Expert Systems with Applications 42 (2015) 5754–5778

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An efficient approach for mining association rules from high utility itemsets

Jayakrushna Sahoo[a], Ashok Kumar Das[b], A. Goswami[a,*]

[a] Department of Mathematics, Indian Institute of Technology, Kharagpur 721 302, India
[b] Center for Security, Theory and Algorithmic Research, International Institute of Information Technology, Hyderabad 500 032, India

ARTICLE INFO

Article history:
Available online 11 March 2015

Keywords:
Data mining
High utility itemset mining
Association rule mining
Condensed representations
Non-redundant association rules

ABSTRACT

Traditional association rule mining based on the support–confidence framework provides the objective measure of the rules that are of interest to users. However, it does not reflect the semantic measure among the items. The semantic measure of an itemset is characterized with utility values that are typically associated with transaction items, where a user will be interested to an itemset only if it satisfies a given utility constraint. In this paper, we first define the problem of finding association rules using utility-confidence framework, which is a generalization of the amount-confidence measure. Using this semantic concept of rules, we then propose a compressed representation for association rules having minimal antecedent and maximal consequent. This representation is generated with the help of high utility closed itemsets (HUCI) and their generators. We propose the algorithms to generate the utility based non-redundant association rules and methods for reconstructing all association rules. Furthermore, we describe the algorithms which generate high utility itemsets (HUI) and high utility closed itemsets with their generators. These proposed algorithms are implemented using both synthetic and real datasets. The results demonstrate better efficiency and effectiveness of the proposed HUCI-Miner algorithm compared to other well-known existing algorithms. In addition, the experimental results show better quality in the compressed representation of the entire rule set under the considered framework.

1. Introduction

An expert system is a computer system, which emulates, or acts in all respects, with the decision-making capabilities of a human expert (Mishra, Das, & Mukhopadhyay, 2014). In general, there are three components associated with an expert system, which are (1) knowledge base, (2) inference engine, and (3) user interface (Huynh-Thi-Le, Le, Vo, & Le, 2015). The central of expert systems is the knowledge base as it has the problem solving knowledge of the particular application (Sadik, 2008). Alonso, Martinez, Perez, and Valente (2012) pointed out the cooperation between expert knowledge and data mining discovered knowledge. They also found that the expert knowledge and discovered knowledge are two powerful tools that can be combined together. Data mining techniques are useful in order to discover efficiently the hidden interesting and useful information from large databases, where the implication of interesting and useful information depends on the problem formulation and the application domain. An important data mining task that has received considerable research attention in recent years is the discovery of association rules from the transactional databases (Agrawal & Srikant, 1994; Han, Pei, & Yin, 2000; Park, Chen, & Yu, 1995; Webb, 2006). The traditional association rules mining (ARM) techniques depend on support confidence framework in which all items are given same importance by considering the presence of an item within a transaction, but not the profit of item in that transaction. The goal of such techniques is to extract all the frequent itemsets, where the itemsets having the given minimum support such that the support is the percentage of transactions containing the itemset, which generate all the valid association rules $A \rightarrow B$ from frequent itemset $A \cup B$ whose confidence has at least the user defined confidence such that the confidence is the percentage of transactions containing itemset B among the set of transactions containing A. In other words, given a subset of the items in an itemset, we need to predict the probability of the purchase of the remaining items in a transactional database. In general, from confidence of a rule generated from an itemset, we can know the percentage of number transactions of the items, which is sold together with remaining items of that itemset. However, we may not know the percentage of its profit obtained. Therefore, if we can know the percentage of the items' profit, we

* Corresponding author. Tel.: +91 3222 283650; fax: +91 3222 255303.
E-mail addresses: jayakrushnas@gmail.com (J. Sahoo), iitkgp.akdas@gmail.com, ashok.das@iiit.ac.in (A.K. Das), goswami@maths.iitkgp.ernet.in (A. Goswami).

http://dx.doi.org/10.1016/j.eswa.2015.02.051
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

are in a position to find out a rule, which is more valuable than support and confidence, and as a result, it can allow us to permit with more accurate financial analysis and decisions. Nevertheless, this support–confidence framework does not provide the semantic measure of the rule but only it provides the statistical measure as the relative importance of items is not considered. However, such measure is not an adequate measure to the decision maker as the itemset cannot be measured in terms of stock, cost or profit, called utility. Consider a sales manager who aims to promote itemsets to increase the item selling. The following example is evident that a support–confidence based framework for association rule mining may mislead the manager in the decision making for determining the financial implications of an itemset.

Example 1. Consider the transaction database $\mathcal{D}$ shown in Table 1 that includes nine transactions t_1 through t_9 and eight items A through H. The numbers in the transaction database, which are bracketed, indicate the sales quantity for each item. Table 2 provides the unit profit for each item. The support and utility of the itemset DEF can be calculated using Tables 1 and 2 as 4 and 36, respectively, as the transactions containing DEF are t_2, t_3, t_4 and t_7. Since t_2 includes one D, four Es and five Fs, t_3 includes one D, five Es and one F, t_4 contains one D, two Es and six Fs, and t_7 contains one D, one E and four Fs, a total of four Ds, 12 Es and 16 Fs appear in transactions containing the itemset *DEF*. Using Table 2, the profit of items D, E and F are respectively $2, 1$ and 1. Thus, the profit of the itemset DEF is 36. Using the standard confidence (Agrawal & Srikant, 1994), the confidence of the rule $D \rightarrow EF$ is $4/5 = 80\%$ as only 5 transactions containing in the item D, which are t_2, t_3, t_4, t_7 and t_8. Again, the confidence of the rule $F \rightarrow DE$ is $4/6 = 67\%$. The total utility of items D and F are then 22 and 20, respectively. The contribution of items D and F towards to the total profit of itemset *DEF* are 8 and 16, respectively. Therefore, if we consider the minimum confidence as 70%, the rule $F \rightarrow DE$ is an invalid rule, but the contribution of F from its utility is more than the contribution of item D towards to the total profit of itemset *DEF*. This clearly indicates that the selling of itemset *DEF* contributes a great portion to the total utility of F to 16 out of 20, and hence, the rule, $D \rightarrow EF$, having confidence above the user defined threshold, may mislead to the manager towards the value based decision making.

The support–confidence framework for association rule mining approach explained in Example 1 does not provide any additional knowledge to the manager except the measures that reflects the statistical correlation among items. In addition, it does not reflect their semantic implication towards the mining knowledge. In other words, the support–confidence model may not measure the usefulness of a rule in accordance with a user's objective (for example, profit).

In order to address the above shortcoming of support confidence framework, several researchers have focused on weighted association rule (Cai, Fu, Cheng, & Kwong, 1998; Ramkumar, Ramkumar, & Shalom, 1998; Tao, Murtagh, & Farid, 2003; Wang, Yang, & Yu, 2000). In such framework, the weights of items (the importance of items to the user) are considered and it also varies differently in application domains. However, this framework has two pitfalls. Firstly, these schemes still consider the support of an itemset to measure their importance and secondly, these models do not employ the quantities or prices of items purchased. Considering both quantities of items in a transaction and weights of items Carter, Hamilton, and Cercone (1997) proposed a share-confidence model to discover association rule among numerical attributes which are associated with items in a transaction. Carter et al.'s share-confidence model deals with the amount-share that is a fraction of total weight but not the utility value, such as the net profit, total cost (Geng & Hamilton, 2006). As a result, this model does not accomplish to conventional utility mining (Lin, Hong, & Lu, 2011; Liu, Liao, & Choudhary, 2005; Yao, Hamilton, & Butz, 2004) in which the requirements of decision makers are used to extract the itemsets with high utility, the utility of itemset is no less than the user specified minimum utility threshold, which are composed of weights and purchased quantities. The weight represents the importance of distinct items known as *external utility*, and the purchased quantity in each transaction is known as *internal utility* of the items. The product of external utility with sum total of internal utility of an item is called *utility* of the item. As utility does not satisfy downward closure property (Liu et al., 2005), most of the methods proposed in the literature are applied to find the candidate high utility itemsets first and then to identify actual high utility itemsets by an additional database scan. Some researchers proposed methods to find high utility itemsets without candidate generations (Fournier-Viger, Wu, Zida, & Tseng, 2014b; Liu & Qu, 2012) to avoid additional database scan. However, the discovering process of high utility itemsets takes more execution time and remains a challenge to formulate more effective algorithms. In this paper, we proposed an effective algorithm which is more than two times faster than the state-of-the-art algorithm for discovering high utility itemsets.

Table 1
An example transaction database $\mathcal{D}$.

T_{id}	Transaction
t_1	$A(4), C(1), E(6), F(2)$
t_2	$D(1), E(4), F(5)$
t_3	$B(4), D(1), E(5), F(1)$
t_4	$D(1), E(2), F(6)$
t_5	$A(3), C(1), E(1)$
t_6	$B(1), F(2), H(1)$
t_7	$D(1), E(1), F(4), G(1), H(1)$
t_8	$D(7), E(3)$
t_9	$G(10)$

Table 2
Utility table.

Item	Utility
A	3
B	4
C	5
D	2
E	1
F	1
G	2
H	1

1.1. Motivating examples of applications for utility-confidence framework

The share-confidence (we call it as the utility-confidence) model can be applied in various applications, including online purchases in e-commerce (Shie, Yu, & Tseng, 2013), retail sales (Barber & Hamilton, 2003; Hilderman, Hamiliton, Carter, & Cercone, 1998), cross-selling (Lee, Park, & Moon, 2013) and profit mining (Chen, Zhao, & Yao, 2007; Wang, Zhou, & Han, 2002). Note that the utility-confidence framework is also applicable to the market share rule (Zhang, Padmanabhan, & Tuzhilin, 2004), where the profit is obtained transaction-wise, but not item-wise. In this paper, we provide the examples, which are from the retail transaction datasets. To increase the profit, the manager decides to reward the customers who purchased more than some value and grant a discount on the purchase, or the manager offering a shipping discount may encourage buyer to buy additional items by which shipping is free