An Efficient Algorithm for Mining Association Rules using Confident Frequent Itemsets

Basheer Mohamad Al-Maqaleh
Faculty of Computer Sciences & Information Systems,
Thamar University, Yemen
basheer.almaqaleh.dm@gmail.com

Saleem Khalid Shaab
Thamar Community College, Thamar,
Yemen
saleem_shaab@yahoo.com

***Abstract*— Identifying frequent itemsets is one of the most important issues faced by the knowledge discovery and data mining community. There have been a number of successful algorithms developed for extracting frequent itemsets in very large databases. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational datasets. A problem with such a process is that the solution of interesting patterns has to be performed only on frequent itemsets. Pushing constraints in frequent itemsets mining can help pruning the search space. In this paper, an efficient algorithm is proposed to integrate confidence measure during the process of mining frequent itemsets, which generates confident frequent itemsets. Consequently, the suggested algorithm generates strong association rules from these confident frequent itemsets. This technique has been implemented and the experimental results show the usefulness and effectiveness of the proposed algorithm.**

Keywords- KDD; data mining; confident frequent itemsets; Apriori algorithm.

I. INTRODUCTION

The advances in data collection have generated an urgent need for techniques that can intelligently and automatically analyze and mine knowledge from huge amounts of data. The Knowledge Discovery in Databases (KDD) is the process to exploit the possibilities of extracting the knowledge implicit in the collected data. The field of KDD integrates techniques from artificial intelligence, mathematics and statistics for the discovery of interesting, previously unknown and potentially useful information from large datasets [16]. Data mining is a step of KDD in which patterns or models are extracted from data by using some automated techniques [15].

Association rule mining is receiving increasing attention. It has been mainly developed to identify the relationships strongly associated among itemsets that have high-frequency and strong-correlation [14], [15]. Association rule mining is an unsupervised learning because it extracts rules without any prior class information. The task of association rules mining usually performed into two steps [12]. The first step aims at finding all frequent itemsets that satisfy the minimum support (*minsup*) constraint with the frequent itemset property (any subset of a frequent itemset is frequent; if an itemset is not frequent, none of its supersets can be frequent) for efficiency reasons. The second step involves generating association rules that satisfy the minimum confidence (*minconf*) constraint from the frequent itemsets.

Fundamentally, all association rules meet a *minsup* threshold that defines a frequent itemset. If these association rules further meet a *minconf* threshold, then they are called strong association rules[15]. Several algorithms have been proposed to address association rules mining [12], [13], [14]. Association rule mining can produce a huge amount of patterns that are most of the time not useful to the users. It is, hence, impossible for an expert to assess these patterns. This is the case with the well-known Apriori algorithm [12], [13]. One of the methods used to cope with such an amount of output depends on using constrained association rules mining [3], [6], [11], [17] that helps to reduce the number of uninteresting discovered rules.

Constrained itemsets mining is a hot research theme in data mining. Recent studies show that constraint pushing may substantially improve the performance of frequent itemsets mining [18], [19], [20], [21]. It has been observed that every often user wants to restrict the set of frequent itemsets to be discovered by adding extra constraints like Lift measure [4], and J measure [5] or based on some type of knowledge either before the mining (pre-processing) or after the mining (post-processing) [6], [7], [8], [9], [10]. The pre-processing approaches limit the potential for discovering 'surprising' information in the data. It is clear that additional constraints for itemsets can be verified in a post-processing step, after all itemsets exceeding a given *minsup* threshold have been discovered. The post-processing approaches, on the other hand, sacrifice processing speed for many rules are generated which are then pruned [17]. Nevertheless, such a solution cannot be considered satisfactory since users providing advanced selection criteria may expect that the data mining system will exploit them in the mining process to improve performance [17].

As mentioned above, the Apriori algorithm works into two steps. But in the proposed algorithm, the confidence measure is pushed into the mining of frequent itemsets. During the frequent itemsets generation, the proposed

978-0-7695-4941-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ACCT.2013.16

algorithm computes the confidence and prunes the items that have confidence less than the *minconf* threshold. Further iteration of the algorithm is performed only for the itemsets that have support and confidence higher than a user specified thresholds. These itemsets are called confident frequent itemsets. The method described in this paper pushes the confidence measure within the mining algorithm of frequent itemsets in order to eliminate search space that is uninteresting to the user or to emphasize the search space that is interesting to the user.

Roddick and Rice [2] investigate various ways to set thresholds that define interestingness in rules, such as content dependent and independent ones and links the interest in a pattern to its low probability to occur. The work suggested by Wang, He, Cheung and Chin[1] eliminate the problem of setting *minsup* and find directly association rules with a *minconf* threshold with the drawback that from some databases too many rules might be generated. Pei, Han and Lakshmanan [3] proposed a convertible constraint, which can be pushed deep into frequent pattern mining. Dense Miner [11] applies all of *minconf* and minimum improvement to constraint the search space. An incremental association rules mining algorithm that integrates shocking interestingness criterion during the process of building the model is proposed in [18].

The main challenge in this paper is how to integrate the constraint of confidence measure earlier in the mining procedure(push this constraint deeply into the frequent itemsets generation process) rather than using the simple approach of running a traditional algorithm then using a post-processing pass to filter the generated rules.

This paper is organized as follows. In Section II we briefly recall the basic concepts of association rules. The proposed algorithm for discovering strong association rules is presented in Section III. The experimental results are shown in Section IV to evaluate the performance of the proposed algorithm. And, in Section V the conclusion and future work are described.

II. ASSOCIATION RULE

An association rule is defined as follows:
Let $I = \{i_1, \ldots, i_n\}$ be a set of items, and $T = \{t_1, \ldots, t_m\}$ a set of transactions, where each transaction t_i consists of a subset of items in I. An association rule is then an implication of the form:

$$A \rightarrow B, A \in I, B \in I, A \cap B = \emptyset.$$

The support for an itemset is defined as the ratio of the total number of transactions which support this itemset to the total number of transactions in the database. An itemset A has support s in T if s% of the transactions in T contains A. An itemset A is frequent if its support is higher than the user specified *minsup*.

The confidence of rule $A \rightarrow B$ is the probability that when itemset A occurs in a transaction in T, itemset B also occurs in the same transaction. The rule $A \rightarrow B$ holds in T with confidence c if c% of transactions in T that contain A also contain B. An example of association rules is: 60% of transactions that contain coffee also contain sugar; 5% of all transactions contain both of these items. Here 60% is called the confidence of the rule, and 5% the support of the rule.

The problem of mining association rules is to generate all association rules that consist of frequent itemsets and the confidence greater than the user-specified *minconf.* The discovery of association rules is thus important in understanding the underlying relationships between a large numbers of possible combinations of items [15].

III. THE PROPOSED ALGORITHM

The problem of mining frequent itemsets plays an essential role in mining association rules, but it is not sufficient to mine all frequent itemsets. Instead, it is sufficient to mine the set of confident frequent itemsets.

The proposed algorithm uses confidence measure as a constraint during the model building in order to discover association rules. Instead of using the confidence measure as a post-processing step as in the Apriori algorithm, this measure is pushed into the mining of frequent itemsets to form a constraint in order to discover only confident frequent itemsets. For every stage of frequent itemests generation, frequent sub- itemsests are generated from every frequent itemset at that stage. These frequent sub- itemsests are evaluated using confidence (conf) measure (use "(1)") and prune the frequent sub- itemsests that do not satisfy this measure resulting in a set of confident frequent itemsets in this stage.

$$conf(\text{frequent sub} - \text{itemset}) = \frac{\text{sup.count.fre.itmst}}{\text{sup.count.fre.sub} - \text{itmst}} \quad (1)$$

where sup.count.fre.itmst and sup.count.fre.sub-itmst is the number of transactions containing frequent itemset and frequent sub-itemset respectively.

The proposed algorithm expands the current confident frequent itemsets to the next level frequent itemsets like a normal Apriori algorithm. This approach approves that only confident frequent itemsets are eligible to be candidate during the next iteration of frequent itemsets generation.

The confident frequent itemsets in the proposed algorithm can substantially reduce the number of patterns generated in frequent itemset mining while preserving the complete information regarding the set of frequent itemsets. That is, from the set of confident frequent itemsets, we can easily derive the set of frequent itemsets and their support in further iterations.

Based on this, many of frequent itemsets will relate to unconfident frequent itemsets. If only confident frequent itemsets are extracted, the search space can be greatly reduced. As a result, the proposed algorithm uses the confident frequent itemsets to generate strong association rules. This can improve the quality of the extracted association rules and make them more interesting and easier to understand.

Figure 1 shows the proposed algorithm.

create L_1 = set of frequent(supported) itemsets of cardinality one
set k to 2
while ($L_{k-1} \neq \emptyset$)
{
create C_k from L_{k-1}
prune all the itemsets in C_k that are not frequent, to create L_k
for every frequent itemset in L_k
{
generate all frequent sub-itemets
compute confidence for every frequent sub-itemets
if confidence for all frequent sub-itemets < *minconf* **then**
delete that frequent itemset from L_k
}//for
increase k by 1
}// while
The set of all confident frequent itemsets is $L_1 \cup L_2 \cup \cdots \cup L_k$. Use these confident frequent itemsets to generate strong association rules.

Figure 1. The pseudo-code of the proposed algorithm.

IV. EXPERIMINTAL RESULTS

The proposed approach is implemented, and to evaluate the performance of the proposed algorithm, the algorithm is applied to some real-world datasets from the UCI datasets repository [22]. We evaluated the performance of the proposed algorithm and compared it with the Apriori algorithm which was implemented in a public domain tool called Weka: http://www.cs.waikato.ac.nz/ml/weka/ index.html. WEKA is a collection of machine learning algorithms for data mining tasks. We used the default parameters of the Apriori algorithm to make the comparison fair. The performance of the proposed algorithm on different datasets is demonstrated below:

A. Experiment 1

Mushroom dataset was used for this experiment. This dataset has 8124 examples, and 23 nominal attributes. Table I presents the final strong association rules discovered by the proposed algorithm with the following thresholds: *minsup*=0.82 and *minconf*=1.00.

TABLE I. THE RESULT FOR THE MUSHROOM DATASET

No.	Mind strong association rules	*minsup*	*minconf*
1	gill_spacing=f ∧ ring_number=w → veil_color=p	1.00	1.00
2	gill_spacing=f ∧ gill_size=c → veil_color =p	0.82	1.00
3	gill_color=b ∧ gill_spacing=f → veil_color =p	0.87	1.00
4	gill_color=b ∧ veil_color=p → ring_number =w	0.88	1.00
5	ring_type=o ∧ ring_number=w → veil_color=p	0.97	1.00
6	ring_type=o → gill_spacing=f ∧ veil_color=p	0.97	1.00
7	ring_type=o ∧ gill_color=b → veil_color =p	0.85	1.00

The Apriori algorithm with *minsup*=0.82 and *minconf*=1.00 would generate 11 strong association rules for the Mushroom dataset.

B. Experiment 2

Voting dataset was used for this experiment. This dataset has 435 examples, and 17 nominal attributes. From this dataset, the proposed scheme discovered 4 strong association rules with the following thresholds: *minsup*=0.50 and *minconf*=0.70, which are given in Table II.

TABLE II. THE RESULT FOR THE VOTING DATASET

No.	Mind strong association rules	*minsup*	*minconf*
1	duty_free_exports =n → crime =y	0.51	0.81
2	crime =y → duty_free_exports=n	0.51	0.79
3	duty_free_exports =n → regligious _groups_in_schools=y	0.50	0.79
4	regligious_groups_in_schools =y → duty_free_exports =n	0.50	0.77

The Apriori algorithm with *minsup*=0.50 and *minconf*=0.70 would generate 16 strong association rules from the Voting dataset.

C. Experiment 3

Zoo dataset was used for this experiment. This dataset has 101 examples, and 18 attributes. The attributes were nominal. Table III presents the final strong association rules discovered by the proposed algorithm with the following thresholds: *minsup*=0.65 and *minconf*=0.90.

TABLE III. THE RESULT FOR THE ZOO DATASET

No.	Mind strong association rules	*minsup*	*minconf*
1	breathes=1 ∧ backbone=1→ venomous =0	0.66	0.97
2	breathes=1 ∧ venomous=0 → fins=0	0.70	0.95
3	breathes=1 ∧ fins=0 → venomous=0	0.70	0.93
4	venomous=0 ∧ fins=0 → breathes=1	0.70	0.92
5	venomous=0 ∧ fins=0 → airbone=0	0.66	0.92
6	airbone=0 ∧ feathers=0 → venomous=0	0.60	0.92
7	venomous=0 ∧ airbone=0 → feathers=0	0.66	0.94

The Apriori algorithm with *minsup*=0.65 and *minconf*=0.90 would generate 24 strong association rules for the Zoo dataset.

D. Experiment 4

In this experiment Weather dataset is used. This dataset has 14 examples, and 5 nominal attributes. The proposed algorithm generates 5 strong association rules with the following thresholds: *minsup*=0.20 and *minconf*=0.70 as shown in Table IV.

TABLE IV. THE RESULT FOR THE WEATHER DATASET

No.	Mind strong association rules	*minsup*	*minconf*
1	outlook=overcast → play=yes	0.29	1.00
2	play=no → humidity=high	0.29	0.80
3	humidity=normal → play=yes	0.43	0.86
4	outlook=sunny ∧ play=yes → humidity=high	0.21	1.00
5	tempreture=cool ∧ play=yes → humidity=normal	0.21	1.00

In the same constraints the Apriori algorithm would generate 17 strong association rules form the Weather dataset.

The experimental results show that the proposed algorithm is powerful and outperforms rather than, substantially, the Apriori algorithm. The large number of rules generated by the Apriori algorithm makes manual inspection of the rules very difficult. Hence, automated assistance is needed. In the proposed algorithm the confidence measure is pushed in association rule mining to reduce the huge search space. So, the proposed algorithm would decrease the number of extracted association rules. Figure 2 depicts the comparative performance of the two algorithms.

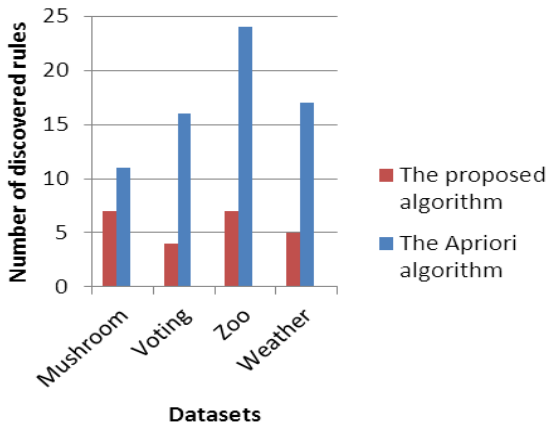

Figure 2. Number of discovered rules by the proposed algorithm and the Apriori algorithm.

V. CONCLUSTION AND FUTURE WORK

The task of data mining is to produce interesting patterns and extract useful knowledge for human users from a database. Frequent itemsets mining is one of the most important areas of data mining. The proposed algorithm shows that integrating confidence measure during the process of mining frequent itemsets may substantially improve the performance of association rules mining by reducing the search space. The integration of confidence measure and frequent pattern growth mining into one unified framework, leads to further improvement of mining efficiency. The experimental results show the effectiveness of the proposed algorithm in reducing the number of discovered rules comparing with the Apriori algorithm.

One of the most important future research directions would be the automated discovery of interesting association rules from large datasets using multi-objective genetic algorithm.

REFERENCES

[1] K. Wang, Y. He, D.W-L Cheung, and F. Chin, "Mining confident rules without support requirement," In Proceedings of the 10th International Conference on Information and Knowledge Management ACM CIKM, pp. 89-96, 2001.

[2] J. Roddick and S. Rice, "What's interesting about cricket? : on thresholds and anticipation in discovered rules," ACM SIGKDD Explorations Newsletter, vol.3,issue 1,pp.1-5, 2001.

[3] J. Pei, J. Han, and L. V. S. Lakshmanan "Mining frequent item sets with convertible constraints," In Proceedings of 17th International Conference on Data Engineering (ICDE'01), IEEE Computer Society, pp. 433-442,2001.

[4] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: generalizing association rules to correlations," In: ACM SIGMOD/ PODS '97 Joint Conference,pp. 265–276,1997.

[5] K. Wang, S. Hock , W. Tay, and B. Liu, "Interestingness-based interval merger for numeric association rules," In Proceedings of the 4th International Conference on Knowledge Discovery & Data Mining, AAAI Press , pp. 121-127,1998.

[6] Tseng and Shin-Mu, "Mining association rules with interestingness constraints in large databases," International Journal of Fuzzy Systems, vol. 3, no. 2, pp. 415-421, 2001.

[7] B. Liu, W. Hsu, S. Chen, "Discovering conforming and unexpected classification rules," IJCAI-97 Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-97), Nagoya, Japan ,1997.

[8] B. Padmanabhan and A. Tuzhilin, "A belief-driven method for discovering unexpected patterns," American Association for Artificial Intelligence, AAAI Press ,pp. 94-100, 1998.

[9] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," KDD '99 Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA ,1999.

[10] M. Zaki, "Generating non-redundant association rules," In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, pp.34-43,2000.

[11] R. Bayardo, R. Agrawal, and D. Gunopulos, "Constraint based rule mining in large, dense databases," In ICDE, pp. 188-197,1999.

[12] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," In the Proceedings of International Conference on Management of Data(ACM-SIGMOD '93), Washington, DC, pp. 207–216,1993.

[13] R. Agrawal. and R. Srikant, "Fast algorithms for mining association rules," In Proceedings of the 20th International Conference on Very Large Databases (VLDB '94), Santiago de Chile, Chile, pp. 478-499,1994.

[14] G. Webb, "Efficient search for association rules," In the Proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 99-107,2000.

[15] J. Han, and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2006.

[16] U. Fayyad and R. Uthurusamy, "Data mining and knowledge discovery in databases," Communications of the ACM, vol. 39, no. 11, pp.24–34,1996.

[17] M. Wojciechowski and M. Zakrzewicz, "Dataset filtering techniques in constraint-based frequent pattern mining," In Proceedings of the

ESF Exploratory Workshop on Pattern Detection and Discovery, Springer-Verlag London, UK, pp. 77-91, 2002.

[18] E. Yafi, A. S. Al-Hegami, A. Alam, and R. Biswas, "Incremental mining of shocking association patterns," International Journal of Computer and Information Engineering , 3(2), pp. 113-117, 2009.

[19] C. K-S. Leung1, B. Hao, F. Jiang, "Constrained frequent itemset mining from uncertain data streams ," In Proceedings of 26th IEEE Interntional Conference on Data Engineering Workshops (ICDEW2010), pp. 120-127.

[20] G. Almodaifer, A. Hafez, and H. Mathkour, "Discovering Medical Association Rules from Medical Datasets," Interntionl Symposium on IT in Medicine and Education (ITME), vol. 2, pp. 43-47, 2011.

[21] A. N. Tran, T. C. Truong, B. H. Le, and H. V. Duong, "Mining Association Rules Restricted on Constraint," In 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp. 1-6, 2012.

[22] UCI Repository of Machine Learning Databases, Department of Information and Computer Science University of California, 1994.[http://www.ics.uci.edu/ ~mlearn/MLRepositry.html.