a], Yeong-Chyi Lee [b,*], Min-Thai Wu [c]

[a] Dept. of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
[b] Dept. of Information Management, Cheng Shiu University, Kaohsiung, Taiwan
[c] Dept. of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan

ARTICLE INFO

Keywords:
Data mining
Fuzzy set
Genetic algorithm
Parallel processing
Association rule

ABSTRACT

Data mining is most commonly used in attempts to induce association rules from transaction data. In the past, we used the fuzzy and GA concepts to discover both useful fuzzy association rules and suitable membership functions from quantitative values. The evaluation for fitness values was, however, quite time-consuming. Due to dramatic increases in available computing power and concomitant decreases in computing costs over the last decade, learning or mining by applying parallel processing techniques has become a feasible way to overcome the slow-learning problem. In this paper, we thus propose a parallel genetic-fuzzy mining algorithm based on the master–slave architecture to extract both association rules and membership functions from quantitative transactions. The master processor uses a single population as a simple genetic algorithm does, and distributes the tasks of fitness evaluation to slave processors. The evolutionary processes, such as crossover, mutation and production are performed by the master processor. It is very natural and efficient to run the proposed algorithm on the master–slave architecture. The time complexities for both sequential and parallel genetic-fuzzy mining algorithms have also been analyzed, with results showing the good effect of the proposed one. When the number of generations is large, the speed-up can be nearly linear. The experimental results also show this point. Applying the master–slave parallel architecture to speed up the genetic-fuzzy data mining algorithm is thus a feasible way to overcome the low-speed fitness evaluation problem of the original algorithm.

1. Introduction

As information technology (IT) progresses rapidly, its capacity to store and manage data in databases is becoming important. Though IT development facilitates data processing and eases demands on storage media, extraction of available implicit information to aid decision making has become a new and challenging task. Vigorous efforts have thus been devoted to designing efficient mechanisms for mining information and knowledge from large databases. As a result, data mining, first proposed by Agrawal, Imielinksi, and Swami (1993), has become a central field of study in the database and artificial intelligence areas.

Deriving association rules from transaction databases is most commonly seen in data mining (Chen, Han, & Yu, 1996; Famili, Shen, Weber, & Simoudis, 1997; Han & Fu, 1995). It discovers relationships among items such that the presence of certain items in a transaction tends to imply the presence of certain other items. In the past, Agrawal and his co-workers proposed a method (Srikant & Agrawal, 1996) for mining association rules from data sets using quantitative and categorical attributes. Their proposed method first determines the number of partitions for each quantitative attribute, and then maps all possible values of each attribute onto a set of consecutive integers.

Recently, the fuzzy set theory (Zadeh, 1965) has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning (Kandel, 1992). Many fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains (Hong & Lee, 1996; Yuan & Shaw, 1995). Several fuzzy mining algorithms for managing quantitative data have also been proposed (Cai, Fu, Cheng, & Kwong, 1998; Kaya & Alhajj, 2003; Luan, Sun, Zhang, Yu, & Zhang, 2012; Mangalampall & Pudi, 2009; Mohamadlou, Ghodsi, Razmi, & Keramati, 2009; Ouyang & Huang, 2009; Wang, Su, Liu, & Cai, 2012), where the membership functions were assumed to be known in advance. The given membership functions may, however, have a critical influence on the final mining results. Genetic algorithms (GAs) Holland, (1975) have also recently been used in the field of data mining since they are powerful search techniques in solving difficult problems and can provide feasible solutions in a limited amount of time. Hong et al. thus proposed a GA-based fuzzy data-mining method (Hong, Chen, Wu, & Lee, 2006) for extracting both association rules and membership functions from quantitative transactions. In that method, the fitness evaluation is based on the suitability of derived membership

* Corresponding author. Tel.: +886 77310606.
E-mail addresses: tphong@nuk.edu.tw (T.-P. Hong), yeongchyi@csu.edu.tw (Y.-C. Lee), d53040015@student.nsysu.edu.tw (M.-T. Wu).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.090

functions and the number of large itemsets. The evaluation for fitness values is, however, quite time-consuming.

Due to dramatic increases in available computing power and concomitant decreases in computing costs over last decades, learning or mining by applying parallel processing techniques has become a feasible way of overcoming the problem of slow learning (Cordón, Herrera, & Villar, 2001; Herrera, Lozano, & Verdegay, 1997; Wang, Hong, & Tseng, 1998). Several parallel approaches to speed up the process of data-mining were also proposed (Agrawal & Shafer, 1996; Chen, Wang & Chen, 2012; Joshi, Han, Karypis, & Kumar, 2000; Veloso, Meira, & Parthasarathy, 2003). In addition, some parallel methods with genetic algorithms were also suggested (Abramson & Abela, 1992; Araujo, Lopes, & Freitas, 1999). They have been applied to solving timetable scheduling and discovering classification rules.

Among the parallel architectures, the master–slave architecture is particularly easy to implement. It also usually promises substantial gains in performance (Cantu-Paz, 1998). The master processor allocates the tasks to the slave processors and collects the results from them. It can also do its own work if necessary. As mentioned before, the fitness evaluation in genetic-fuzzy data mining is usually very time-consuming. In this paper, we thus extend our previous work (Hong et al., 2006) by using the master–slave parallel architecture to dynamically adapt membership functions and to use them to deal with quantitative transactions in fuzzy data mining. It is very natural and efficient to design a parallel GA-fuzzy mining algorithm based on the master–slave architecture. The master processor uses a single population as a simple genetic algorithm does, and distributes the tasks of fitness evaluation for suitability of membership functions and large itemsets to slave processors. The evolutionary processes, such as crossover, mutation and production are performed by the master processor. We expect that by appropriately allocating the tasks among the different types of processors, the efficiency of the proposed genetic-fuzzy mining algorithm can greatly be raised.

The remainders of this paper are organized as follows. Related works about parallel processing applied to data mining and genetic algorithms are reviewed in Section 2. A parallel genetic-fuzzy mining framework based on the master–slave architecture is described in Section 3. The adopted representation of chromosomes and the genetic operators used in this paper are stated in Section 4. A parallel genetic-fuzzy mining algorithm is proposed in Section 5. A simple example for demonstrating the proposed algorithm is given in Section 6. The time complexity of the proposed algorithm is analyzed in Section 7. The experimental results are shown in Section 8. Finally, conclusion and future work are given in Section 9.

2. Review of related works

Some related works about data mining and genetic algorithms are first reviewed below.

2.1. Data mining

The goal of data mining is to discover important associations among items such that the presence of some items in a transaction will imply the presence of some other items. To achieve this purpose, Agrawal and his co-workers proposed several mining algorithms based on the concept of large itemsets to find association rules in transaction data (Agrawal & Srikant, 1994; Agrawal et al., 1993). They divided the mining process into two phases. In the first phase, candidate itemsets were generated and counted by scanning the transaction data. If the number of an itemset appearing in the transactions was larger than a pre-defined threshold value (called minimum support), the itemset was considered a large itemset. Itemsets containing only one item were processed first. Large itemsets containing only single items were then combined to form candidate itemsets containing two items. This process was repeated until all large itemsets had been found. In the second phase, association rules were induced from the large itemsets found in the first phase. All possible association combinations for each large itemset were formed, and those with calculated confidence values larger than a predefined threshold (called minimum confidence) were output as association rules.

Srikant and Agrawal then proposed a mining method (Srikant & Agrawal, 1996) to handle quantitative transactions by partitioning the possible values of each attribute. Hong et al. proposed a fuzzy mining algorithm to mine fuzzy rules from quantitative data (Hong, Kuo, & Chi, 2001). They transformed each quantitative item into a fuzzy set and used fuzzy operations to find fuzzy rules. Cai et al. proposed weighted mining to reflect different importance to different items (Cai et al., 1998). Each item was attached a numerical weight given by users. Weighted supports and weighted confidences were then defined to determine interesting association rules. Yue et al. then extended their concepts to fuzzy item vectors (Yue, Tsang, Yeung, & Shi, 2000). Fuzzy mining has also been widely adopted in many research fields, such as sequential pattern mining, intrusion detection, biological knowledge extraction, and so on (Chen & Huang, 2008; Lopez, Blanco, Garcia, & Marin, 2007; Romsaiyud1 & Premchaiswadi, 2011; Tajbakhsh, Rahmati, & Mirzaei, 2009; Wang et al., 2012; Watanabe & Fujioka, 2012). In the above fuzzy mining approaches, the membership functions were assumed to be known in advance. Wang et al. used GAs to tune membership functions for intrusion detection systems based on similarity of association rules (Wang & Bridges, 2000). Kaya and Alhajj (2003) proposed a GA-based clustering method to derive a predefined number of membership functions for getting a maximum profit within an interval of user specified minimum support values.

As to parallel data mining, Agrawal and Shafer proposed three parallel mining algorithms based on the Apriori algorithm for speeding up the mining process (Agrawal & Shafer, 1996). The first one was called count distribution, in which the counting task for itemsets was distributed in different processors. The second one was called data distribution, in which itemsets were distributed in different processors and the results were broadcast to each processor for generating globe candidate itemsets in the next phase. The third one was called candidate distribution, which reduced the problem of synchronization between processors by repartitioning transactions according to the itemsets allocated to distinct processors. In addition, two parallel algorithms for mining frequent itemsets were also proposed based on the data-distribution and the candidate-distribution approaches by using the lattice data structure (Veloso et al., 2003).

2.2. Genetic algorithms

Genetic algorithms (GAs) have become increasingly important for researchers in solving difficult problems since they could provide feasible solutions in a limited amount of time (Homaifar, Guan, & Liepins, 1993). They were first proposed by Holland (1975) and have been successfully applied to the fields of optimization, machine learning, neural network, fuzzy logic controllers, and so on (Alcala, Alcala-Fdez, Gacto, & Herrera, 2007; Gautam, Khare & Pardasani, 2010). GAs are developed mainly based on the ideas and the techniques from genetic and evolutionary theory (Grefenstette, 1986). According to the principle of survival of the fittest, they generate the next population by several operations, with each individual in the population representing a possible solution. There are three principal operations in a genetic algorithm.