UFIMT: An Uncertain Frequent Itemset Mining Toolbox

Yongxin Tong  
HKUST  
yxtong@cse.ust.hk

Lei Chen  
HKUST  
leichen@cse.ust.hk

Philip S. Yu  
University of Illinois at Chicago  
psyu@cs.uic.edu

Motivations

Unlike the frequent itemset in deterministic data has a unique definition, the frequent itemset under uncertain environments has two different definitions. Most existing works focus on one of the definitions and no comprehensive study is conducted to compare the two different definitions. Thus, we hope to

1. Present a novel prototype system, UFIMT, for mining frequent itemsets over uncertain databases.
2. Clarify the relationship of the two definitions of frequent itemsets over uncertain data.
3. Provide uniform baseline implementations for the existing representative algorithms for mining frequent itemsets under uncertain databases.
4. Propose an objective and sufficient experimental evaluation platform.

Basic Concepts

Definition 1 (Expected Support). Given an uncertain database UDB which includes N transactions, and an itemset X, the expected support of X is:

$$\text{esup}(X) = \sum_{i=1}^{N} p_i(X)$$

where $p_i(X)$ is the probability of the itemset X appearing in the i-th transaction.

Definition 2 (Expected-Support-based Frequent Itemset). Given an uncertain database UDB, and a minimum expected support ratio, min_esup, an itemset X is an expected support-based frequent itemset if and only if $\text{esup}(X) > N \times \text{min}_{\text{esup}}$.

Definition 3 (Frequent Probability). Given an uncertain database UDB which includes N transactions, a minimum support ratio min_sup, and an itemset X, X’s frequent probability, denoted as Pr(X), is:

$$\text{Pr}(X) = \text{Pr}(\mu(\text{sup}(X)) > N \times \text{min}_{\text{sup}})$$

Definition 4 (Probabilistic Frequent Itemset). Given an uncertain database UDB, a minimum support ratio min_sup, and a probabilistic frequent threshold $p_{ft}$, an itemset X is a probabilistic frequent itemset if X’s frequent probability is larger than the probabilistic frequent threshold, namely, $\text{Pr}(X) > p_{ft}$.

<table>
<thead>
<tr>
<th>Type</th>
<th>Algorithms</th>
<th>Highlights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Support-based Frequent Algorithms</td>
<td>UApriori</td>
<td>Apriori-based search strategy</td>
</tr>
<tr>
<td></td>
<td>UFP-growth</td>
<td>UFP-tree index structure; Pattern growth search strategy</td>
</tr>
<tr>
<td></td>
<td>UH-Mine</td>
<td>UH-struct index structure; Pattern growth search strategy</td>
</tr>
<tr>
<td>Exact Probabilistic Frequent Algorithms</td>
<td>DP</td>
<td>Dynamic programming-based exact algorithm</td>
</tr>
<tr>
<td></td>
<td>DC</td>
<td>Divide-and-conquer-based exact algorithm</td>
</tr>
<tr>
<td>Approximation Probabilistic Frequent Algorithms</td>
<td>PDUApriori</td>
<td>Poisson-distribution-based approximation algorithm</td>
</tr>
<tr>
<td></td>
<td>NDUApriori</td>
<td>Normal-distribution-based approximation algorithm</td>
</tr>
<tr>
<td></td>
<td>NDUH-Mine</td>
<td>Normal-distribution-based approximation algorithm; UH-struct index structure</td>
</tr>
</tbody>
</table>

Table 1. A Summary of Representative Algorithms in UFIMT

Architecture and Demonstration

The system architecture and a screenshot of UFIMT is shown in Figure 1 and Figure 2. Moreover, two figures of experimental comparisons over UFIMT are illustrated in Figure 3 and 4.