An improved approach to find membership functions and multiple minimum supports in fuzzy data mining ☆

Chun-Hao Chen [a], Tzung-Pei Hong [b,c,*], Vincent S. Tseng [a]

[a] Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 701, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC
[c] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan, ROC

ARTICLE INFO

ABSTRACT

Fuzzy mining approaches have recently been discussed for deriving fuzzy knowledge. Since items may have their own characteristics, different minimum supports and membership functions may be specified for different items. In the past, we proposed a genetic-fuzzy data-mining algorithm for extracting minimum supports and membership functions for items from quantitative transactions. In that paper, minimum supports and membership functions of all items are encoded in a chromosome such that it may be not easy to converge. In this paper, an enhanced approach is proposed, which processes the items in a divide-and-conquer strategy. The approach is called divide-and-conquer genetic-fuzzy mining algorithm for items with Multiple Minimum Supports (*DGFMMS*), and is designed for finding minimum supports, membership functions, and fuzzy association rules. Possible solutions are evaluated by their requirement satisfaction divided by their suitability of derived membership functions. The proposed GA framework maintains multiple populations, each for one item's minimum support and membership functions. The final best minimum supports and membership functions in all the populations are then gathered together to be used for mining fuzzy association rules. Experimental results also show the effectiveness of the proposed approach.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Data mining is commonly used for inducing association rules from transaction data. An association rule is an expression $X \rightarrow Y$, where X is a set of items and Y is a single item. It means in the set of transactions, if all the items in X exist in a transaction, then Y is also in the transaction with a high probability (Agrawal & Srikant, 1994). Most previous studies focused on binary-valued transaction data. Transaction data in real-world applications, however, usually consist of quantitative values. Designing a sophisticated data-mining algorithm able to deal with various types of data presents a challenge to workers in this research field.

Fuzzy set theory has been used in intelligent systems for a long time because of its simplicity and similarity to human reasoning (Chen, Mikulcic, & Kraft, 2000; Siler & James, 2004; Zhang & Liu, 2006). The theory has been applied in fields such as manufacturing,

engineering, diagnosis, economics, among others (Heng et al., 2006; Ishibuchi & Yamamoto, 2005; Liang, Wu, & Wu, 2002). Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains (Casillas, Cordon, del Jesus, & Herrera, 2005; Hong et al., 2001; Rasmani & Shen, 2004).

Most of the previous approaches set a single minimum support threshold for all the items or itemsets and identify the relationships among binary transactions. In real applications, different items may have different criteria to judge their importance and quantitative data may exist. We can thus divide the fuzzy data mining approaches into two kinds, namely *single-minimum-support fuzzy-mining* (*SSFM*) and *multiple-minimum-support fuzzy-mining* (*MSFM*) problems. Several mining approaches (Chan & Au, 1997; Hong, Kuo, & Chi, 1999, 2001; Kuok, Fu, & Wong, 1998; Yue, Tsang, Yeung, & Shi, 2000) have been proposed for the *SSFM* problem. Chan and Au proposed an F-APACS algorithm to mine fuzzy association rules (Chan & Au, 1997). They first transformed quantitative attribute values into linguistic terms and then used the adjusted difference analysis to find interesting associations among attributes. Kuok et al. (1998) proposed a fuzzy mining approach to handle numerical data in databases and derived fuzzy association rules. At nearly the same time, Hong et al. (1999) proposed a fuzzy mining algorithm to mine fuzzy rules from quantitative transac-

tion data. Basically, these fuzzy mining algorithms first used membership functions to transform each quantitative value into a fuzzy set in linguistic terms and then used a fuzzy mining process to find fuzzy association rules. Yue et al. (2000) then extended the above concept to find fuzzy association rules with weighted items from transaction data. They adopted Kohonen self-organized mapping to derive fuzzy sets for numerical attributes. As to the *MSFM* problem, Lee, Hong, and Lin (2004) proposed a mining algorithm which used multiple minimum supports to mine fuzzy association rules. They assumed that items had different minimum supports and the minimum support for an itemset was set as the maximum of the minimum supports of the items contained in the itemset. Under the constraint, the characteristic of level-by-level processing was kept, such that the original Apriori algorithm could easily be extended to finding large itemsets.

In the above approaches, the membership functions were assumed to be known in advance. Although many approaches for learning membership functions were proposed (Cordón, Herrera, & Villar, 2001; Roubos & Setnes, 2001; Setnes & Roubos, 2000; Wang, Hong, & Tseng,1998, 2000), most of them were usually used for classification or control problems. For fuzzy mining problems, Kaya et al. (2003) proposed a GA-based approach to derive a predefined number of membership functions for getting a maximum profit within a user specified interval of minimum supports. Hong, Chen, Wu, and Lee (2006) also proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions. It maintained a population of sets of membership functions and used the genetic algorithm to automatically derive the resulting one. Its fitness function considered the number of large 1-itemsets and the suitability of membership functions. The suitability measure was used to reduce the occurrence of bad types of membership functions.

The above mentioned approaches, however, were mainly proposed for the *SSFM* problem. As to the *MSFM* problem, we proposed in the past a genetic-fuzzy mining algorithm for items with multiple minimum supports (called the *GFMMS algorithm*) for solving it (Chen, Hong, Tseng, & Lee, 2009). The minimum supports and sets of membership functions of all the items were encoded into a chromosome. Each chromosome was then evaluated by the criteria of requirement satisfaction and suitability of membership functions. Since the chromosome was quite long in this way, lots of processing time was spent to learn global nearly optimal solutions.

Recently, the divide-and-conquer strategy has been used in the evolutionary computation community with a very good effect. Many algorithms based on it have been proposed in different applications (Au, Chan, & Yao, 2003; Darwen & Yao, 1997; Khare, Yao, Sendhoff, Jin, & Wersing, 2005; Yao, 2003). In this paper, we thus propose an enhanced *GFMMS* algorithm, namely divide-and-conquer genetic-fuzzy mining algorithm for items with Multiple Minimum Supports (*DGFMMS*), that can divide-and-conquer the derivation process of the minimum supports and membership functions of different items. The proposed algorithm maintains multiple populations, each for one item's minimum support and membership functions. The fitness of each set of membership functions is evaluated by the requirement satisfaction which is used to reflect the closeness of the derived strength of fuzzy regions of large 1-itemsets for chromosome to its Required Strength of Fuzzy regions (*RSF*) and by the suitability of the derived membership functions. The final best sets of membership functions in all the populations are then gathered together to be used for mining fuzzy association rules. Experiments also were made to show the effective of the proposed approach.

The remaining parts of this paper are organized as follows. The proposed divide-and-conquer genetic-fuzzy mining framework for items with multiple minimum supports is introduced in Section 2. The adjustment process of minimum supports and membership

functions is explained in Section 3. The details of the proposed algorithm for mining multiple minimum supports, membership functions and fuzzy association rules are described in Section 4. An example to illustrate the proposed algorithm is given in Section 5. Experiments to demonstrate the performance of the proposed algorithm are stated in Section 6. Conclusions and future works are given in Section 7.

2. The proposed framework

In this paper, the GA and the fuzzy concepts are used together to discover suitable minimum supports, membership functions and useful fuzzy association rules from quantitative transactions. A GA-based framework with the divide-and-conquer strategy is proposed for searching for minimum supports and sets of membership functions suitable for the mining problems. The final best minimum supports and membership functions for items in all the populations are then gathered together to be used for mining fuzzy association rules. The proposed framework is shown in Fig. 1.

The proposed framework in Fig. 1 is divided into two phases: phase of mining minimum supports and membership functions and phase of mining fuzzy association rules. Assume the number of items is m. In the first phase, the clustering approach is first used for deriving initialization information which is used for getting better initial populations as used in Chen et al. (2009). The initialization information includes an appropriate number of linguistic terms, a range of possible minimum supports, and a set of initial membership functions of each item. It then maintains m populations of minimum supports and membership functions, with each population for an item I_j $(1 \leqslant j \leqslant m)$. Each chromosome in a population represents a possible minimum support and membership functions for that item. The chromosomes in the same population are of the same length. Each chromosome is evaluated by the requirement satisfaction and the suitability of membership functions, which are defined later. The evaluation results are then utilized to choose appropriate chromosomes for mating. The offspring then undergo recursive evolution until a good minimum support and membership functions (the chromosome with the highest fitness value) have been obtained. Next, in the phase of mining fuzzy association rules, the obtained minimum supports and membership functions of all the items are gathered together and used to mine the fuzzy interesting association rules from the given quantitative database (Lee et al., 2004). The details are described in the next section.

3. The proposed divide-and-conquer genetic-fuzzy mining approach

3.1. Chromosome representation

It is important to encode minimum supports and membership functions as string representation for GAs to be applied to our problem. Several possible encoding approaches were described in the past (Cordón et al., 2001; Parodi et al., 1993; Wang et al., 1998, Wang, Hong, & Tseng, 2000). In this paper, each individual consists of two parts, respectively for a minimum support and a set of membership functions. The first part encodes minimum support of a certain item by the real-number schema. Thus, the minimum support of an item I_j is encoded with a real number α_j. The second part handles the set of membership functions for an item. It also adopts the real-number schema. Assume the membership functions are triangular. Three parameters are thus used to represent a membership function. Fig. 2 shows an example for item I_j, where R_{jk} denotes the membership function of the kth linguistic term for I_j and c_{jkp} indicates the pth parameter of fuzzy region