Information Sciences 354 (2016) 60–85

Contents lists available at ScienceDirect

Information Sciences

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

Mining interesting patterns from uncertain databases

Akiz Uddin Ahmed[a], Chowdhury Farhan Ahmed[b,*], Md. Samiullah[a], Nahim Adnan[a], Carson Kai-Sang Leung[c]

[a] *Department of Computer Science and Engineering, University of Dhaka, Bangladesh*
[b] *ICube Laboratory, University of Strasbourg, France*
[c] *Department of Computer Science, University of Manitoba, Canada*

ARTICLE INFO

Article history:
Received 28 October 2014
Revised 13 February 2016
Accepted 3 March 2016
Available online 9 March 2016

Keywords:
Data mining
Frequent patterns
Uncertain database
Correlated patterns
Weighted patterns

ABSTRACT

Due to a growing demand for efficient algorithms for mining frequent itemsets from uncertain databases, several approaches have been proposed in recent years, but all of them use support-based constraints to prune the combinatorial search space. Most real life databases contain data whose correctness is uncertain. The support-based constraint alone is not enough, because the frequent itemsets may have weak affinity. Even a very high minimum support is not effective for finding correlated patterns with increased weight or support affinity. There are a few approaches in precise databases that propose new measures to mine correlated patterns, but they are not applicable in uncertain databases because certain and uncertain databases differ both semantically and computationally. In this paper, we propose a new strategy: Weighted Uncertain Interesting Pattern Mining (WUIPM), in which a tree structure (WUIP-tree) and several new measures (e.g., *uConf*, $wUConf$) are suggested to mine correlated patterns from uncertain databases. To our knowledge, ours is the first work specifically to consider weight or importance of an individual item alongside correlation between items of patterns in uncertain databases. Additionally, we propose a new metric, prefix proxy value, *pProxy* for our WUIP-tree that helps improve the mining performance. A comprehensive performance study shows that our strategy (a) generates fewer but valuable patterns and (b) is faster than existing approaches even when affinity measures are not applied.

1. Introduction

For being a base step in knowledge discovery in data mining, many efficient frequent pattern-mining methods were discovered for various fields like itemset mining (e.g., Apriori [4], FP-growth [16]), sequential pattern mining (e.g., GSP [5], SPADE [34], Prefix-Span [27]), data stream mining [7,19,30], high utility pattern mining [6–8], frequent graph mining (e.g., AGM [17], FSG [15], gSpan [32]), correlated pattern mining (e.g., gConfidence [29], graph correlation [18], interest measures [26]), behavior mining [12], and time series (e.g., STNR [28] and flexible periodic pattern [25]). Most of the approaches use a support-based constraint to prune the combinatorial search space and are only applicable to precise databases.

[*] Corresponding author. Tel.: +33 629 271 568.
E-mail addresses: shawpan.du@gmail.com (A.U. Ahmed), cfahmed@unistra.fr, farhan@cse.univdhaka.edu (C.F. Ahmed), samiullah@cse.univdhaka.edu (Md. Samiullah), nahimadnan@gmail.com (N. Adnan), kleung@cs.umanitoba.ca (C.K.-S. Leung).

http://dx.doi.org/10.1016/j.ins.2016.03.007
0020-0255/© 2016 Elsevier Inc. All rights reserved.

However, many real life scenarios exist where an item cannot be labeled as present or absent. These kinds of uncertain databases contain transactions t_i, which represent a set $\{a_1\colon p_1, a_2\colon p_2, \ldots, a_n : p_n\}$ of n items where each item a_j is paired with an existential probability value p_j $(0 < p_j < 1)$, depicting the likelihood of presence for the item a_j.

In uncertain databases, the *expected support count* of an itemset is used instead of the actual *support count* of the precise databases. The *expected support count* of an itemset I, in an uncertain database *UDB* is calculated as,

$$ExpSup(I) = \sum_{i=1}^{|UDB|} \left(\prod_{x \in I} p(x, t_i) \right)$$

where $|UDB|$ is the number of transactions in *UDB* and $p(x, t_i)$ is the existential probability of an item x in a transaction t_i. An itemset is considered *frequent* if its *expected support* satisfies a predefined threshold. From the above formula, it is clear that there is a significant computational difference between frequent itemsets of certain and uncertain databases.

Several algorithms (e.g., U-Apriori [14], UF-growth [20], UFP-growth [2], CUF-growth [21], CUF-growth* [21], PUF-growth [22], U2P-Miner [24], etc.) have been proposed in recent years to mine frequent itemsets from uncertain databases. Some of them also deal with efficient frameworks for ranking top-*k* query processing in uncertain databases [1].

CUF-growth* [21] has outperformed UFP-growth [2] by introducing limiting values and thus has a compact tree structure but generates many false positives. Then, PUF-growth [22] has successfully achieved a better running time by introducing limiting values for prefix branches of the tree and reducing false positives, but this can be further improved as shown in our proposed algorithm. All these have focused only in mining frequent patterns from uncertain databases. However, the frequent patterns are large in number and most of the time contain redundant information, but correlation analysis among the items can eliminate redundant and less significant itemsets from the mined patterns.

Earlier, WIP [33] and Hyperclique Miner [31] have shown interest in mining correlated and weighted correlated patterns, but the basic difference between certain and uncertain databases makes these approaches obsolete for uncertain databases. As a result, there is no suitable approach for mining affinity patterns from uncertain databases that might have been more useful to users.

1.1. Motivating examples

Let us consider a medical diagnosis database where a sample record may look like {fever : 70%, flu : 60%, AIDS : 50%, leukemia : 40%} where the patient is reckoned to have fever with 70% probability and so on. Frequent pattern mining algorithms will extract the most frequent itemsets given a *minimum support* threshold as interesting patterns, but diseases like AIDS and leukemia are expected to occur far less than fever and flu in a common diagnosis database. Therefore, patterns with AIDS and leukemia will be regarded as non-interesting unless the *minimum support* is very low. However, these kinds of diseases are more critical than others, and hence, experts in medical science want to know more about the behaviors of the diseases. This problem can be solved using a very low *minimum support*, but having a low *minimum support* will introduce other problems like mining too many redundant or non-valuable patterns, which violates the primary goal of data mining. As a result, we can conclude that AIDS and leukemia must be given more importance/weight than fever and flu. Now, let us assume that we want to find the correlated diseases; e.g., fever and AIDS are correlated illnesses. In frequent pattern mining, it is very usual to extract {fever, AIDS, leukemia} as a frequent pattern, which indicates the items are equally correlated, but this is not the scenario because {fever, AIDS} or {fever, leukemia} are correlated, but AIDS and leukaemia are not. Therefore, a new way to mine correlation between items is mandatory.

In underground coal mining, monitoring the environment using wireless sensor networks has been a crucial task in such places for keeping a secured working environment for coal miners [23]. Many environmental factors require close observation such as the amount of gas, water and dust. As the system largely depends on sensor data that is usually error prone, it can be expressed with associated probability values; e.g., {gas toxicity : 40%, water rise : 35%, intolerable dust : 15%, collapsed hole : 10%}. From a set of such uncertain data, frequent patterns can be found that indicate an alarming chance of collapse in a certain region. Moreover, correlated patterns may be found that indicate some other incident that follows a particular set of events; e.g., increasing vibrations may lead to temperature rise. Sometimes environmental factors like a change in underground water levels may instantiate a severe collapse in coal mining, which is a rare event. To incorporate such factors, weighted affinity pattern mining can be implemented to ensure safety measures.

Finding similarity patterns of moving objects [13] (e.g., rescue or military operations, searching objects in space using image data), predicting paths for cyclones or missiles, behavior mining in user behavior analytics systems and predicting market trends in a trend analysis system produces data that is uncertain. Hence, we need uncertain pattern mining techniques other than the techniques for robust and accurate data. Say, by assigning ranking values in mining techniques to find patterns on trajectories, we can find the different groups of objects as frequent patterns that move together most of the time. While going from one place to another, it can be determined who is following which object, or which objects lead the path of another object; i.e., finding weight affinity patterns. As another instance, mining behavior from social networking sites that estimates someone to be a student with 80% probability creates a database with tuples like {student : 0.8, photographer : 0.7, cyclist : 0.3}. From this type of database, it may be discovered that {student, photographer} is not only a frequent but also a correlated itemset and pages on books or photography should be suggested to that particular user. Thus, correlated and weighted correlated patterns of such types of data can suggest the most related books and/or materials