WEIGHTED ASSOCIATION RULE MINING- A REVIEW

M. KALPANA DEVI, M. USHA RANI
1 Research Scholar, 2 Associate Professor, Head, Dept. of Computer Science, SPMVV, Tirupati

ABSTRACT

Traditional association rule mining consider support confident measures to find out frequent item sets, it assumes all items are having equal significance. Where as weighted association rule mining assigns weights to items based on different aspects. Because researchers are more concerned with qualitative aspects of attributes (e.g. significance), as compared to considering only quantitative ones (e.g. number of appearances in a database etc). Because qualitative properties are required in order to fully exploit the attributes present in the dataset. In last few years a few number of weighted associative rule mining algorithms have been proposed, i.e. WAR, WARM, WFIM, WIP, WUARM, FWARM and others. These algorithms employ different rule discovery, rule ranking, rule pruning, rule prediction methods. This paper focuses on surveying and comparing the weighted associative rule mining techniques with regards to the above criteria.

1. INTRODUCTION

Associative rule mining is a branch of a larger area of scientific study known as data mining. Data mining is one of the main phases in knowledge discovery from databases (KDD), which extracts useful patterns from data. Association rule mining [1] is a popular data mining technique because of its wide application in marketing and retail communities as well as other more diverse fields. Classical association rule mining employs support and confidence measures which treats every transaction equally. In contrast, different transactions have different weights in real-life data sets. For example, in retail mining application, frequent item sets identified by the standard association rule mining algorithm may contribute only a small portion of the overall company profit because high profit and luxury items normally do not frequently appear in transactions and thus do not appear in rules with high support count values. To solve this problem weighted association rule mining has been evolved [2]. In the last few years so many algorithms have been successfully proposed for mining association rules with weighted settings.

2. WEIGHTED ASSOCIATION RULE MINING

Weighted association rule mining approach generalizes the traditional association rule mining by assigning weights to the items based on their significance.

Given a set of items I = {i_1, i_2,...,i_n} and a database of transactions D = {t_1, t_2,...,t_m} where t_i = {i_1, i_2,...,i_p}, p ≤ m \text{and} I_i \in D, if X \subseteq I with K = |X| is called a k-itemset or simply an itemset. Let a database D be a multi-set of subsets of I. Each T \in D supports an itemset X \subseteq I if X \subseteq T holds. An association rule is an expression X \rightarrow Y, where X, Y are item sets and X \cap Y = \emptyset holds. Number of transactions T supporting an item X w.r.t D is called support of X, \text{Supp}(X). The strength or confidence (c) for an association rule X \rightarrow Y is the ratio of the number of transactions that contain X \cup Y to the number of transactions that contain X, \text{Conf}(X \rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}. A weight of an item is a non-negative real number that shows the importance of each item. A pair (x, w) is called a weighted item where x \in I is an item and w \in W is the weight associated with x. A transaction is a set of weighted items, each of which may appear in multiple transactions with different weights.

Ramkumar et al. [2] introduced weighted support of association rules based on the costs assigned to both items as well as transactions. An algorithm called WIS was proposed to derive the rules that have a weighted support larger than a given threshold. Cai et al. [3] defined weighted support in a similar way except that they only took item weights into account. The definition broke the downward closure property. As a result, the proposed mining algorithm became more complicated and time consuming. Tao et al. [4] provided another definition to retain the “weighted downward closure property”. In conclusion, the methodology of WARM is to assign weights to items, invent new measures (weighted support) based on these weights, and develop the corresponding mining algorithms.

3. SOLUTION SCHEME

The problem of WARM can be divided into two steps

1. Assign weight for each item, based on its significance. Calculate weighted support of each itemset. The weighted support is the fraction of the weight of the transactions that contains both A and B relative to the weight of all transactions.

2. Using large itemsets, generate association rules that have above the user specified minimum weighted confidence.
Algorithms proposed by various authors for mining association rules with weighted settings are as follows.

3.1 Efficient mining of weighted association rules (WAR)
Wei Wang et al. proposed an efficient mining methodology for Weighted Association Rules (WAR) [12]. The idea is a numerical attribute can be assigned for every item which in turn judges the weight of the item in a particular weight domain. WAR uses a two-fold approach where the frequent itemsets are generated through standard association rule mining algorithms without considering weight. Post-processing is then applied on the frequent itemsets during rule-generation to derive the maximum WARs. It focuses on how weighted association rules can be generated by examining the weighting factors of the items included in generated frequent itemsets. Therefore, we could classify this type of weighted association rule mining methods as a technique of post processing association rules.

3.2 Association rule mining using weighted support and significance framework (WARM)
In the WARM context, an itemset is significant if its weighted support is above a pre-defined minimum weighted support threshold. In fact, the threshold values specified by the user are from the margin of significance of cost point of view. This method may be more meaningful than only specifying relatively arbitrary support threshold. The weighted support of an itemset can be defined as the product of the total weight of the itemset (sum of the weights of its items) and the weight of the fraction of transactions that the itemset occurs in.

**Weighted support:** Weighted support WSP of an itemset, A set of transactions T respects a rule R in the form A → B, where A and B are non-empty sub-itemsets of the item space I and they share no item in common. Its weighted support is the fraction of the weight of the transactions that contains both A and B relative to the weight of all transactions. This can be formulated as:

$$\text{wsp (AB)} = \frac{\sum_{i=1}^{k} \text{weight}(t_i)}{\sum_{k=1}^{\text{WST}}}$$

The goal of the weighted association rule mining is then changed to determining all rules that are above a user specified minimum weighted support threshold holding a minimum user specified confidence. In order to calculate weighted support of an itemset, we need a method to evaluate transaction weight. The transaction weight can be derived from weights of the items presented in the transaction. One may formulate it easily as the average weight of the items presented in the transaction.

$$\text{weight (t_i)} = \frac{\sum_{i=1}^{\text{WStok}} \text{weight (item (i))}}{\text{WS, (t_i)}}$$

The itemset is then validated as significant if its weighted support is above the pre-defined minimum weighted support.

3.3 Fuzzy Weighted Association Rule Mining (FWARM)
WAR uses a post-processing approach by deriving the maximum weighted rules from frequent itemsets. Post WAR doesn’t interfere with the process of generating frequent itemsets but focuses on how weighted AR’s can be generated by examining weighting factors of items included in generated frequent itemsets. Similar techniques were proposed for weighted fuzzy quantitative association rule mining [6, 7, 8]. In [9], a two-fold pre processing approach is used where firstly, quantitative attributes are discretised into different fuzzy linguistic intervals and weights assigned to each linguistic label. A mining algorithm is applied then on the resulting dataset by applying two support measures for normalized and un-normalized cases. The closure property is addressed by using the z-potential frequent subset for each candidate set. An arithmetic mean is used to find the possibility of frequent k+1 itemset, which is not guaranteed to validate the downward closure property. Maybin and Sulaiman [10] proposed fuzzy weighted support and confidence framework (FWARM) algorithm belongs to the breadth first traversal family of ARM algorithms, developed using tree data structures [13] and works in a fashion similar to the Apriori algorithm. It was proposed to mine weighted boolean and quantitative data (by fuzzy means) to address the issue of invalidation of downward closure property and also it shows that using the proposed framework, rules can be generated efficiently with a valid downward closure property without biases made by pre- or post-processing approaches.

3.4 Weighted Frequent Itemset Mining (WFIM)
WFIM [11] is the weighted frequent itemset mining algorithm to use a pattern growth algorithm. The main approach is to push weight constraints into the pattern growth algorithm and provide ways to keep the
downward closure property. WFIM adopts an ascending weight ordered prefix tree. The tree is traversed bottom-up because the previous matching can not maintain the downward closure property. A support of each itemset is usually decreased as the length of an itemset is increased, but the weight has a different characteristic. WFIM adapts the divide and conquer approach for mining weighted frequent itemsets. It divides mining the FP-tree into mining smaller FP trees.

WFIM focused on the downward closure property while maintaining algorithm efficiency. Patterns generated by WFIM have weak support and/or weight affinity patterns. WFIM uses a weight range to adjust the number of patterns. However, WFIM does not provide ways to remove patterns that include items with different support and/or weight levels. It would be better if the weak affinity patterns could be pruned first, resulting in fewer patterns after mining.

3.5 Mining Weighted Interesting Patterns (WIP)

WIP (Weighted Interesting Pattern mining) based on mining weighted frequent patterns. Unil Yun and John J. Leggett [12] define the concept of a weighted hyperclique pattern that uses a new measure, called weight-confidence, to consider weight affinity and prevent the generation of patterns with substantially different weight levels. The weight confidence is used to generate patterns with similar levels of weights and the h-confidence serves to identify strong support affinity patterns. The main approach of WIP is to push weight confidence and/or h-confidence into the weighted frequent pattern mining algorithm based on the pattern growth approach and prune uninteresting patterns. A level of weight and/or support is considered to reflect the overall weight and/or support affinity among items within the pattern. In WIP, a new measure of weight confidence is defined. WIP also divides mining the FP-tree into mining smaller FP trees as in WFIM. In WIP, an ascending weight order method and a bottom-up traversal strategy are used in mining weighted interesting patterns.

3.6 Mining Weighted Association Rules without Pre-assigned Weights

Item set evaluation by support in classical association rule mining [1] is based on counting. In this section, we will introduce a link-based measure called w-support and formulate association rule mining in terms of this new concept. In this paper, author introduced w-support, a new measure of item sets in databases with only binary attributes. The basic idea behind w-support is that a frequent item set may not be as important as it appears, because the weights of transactions are different. These weights are completely derived from the internal structure of the database based on the assumption that good transactions consist of good items. This assumption is exploited by extending Kleinberg’s HITS model and algorithm [13] to bipartite graphs. Therefore, w-support is distinct from weighted support in weighted association rule mining (WARM) [2], [3], where item weights are assigned. Furthermore, a new measurement framework of association rules based on w-support is proposed.

The w-support of an item set X is defined as

\[ W\text{supp}(X) = \frac{\sum_{T \subset X \cap T \subset D} \text{hub}(T)}{\sum_{T \subset D} \text{hub}(T)} \]

where hub (T) is the hub weight of transaction T. An item set is said to be significant if its w-support is larger than a user specified value.

Experimental results show that w-support can be worked out without much overhead, and interesting patterns may be discovered through this new measurement. Compared with Apriori [14], the proposed mining algorithm requires an additional iterative procedure to compute the hub weights of all transactions. The database is scanned exactly once in each iteration. Therefore, the convergence rate of the hub weights is critical to the performance. This method works at the cost of three or four additional database scans over the traditional techniques.

3.7 Weighted Utility Framework for Mining Association Rules (WUARM)

Weighted Utility ARM (WUARM), considers the varied significance and different frequency values of individual items as their weights and utilities. Thus, weighted utility mining focuses on identifying the itemsets with weighted utilities higher than the user specified weighted utility threshold. Weighted Utility association rule mining (WUARM)[15] is the extension of weighted association rule mining in the sense that it considers items weights as their significance in the dataset and also deals with the frequency of occurrences of items in transactions. Thus weighted utility association rule mining is concerned with both the frequency and significance of itemsets. Here weighted utility mining is helpful in identifying the most valuable and high selling items which contribute more to the company’s profits. Weighted Utility of an item set depends upon two factors and weighted utility support can be defined as follows:

**Transactional Utility**: It is the frequency of occurrences or quantity of an item in a transaction.

**Item significance**: It is the value representing significance of an item (value, profit etc) and it holds across the dataset.

**Weighted Utility Support** _wus_ of an itemset X→Y is the fraction of transaction weighted utilities that contain both X and Y relative to the transactional weighted utility of all transactions. It can be formulated as:
In this paper, I have surveyed various WARM approaches. Authors identified the limitation of the traditional Association Rule Mining model, in particular, its incapacity for treating units differently and proposed that weight can be integrated in the mining process to solve this problem. To identify the challenge faced when making improvement towards using weight, in particular the invalidation of downward closure property, a set of new concepts are proposed to adapt weighting in the new setting. Invented new measures (weighted support) based on these weights, and develop the corresponding mining algorithms.

REFERENCES


[12] Unil Yun and John J. Leggett ‘WIP: mining Weighted Interesting Patterns with a strong weight and support affinity’ Texas A&M University College Station, Texas 77843, USA

