Applied Soft Computing 30 (2015) 767–777

Contents lists available at ScienceDirect

Applied Soft Computing

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

Fuzzy utility mining with upper-bound measure

Guo-Cheng Lan [a], Tzung-Pei Hong [b,c,*], Yi-Hsin Lin [b], Shyue-Liang Wang [d]

[a] Computational Intelligence Technology Center, Industrial Technology Research Institute, Hsinchu 310, Taiwan
[b] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[c] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan
[d] Department of Information Management, National University of Kaohsiung, Kaohsiung 811, Taiwan

ARTICLE INFO

Article history:
Received 3 February 2013
Received in revised form 27 January 2015
Accepted 27 January 2015
Available online 16 February 2015

Keywords:
Data mining
Fuzzy data mining
Fuzzy utility mining
High fuzzy utility itemset
Upper bound

ABSTRACT

Fuzzy utility mining has been an emerging research issue because of its simplicity and comprehensibility. Different from traditional fuzzy data mining, fuzzy utility mining considers not only quantities of items in transactions but also their profits for deriving high fuzzy utility itemsets. In this paper, we introduce a new fuzzy utility measure with the fuzzy minimum operator to evaluate the fuzzy utilities of itemsets. Besides, an effective fuzzy utility upper-bound model based on the proposed measure is designed to provide the downward-closure property in fuzzy sets, thus reducing the search space of finding high fuzzy utility itemsets. A two-phase fuzzy utility mining algorithm, named *TPFU*, is also proposed and described for solving the problem of fuzzy utility mining. At last, the experimental results on both synthetic and real datasets show that the proposed algorithm has good performance.

1. Introduction

Data mining on knowledge discovery is a very important phase for extracting interesting rules and patterns from various data applications. Traditional association-rule mining was first proposed to find the relationship between items from a set of data [1,2], but transactions in retail databases usually included not only bought items but also the bought quantities of the items. Since traditional association-rule mining techniques were not sufficient to be used to handle such quantitative data, Srikant and Agrawal then proposed a new issue named quantitative association rule mining, which partitioned several attribute value ranges for each attribute to find useful quantitative rules. For example, assume a quantitative rule "{age: [20∼29] → Car: [0,1]}", which most customers with age ranging 20–29 usually do not buy cars or buy one car [21]. However, how to decide the suitable intervals for the domain values in each attribute is difficult, and the discovered rules are not easily to be comprehended by decision makers.

Afterward, Chan et al. thus proposed another new issue, namely utility mining, which considered not only quantities of items in transactions but also profits of items in a database, to find itemsets with high utility values in databases [5]. Recently, utility mining has been an emerging issue in data mining due to its practical applications, such as medical data application, mobile data application, behavior data application, stream data application, etc. But, there is a big challenge in utility mining. That is, downward-closure property in association-rule mining cannot be applied to handle the problem of utility mining. To address this, Liu et al. subsequently proposed the two-phase utility mining algorithm (abbreviated as *TP*), which consisted of two main phases to find high utility itemsets in transaction databases [20]. In the first phase, an effective model, called transaction-weighted utilization (abbreviated as *TWU*), was proposed to construct a new downward-closure property [20]. The main principle was that the transaction utility (*tu*) of a transaction was regarded as the upper-bound of any subset in that transaction. The transaction-weighted utility (*twu*) of an itemset was then the summation of transaction utility values of the transactions including the itemset in the database. Next in the second phase, a predefined minimum utility threshold was used to find out itemsets with high utilities, and the itemsets could then be output as auxiliary information. According to definitions in utility mining, however, the output information for a high utility itemset only provided its utility and the items in it for decision makers [20]. Then, we could not acquire more other information from the results discovered by utility mining techniques, such as quantity interval of each item in an itemset [20]. As traditional association-rule mining, in addition, the utility itemset results might be not easily to be comprehended by users.

* Corresponding author at: Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan.
Tel.: +886 75919191.
E-mail addresses: rrfoheiay@gmail.com (G.-C. Lan), tphong@nuk.edu.tw (T.-P. Hong), cccgo123@yahoo.com.tw (Y.-H. Lin).

http://dx.doi.org/10.1016/j.asoc.2015.01.055
1568-4946/

As above mentioned, Wang et al. then proposed a new issue, namely fuzzy utility mining, which combined fuzzy set theory with utility mining to find high fuzzy utility itemsets (*HFU*) from quantitative databases [22]. Moreover, Wang et al.'s also defined a fuzzy utility function to evaluate fuzzy utility of an item in databases. For example, assume there is a quantitative transaction {1*A*, 3*B*}, in which the symbols and the numbers represent the bought items and their sold quantities, respectively. Also, assume the same membership function with three fuzzy regions: *Low*, *Middle*, and *High*, is assumed for the two distinct items, *A* and *B*. With the membership in Fig. 1, the quantities of the two items in the transaction can be converted into the two fuzzy sets f_A = {1/*A.Low*, 0/*A.Middle*, 0/*A.High*} and f_B = {0.6/*B.Low*, 0.4/*B.Middle*, 0/*B.High*}. In addition, assume the profits of the two items, *A* and *B*, are 1 and 8. Take the fuzzy itemset {*A.Low*, *B.Low*} as an example. For the fuzzy term {*B.Low*} of {*A.Low*, *B.Low*}, since the centroid value of the fuzzy region "*Low*" in Fig. 1 is 1, and its profit and membership value are 8 and 0.6, the fuzzy utility of {*B.Low*} can be calculated as 0.6*(1*8), which is 4.8. Similarly, the fuzzy utility of {*A.Low*} can be calculated as 1*(1*1), which is 1. The fuzzy utility of the itemset {*A.Low*, *B.Low*} in the transaction is the summation of the two fuzzy terms in the transaction, which is 5.8 (=1 + 4.8). Finally, the fuzzy utility of an itemset is the summation of fuzzy utilities of the itemset in all transactions including the itemset in a database. As this example notes, however, the common membership degree value for the fuzzy terms in an itemset in fuzzy set theory is not considered in the fuzzy utility function, and the calculated fuzzy utility values for itemsets are also not easily to be explained.

Due to above reasons, this work thus presents a new fuzzy utility function, which considers not only quantities and profits of items but also the minimum operator principle of fuzzy set theory, to evaluate the actual fuzzy utility of an itemset in a set of transactions. However, since downward-closure property in fuzzy itemset mining cannot be kept in fuzzy utility mining with the proposed function, the effective fuzzy utility upper-bound model (abbreviated as *FUUB*) is proposed to avoid information losing. Based on the *FUUB* model, in addition, this work also presents the two-phase fuzzy utility mining approach (abbreviated as *TPFU*). At last, experimental results reveal the proposed *TPFU* approach can have good performance in terms of execution efficiency under various parameter settings.

The rest of parts in this paper is organized as follows. The related works are reviewed in Section 2. The problem to be solved and definitions are described in Section 3. The execution details of the proposed *TPFU* approach are explained in Section 4. An example in Section 4.2 is given to describe how to perform the mining process by using the proposed *TPFU*. The experimental results are then showed in Section 5. Finally, conclusions and future work are discussed in Section 6.

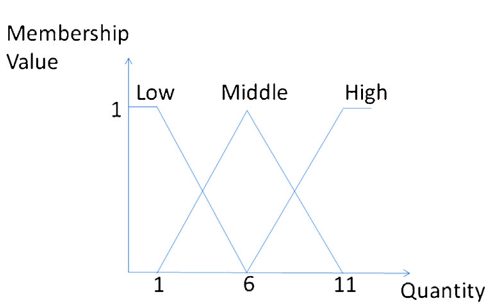

Fig. 1. The membership functions for all the four items in this example.

2. Review of related works

In this section, some related studies on association-rule mining, fuzzy data mining, and fuzzy utility mining are briefly reviewed.

2.1. Association-rule mining

In the field of data mining, discovery of association rules in large databases is one of the important issues due to the consideration of relationship of items [1,2]. For example, assume a product combination "{milks, breads}" has a high frequency in a database. It means the most customers usually buy milks and breads together in the retail. To deal with this problem, a famous algorithm, namely *Apriori*, was first proposed to find frequent itemsets in databases [1]. The *Apriori* algorithm consists of two main phases, finding frequent itemsets and finding association rules. In the first phase, all possible items in a database are regarded as candidate 1-itemsets, and the support of each item is scanned to count the support values of the candidates. Then, the frequent 1-itemsets with supports being larger than or equal to a pre-defined minimum support threshold could be found from the set of candidate 1-itemsets. Next, candidate 2-itemsets with two items are generated from the set of frequent 1-itemsets, and then the process of both the scanning data and counting supports is executed again for the candidate 2-itemsets. The whole mining task can be terminated until no candidate itemsets can be generated in the next pass. After the whole mining process, the final set of frequent itemsets in the database is then found.

2.2. Fuzzy data mining

As mentioned previously, a transaction in retail databases usually also includes quantities bought of items in that transaction in addition to items in the transaction. Then, traditional frequent itemset mining techniques cannot be sufficient to be used to handle such data with quantitative values [8,21]. To deal with such data, Srikant et al. thus proposed a new research issue, named quantitative association rule mining, which partitioned several attribute value ranges for each attribute to find useful quantitative association rules [21]. However, there exist some challenges in quantitative rule mining, such as that it is difficult for how to determine the suitable value ranges for the domain values of each attribute, and it is not easily to be comprehended by users for the rules discovered by the quantitative rule mining techniques.

The fuzzy set theory has widely been used various intelligent systems because of its simplicity and comprehensibility to human reasoning. To reach the goal, Kuok et al. first proposed a new research issue, fuzzy data mining, which applied the concept of fuzzy set theory to the data mining [16]. The main concept behind the issue is that quantitative values in transactions are converted into linguistic regions by fuzzy theory, and a minimum operator in fuzzy theory is applied to obtain the overlap value (minimum value) of membership regions in different items. Different from traditional association rules and quantitative rules, the interesting knowledge with simplicity and comprehensibility for fuzzy data mining could be found from the set of transactions with linguistic regions. To handle this problem, in addition, Hong et al. proposed an effective *Apriori*-based mining algorithm, which adopted a minimum operator in fuzzy theory to count the scalar cardinality value for an itemset in a transaction, to find interesting fuzzy association rules [10]. In addition, Hong et al. proposed an advanced mining approach, which considered the trade-off problem between number of rules and the cost of computation time. The main concept is that fuzzy term with the highest fuzzy count for the items could be kept in the set of fuzzy frequent 1-itemsets, and then a great deal of fuzzy rules could thus be avoided or the execution efficiency could