A Comparative study of CARM and BBT Algorithm for Generation of Association Rules

Rashmi V. Mane
Research Student, Shivaji University, Kolhapur
rvm_tech@unishivaji.ac.in

V.R.Ghorpade
Principal, D.Y.Patil COET, Kolhapur
vijayghorpade@hotmail.com

Abstract

Association rule mining is the technique of data mining to discover user interesting knowledge and forming co-relations among the items. Apriori based, pattern-growth, Fp-tree based are proposed earlier to find the frequent itemsets which are needed for forming associations. Recently two algorithms are proposed for generation of association rules, one of them is CARM and other uses binary technique called BBT algorithm. In this paper we have theoretically analyzed both of these algorithms and comparative study of these algorithms is carried out. Experimental analysis shows that CARM is more time efficient than BBT but it requires large space.

1. Introduction

Data needs a proper mechanism to convert it into knowledge. Data mining techniques are used to discover important and user interesting information from the large dataset. Association rule mining is one of the techniques of data mining which finds a co-relation among the data items. Many algorithms are proposed to find these associations.

Association rule mining needs first to find frequently occurring data items and then finds relation among the data in database. These algorithms basically classified as candidate generation algorithm and frequent pattern growth algorithm. GSP [15], SPADE [13], SPIRIT [12] uses Apriori principle and uses candidate generation approach while FreeSpan, Fp-Growth, PrefixSpan [14] are pattern growth algorithms.

Even though a typical example of association rules mining is market basket analysis to examine history of customer purchase and identifying associations among bought items. Finding relation and association among the data item plays important role in decision support system, telecommunication network, log file analysis, intrusion detection, cross marketing, web log analysis, DNA sequence analysis etc.

Though many algorithms are available for association rule generation, recently two new algorithms are proposed as CARM (Confabulation-inspired association rule mining) [1] and BBT (Binary Based Technique-An Efficient Association Rules Mining Algorithm) [2] are developed for associations.

In this paper we have done comparative study of these two algorithms. The comparison is done as per their working principle and how algorithm finds a rule for a sequential dataset. We have compared the performance of algorithm with respect to its time efficiency and space utilization.

2. Related Work

Lot of work is carried out for generation of association rules. Some of the related work is mentioned here. Maria Luza Antoine, Osmar Zaiane proposed “Mining positive and negative association rules: An approach for confined rules” [16] gives an algorithm which extends support confidence framework with a sliding correlation coefficient threshold. Proposed negative association rule algorithm differs from previously described algorithm as it uses correlation coefficient as measure of interestingness. They have computed correlation coefficient for every pair X, Y of an item i where \( \text{corr}(X, Y) \). They have presented substitution rule mining algorithm. This discovers concrete items having high Chi-Square value and more expected support. Once there itemsets are found, correlation coefficient is computed for each pair. From these pairs negative correlation is generated and desired interesting negative rules are extracted.

Sallam Osman Fageeri, have proposed a binary based Enhanced Semi-Apriori algorithm [2] for discovering frequent itemsets and mining association rules efficiently. In order to minimize the execution time as well as generating large number of candidate sets, this algorithm is implemented using a binary based data structure to store the data and it has minimized the time required for the generation of association rules by using binary map and map inverse concept. This algorithm is efficient than Apriori algorithm but there is overhead of creating binary map and scanning frequency table repeatedly for every new itemset.
Zhixin et al. Recommended an improved classification technique based on predictive association rules [5]. Classification dependent predictive association rules (CPAR) is a type of association classification methods which gives the benefits of associative classification and conventional rule-based classification. CPAR is more effective than conventional rule-based classification because it avoids repeated calculation and multiple rules are generated.

Ken Sun et. al. introduced new focus on Association Rule Mining (ARM) algorithm in 2008 [11]. His algorithm uses w-support, is constraint-based in the sense that all rules must satisfy conditions, such as support and confidence should be greater than or equal to minimum threshold. Main goal of this algorithm is to reduce the number of generated rules.

Yong et al., proposed a mining association rules with new measure criteria [6] in 2005. There are some difficulties in the strong association rules mining depending on support-confidence system. The different measure called Chi-Squared test and cover criteria are introduced for association rules mining, and the Chi-Squared test is used to reduce the amount of rules. The Chi-Squared test and cover of measures are utilized for association rules mining for the purpose of eliminating the itemsets while frequent itemsets or rules are created. Therefore the number of patterns or itemsets reduced and it is easy for user to collect the required association rules.

CARM algorithm requires only one pass through the dataset and it uses a cogency inspired measure for generating association rules while BBT algorithm requires only one scan of the original database, and it uses binary data representation. BBT uses the bit mask for frequent pattern generation. Both CARM and BBT algorithms can be used to find association rules from the different datasets like student dataset, customer transaction dataset, medical dataset etc. CARM uses cogency inspired measure and mines association rules with only one pass through the file [1]. It uses a matrix whose each element represents number of times the link between items is strengthened. Using the link strengths of the items in the matrix we can find frequent itemsets in the datasets. BBT uses binary representation of data and overcomes the drawback of candidate generation [7]. A bit mask is set for every itemset. Logical AND operation is performed between bit mask and binary map of itemset. If it is equal to its bit mask then the frequency of itemset is incremented.

3. Binary Based Technique (BBT) Algorithm

The algorithm scans the original database once, and uses binary based data structure to represent the data as well as vertical database layout concept during the mining process. A bit mask is set for every new itemset. Logical AND operation is performed between bit mask and binary map of itemset. If it is equal to its bit mask then the frequency of itemset is incremented. Using the frequency of the itemsets, we can find frequent itemsets and generate association rules.

It finds 1- frequent itemsets first using traditional method. Then it sets bit mask for frequent itemsets, and it performs ANDing of bit mask of itemsets and their binary map which is stored in the BitSet database. If the result of AND operation is equal to its bit mask then the frequency of the itemset is increased by one in frequency table (FT) [2]. Finally FT is scanned to find the frequent itemsets and to generate association rule using the constraints, support and confidence measure [2, 3]. The algorithm BBT contains two major phases.

Phase I: Discovery of the frequent itemsets.

This phase deals with the first part of association rule mining problem i.e. find all frequent itemsets which are having support greater than or equal to the user-specified minimum support threshold using bit mask and binary map of the itemsets. Frequent itemsets are the itemsets which occurs repeatedly in the database. We find the frequent itemsets in the first phase from the database using BBT algorithm and the output i.e. frequent itemsets are given to the second phase of the algorithm [2].

Phase II: Generation of association rules.

This phase uses the frequent itemsets discovered in Phase I, and starts to generate the association rules. Association rules represent the relationship between the variables or data in the database. Association rules are the rules that have confidence greater than or equal to the user-specified minimum confidence [2].

![Fig. 1. Workflow of BBT algorithm](image)

The algorithm executes in four different steps.

**Step-1**- This procedure scans the original dataset and finds the different items present in the database.
Phase –I includes following steps as-
Phase-I includes following steps as can get 4-frequent itemsets and so on. If the result of the AND operation is same binary 1” in that row for that particular transaction. If that item is not present in the dataset then it places binary “0” for that item in that transaction.

Step-2- The procedure finds the frequency or support of the different itemsets of Length 1 that are present in the database. First of all, it generates the vertical representation of the dataset. It takes the various items present in the dataset and then creates the tabular data which contains the names of the items present in the datasets as the columns. After that it generates the binary dataset. For that it first scans the whole dataset and for every transaction it checks whether an item is present in that transaction or not.

Step-3- This procedure uses the 1-frequent itemsets and generates all possible 2-frequent itemsets. First of all it finds the frequency of all the 1-frequent itemsets from the dataset. Then it compares the frequency of 1-frequent itemsets with the minimum support value given by the user. If the frequency of the 1-frequent itemset is equal to or greater than the minimum support value then it takes that item for further steps of the algorithm, otherwise it discards that item. From the obtained 1-frequent itemsets, it then generates the possible 2-frequent itemsets one by one. Then it finds the binary mask value for the frequent itemsets. After finding the binary mask of the itemset, algorithm performs binary AND operation between the binary mask and binary representation of the each transaction in the BitSet database. If the result of the AND operation is same as the binary mask of that itemset then it increases the frequency of that itemset by one. After performing the AND operation with all the transactions, algorithm calculates the frequency of all the 2-frequent itemsets and compares the frequency of the itemset with the minimum support value. If calculated frequency is satisfying the constraint as equal to or greater than the minimum support value then it considers that frequent itemsets otherwise it discards that itemset.

Step-4- Sets of Size > 2 procedure uses 1-frequent itemsets and 2-frequent itemsets and it will generate the high order frequent itemsets. By appending 1-frequent item to the 2-frequent itemsets, it can obtain 3-frequent itemset. Similarly by appending 1-frequent itemset to the end of 3-frequent itemsets, it can get 4-frequent itemsets and so on.

Phase-II includes following step as

Step-5- The procedure generates Association Rules using the frequent itemsets. For the generation of association rules, it calculates the confidence value of all the frequent itemsets obtained from the given datasets. Then it takes the minimum confidence threshold value from the user and compares the two values. If the calculated confidence is greater than or equal to the minimum confidence threshold value then from that frequent itemset, it generates the association rule.

4. Confabulation-inspired ARM algorithm

CARM approach uses a cogency inspired measure for generating rules. Cogency inspiration can lead to more intuitive rules. Moreover, cogency-related computations only need pair wise item co-occurrences; hence, we can find rules only by one file scan. Rule mining is performed in two main phases:

1. Knowledge acquisition and construction of structure.
2. with the measure of confabulation and cogency generation of rule.

In this algorithm, association rules are generated with only length one item consequent, which means that the consequents of these rules only contain one item. In CARM, items are considered as symbols. There are two modules in this system. Each module contains all items. Let I = {i1, i2, i3, ..., im} be a set of items and T = {t1, t2, t3, ..., tn} be a set of all transactions. L is a m × m matrix that stores knowledge link strength. First, knowledge links with weak strength are established between symbols of two modules. Then, it passes over database. For each transaction ti ∈ T, all 2-subsets of ti are found. These subsets represent all cooccurrences of items belonging to ti. Each 2-subset leads to strengthening the knowledge link between its items. In the simplest form, each element Lij of the matrix L represents the number of times that the link between item i and item j is strengthened. Therefore, principal diagonal stores the total number of times each item appears in database.

During the database scan, for each transaction, we make all 2-subset and then their corresponding value in matrix L is incremented by one.

Fig. 2. Workflow of CARM algorithm

CARM algorithm works in two phases.

Phase I: Knowledge acquisition and structure construction.
In phase I, CARM algorithm constructs knowledge links using matrix for storing knowledge link strengths and discovers Length1-frequent itemsets. Each matrix element represents the number of times that link between items is strengthened, which will be used to find the frequency of the itemsets. Similarly, algorithm finds Length-2-frequent itemsets. This Length-2-frequent itemsets are given as an input to Phase II [1].

**Phase II:** Discovery of further frequent itemsets and association rule generation by confabulation and cogency measure.

In this phase, after finding all frequent length 2-itemsets, the algorithm generates association rules using their support and confidence from the discovered frequent itemsets. Cogency is the probability of the assumed facts being true if the conclusion is true. For computations of cogency pair wise item co-occurrences are needed; hence, it is possible to find rules only by one file scan. It uses Confabulation theory, which is an information processing model of human cognition introduced by Hecht-Nielsen. This theory is based on four fundamental elements as mental object representation, Knowledge links, confabulation and action command origination [1].

5. Experimental work

To compare the performance of BBT and CARM algorithm we have taken a fruit dataset as sequential transaction database. This dataset contains 1220 transactions with eight different items. Following graph shows the time required for two algorithms.

**6. Conclusion**

In this paper we have applied the BBT and CARM algorithm on the datasets fruit market dataset. In Binary-Based technique for mining association rules in large-scale databases, it generates the BitSetDB and perform the AND operation of created mask and the binary representation of each transaction in the BitSetDB. As it needs to scan the whole BitSetDB and perform the logical AND operation every time for finding every frequent itemset, it takes more time to complete the task of Generation of association rules as compared to CARM algorithm. But it takes less memory space as compared to CARM algorithm because it uses the BitSetDB one copy which is updated for every step of the algorithm and it does not create the separate copy every time.

7. References


