Knowledge-Based Systems 96 (2016) 171–187

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Efficient algorithms for mining high-utility itemsets in uncertain databases

Jerry Chun-Wei Lin[a,*], Wensheng Gan[a], Philippe Fournier-Viger[b], Tzung-Pei Hong[c,d], Vincent S. Tseng[e]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen Graduate School, HIT Campus, Shenzhen University, Town Xili, Shenzhen, China
[b] School of Natural Sciences and Humanities, Harbin Institute of Technology, Shenzhen Graduate School, HIT Campus, Shenzhen University, Town Xili, Shenzhen, China
[c] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
[d] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[e] Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 12 June 2015
Revised 26 December 2015
Accepted 27 December 2015
Available online 4 January 2016

Keywords:
High-utility itemset
Uncertain database
Probabilistic-based
Upper-bound
PU-list structure

ABSTRACT

High-utility itemset mining (HUIM) is a useful set of techniques for discovering patterns in transaction databases, which considers both quantity and profit of items. However, most algorithms for mining high-utility itemsets (HUIs) assume that the information stored in databases is precise, i.e., that there is no uncertainty. But in many real-life applications, an item or itemset is not only present or absent in transactions but is also associated with an existence probability. This is especially the case for data collected experimentally or using noisy sensors. In the past, many algorithms were respectively proposed to effectively mine frequent itemsets in uncertain databases. But mining HUIs in an uncertain database has not yet been proposed, although uncertainty is commonly seen in real-world applications. In this paper, a novel framework, named potential high-utility itemset mining (PHUIM) in uncertain databases, is proposed to efficiently discover not only the itemsets with high utilities but also the itemsets with high existence probabilities in an uncertain database based on the tuple uncertainty model. The PHUI-UP algorithm (potential high-utility itemsets upper-bound-based mining algorithm) is first presented to mine potential high-utility itemsets (PHUIs) using a level-wise search. Since PHUI-UP adopts a generate-and-test approach to mine PHUIs, it suffers from the problem of repeatedly scanning the database. To address this issue, a second algorithm named PHUI-List (potential high-utility itemsets PU-list-based mining algorithm) is also proposed. This latter directly mines PHUIs without generating candidates, thanks to a novel probability-utility-list (PU-list) structure, thus greatly improving the scalability of PHUI mining. Substantial experiments were conducted on both real-life and synthetic datasets to assess the performance of the two designed algorithms in terms of runtime, number of patterns, memory consumption, and scalability.

1. Introduction

The main purpose of Knowledge Discovery in Database (KDD) is to discover meaningful and useful information from a collection of data [5–7,12,17]. Frequent itemset mining (FIM) and association rule mining (ARM) [6,7] are some of the most important and common tasks of KDD, since they meet the requirements of numerous domains and applications. ARM typically consists of discovering frequent itemsets (FIs) in a level-wise way using a user-specified minimum support threshold, to then derives association rules (ARs) by also considering a minimum confidence threshold. Many algorithms have been proposed to efficiently mine ARs from precise databases. They can be generally classified into level-wise and pattern-growth approaches. Agrawal et al. first designed the well-known Apriori algorithm to mine ARs in a level-wise way [7]. Han et al. then presented the FP-growth algorithm, which first determines frequent 1-itemsets to build a compressed tree structure, and then discovers the frequent itemsets from the constructed FP-tree without generating candidates [17]. Traditional ARM only considers whether items or itemsets are present or not in transactions. But in real-life applications, several other factors need to be considered such as profit, quantity, cost, and other measures of users' preferences. Considering such factors allows

* Corresponding author. Tel.: +8618824678687.
E-mail addresses: jerrylin@ieee.org (J.C.-W. Lin), wsgan001@gmail.com (W. Gan), philfv@hitsz.edu.cn (P. Fournier-Viger), tphong@nuk.edu.tw (T.-P. Hong), vtseng@cs.nctu.edu.tw (V.S. Tseng).

http://dx.doi.org/10.1016/j.knosys.2015.12.019
0950-7051/

discovering more valuable patterns that let retailers and managers adopt more profitable business strategies than using patterns found by using traditional ARM [39].

High-utility itemset mining (HUIM) [11,27,38,39] considers both the quantity and the profit of items and itemsets to measure how "useful" an item or itemset is. An itemset is called a high-utility itemset (HUI) if its total utility value in a database is no less than a user-specified minimum utility count. The goal of HUIM is to identify items or itemsets in transactions that bring considerable profit to a retailer, although they may not be the most frequent ones. Chan et al. first introduced the concept of HUIM [11]. Yao et al. defined a strict unified framework for mining high-utility itemsets (HUIs) [39]. Liu et al. designed the Two-Phase model [27] to provide an upper bound on the utility of itemsets named the transaction-weighted utility (TWU), which can greatly reduce the number of candidates to be considered for mining HUIs by performing an additional database scan. Several tree-based approaches for efficiently discovering HUIs have been proposed, such as IHUP [9], HUP-growth [24], UP-growth [34] and UP-growth+ [33]. These pattern-growth approaches are faster than previous approaches but may still suffer from long execution time and large memory consumption because they need to generate and maintain a huge number of candidates in memory for mining HUIs. To address the above limitations of traditional HUIM, Liu et al. proposed the HUI-Miner algorithm to directly produce HUIs without performing multiple database scans and without using a candidate generate-and-test approach, by relying on a novel utility-list structure [26]. By enhancing HUI-Miner, the FHM algorithm [14] was shown to outperform previous state-of-the-art HUIM algorithms. The design of more efficient algorithms for mining HUIs from precise databases is still an active research topic.

In real-life applications, uncertainty may be introduced when data is collected from noisy data sources such as RFID, GPS, wireless sensors, and WiFi systems [2,4]. When applied to incomplete or inaccurate data, traditional pattern mining algorithms (e.g., FIM, ARM) cannot be applied to discover the required information. Many algorithms have been developed to discover useful information in uncertain databases. The UApriori algorithm was first proposed to mine frequent itemsets in uncertain databases using a generate-and-test and breadth-first search approach [13]. The uncertain frequent pattern (UFP)-growth algorithm was then designed to mine uncertain frequent itemsets without generating candidates, using a UFP-tree structure [22]. Lin et al. also presented a compressed uncertain frequent pattern (CUFP)-tree [23], and an algorithm to mine uncertain frequent itemsets from the built tree nodes. Development of other algorithm for mining uncertain frequent itemsets in uncertain databases is still in progress [4,25,32,36].

In ref. [16], it had been thoroughly discussed that utility and probability are two different measures used for describing objects (e.g., useful patterns) in real-life applications. The utility is a semantic measure (how the "utility" of a pattern is measured according to a user's a priori knowledge and goals), while probability is an objective measure (the probability of a pattern's existence). Objective interestingness measures (e.g., probability) indicate the support and degree of correlation of a pattern in a given database. However, they do not consider the knowledge of the user who uses the data to discover patterns. Subjective and semantics-based measures (e.g., utility) incorporate the user's background knowledge and goals, and are suitable both for more experienced users and interactive data mining [16]. In HUIM, most algorithms are developed to handle precise databases, which mainly focus on the semantics-based utility measures and do not consider objective probability measures. Thus, traditional HUIM is insufficient to process uncertain data in real-life applications. In real-life applications, an item or itemset is not only present or absent in transactions but also often associated with an existential probability, especially when data is collected from experimental measurements or noisy sensors. Generally speaking, the "utility" of an itemset/pattern represents its importance to the user, i.e., weight, cost, risk, unit profit or value. Most realistic application scenarios are uncertain databases where the "utility" measure can also be considered. For example, in market basket analysis, an uncertain database contains customer transactions, where each transaction record obtained by RFID may be imprecise, i.e., it may contain several items, and be associated with an existence probability [2,4]. For instance, the transaction {*A*:2, *C*:3, *E*:2, 90%} indicates that an event consisting of three items {*A, C, E*} bought with quantities {2, 3, 2} has occurred with an existence probability of 90%. In the field of risk prediction, the risk associated with an event can also be viewed as an occurrence probability. For instance, the event {*B*:1, *D*:3, *E*:1, 75%} indicates that an event consists of three items {*B, D, E*} with occurrence frequencies of {1, 3, 1} and that it occurred with a 75% existence probability. Since the "utility" can be viewed as the user-specified importance, i.e., weight, cost, risk, unit profit or value, HUIM is a useful tool for many real-world applications. And most application scenarios are associated with uncertain databases, e.g., to discover the potential high utility itemsets in market basket analysis; find the potential high risk events in risk prediction system, and mine the potential high risk diseases to make predictions. But numerous discovered HUIs may not be the patterns required by a manager or retailer to take efficient decisions, since traditional HUIM algorithms do not consider existence probabilities. Discovered patterns may be misleading if they have low existential probabilities. In fact, people are always more interested in finding patterns with high existential probabilities than with low existential probabilities. But no algorithm has yet been proposed for mining HUIs in an uncertain database.

In this paper, a novel potential high-utility itemset mining (PHUIM) model is designed to mine potential and meaningful patterns, named highly profitable itemsets with high potential probability (abbreviated as potential high-utility itemsets, PHUIs). Two mining algorithms called PHUI-UP (potential high-utility itemsets upper-bound-based mining algorithm) and PHUI-List (potential high-utility itemsets PU-list-based mining algorithm) are respectively developed based on a level-wise approach and a designed probability-utility-list (PU-list) structure, to mine PHUIs. Major contributions of this paper are summarized as follows:

1. Previous work on HUIM has addressed the issue of mining HUIs efficiently in a precise database. To the best of our knowledge, this is the first paper to address the issue of mining PHUIs in an uncertain database that take both the semantics-based utility measure and the objective probability measure into account.
2. A novel type of patterns named potential high-utility itemset (PHUI) is designed. Moreover, the potential high-utility itemset mining framework (PHUIM) is proposed.
3. Two algorithms, PHUI-UP (potential high-utility itemsets upper-bound-based mining algorithm) and PHUI-List (potential high-utility itemsets PU-list-based mining algorithm), are respectively designed to efficiently mine PHUIs in an uncertain database. They can be used as state-of-the-art algorithms by researchers in future work.
4. PHUI-UP is proposed as a baseline algorithm for mining PHUIs using a level-wise approach in an uncertain database based on an Apriori-like approach and a designed upper-bound model. An improved algorithm named PHUI-List is proposed for discovering PHUIs directly without generating candidates based on a designed vertical data structure, named probability-utility-list (PU-list).
5. Substantial experiments have been conducted on both real-life and synthetic datasets. Results show that the two proposed