ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey of Association Rule Mining Techniques and its Applications

Manoj Ganjir, Jharna Chopra

M.E Student, Dept. of CSE, Shankaracharya Group of Institutions, Bhilai (C.G.), India

Assistant Professor, Dept. of CSE, Shankaracharya Group of Institutions, Bhilai (C.G.), India

ABSTRACT: Data mining has become a major area of study in past few years. Numerous researches had been made within the field of statistics mining. The Association Rule Mining (ARM) is also a big area of study and additionally a statistics mining technique. In this paper a survey is carried out at the distinctive strategies of ARM. In this paper the Apriori algorithm is described and advantages and disadvantages of Apriori algorithm are mentioned. FP- growth algorithm is likewise mentioned and blessings and downsides of FP- boom also are discussed. This paper covers the different strategies available for association rule mining and Simulated Annealing.

KEYWORDS: ARM, Frequent Item set, Pruning, Positive Association Rules, and Negative Association Rules.

I. Introduction

Mobile Information mining is a procedure of separating intriguing learning or examples from extensive databases. There are a few methods that have been utilized to find such sort of information, a large portion of them coming about because of machine learning and measurements. Most of these methodologies concentrate on the revelation of exact information. This learning is valuable to client just when it gives the exact data client needs. Information mining methods are the consequence of a long procedure of examination and item improvement. The primary sorts of techniques performed by information mining systems are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association. Classification separate learning on the premise of already characterized objective trait taking into account different qualities and is regularly spoken to by IF-THEN rules.Dependence Modeling can be seen as a speculation of Classification. The point of Dependence Modeling is to find rules ready to figure the objective characteristic worth, from the estimations of ascertained qualities. The process of partitioning the item set in a set of significant sub-classes (called clusters) is known as Clustering.

A. Association Rule Mining

The notion of mining association rules are as follows. Let S= {S1, S2, S3…………. Sn} be a universe of Items and T= {T1, T2, T3……....Tn} is a set of transactions. Then expression X => Y is an association rule where X and Y are itemsets and X ∩ Y=Φ. Here X and Y are called antecedent and consequent of the rule respectively. This rule holds support and confidence, support is a set of transactions in set T that contain both X and Y and confidence is percentage of transactions in T containing X that also contain Y. An association rule is strong if it satisfies user-set minimum support (minsup) and minimum confidence (minconf) such as support ⩾ minsup and confidence ⩾ minconf. An association rule is frequent if its support is such that support ⩾ minsup. It is assumed to simplify a problem that every item x is purchased once in any given transaction T. generally each transaction T is assigned a number for ex.- Tid. Now it is assumed that X is an itemset then a transaction t is assumed to X iff X ⊆ T. hence it is clear that an association rule is an implication of the form X=>Y where X and Y are subsets of S.

B.Support

Support is a fraction of transactions that contain an itemset. Frequencies of occurring patterns are indicated by support. The probability of a randomly chosen transaction T that contain both itemsets X and Y is known as support. Mathematically it is represented as:-

DOI: 10.15680/IJIRCCE.2015. 0312040

C.Confidence

It measures how often items in Y appear in transactions that contain X. Strength of implication in the rule is denoted by confidence. Confidence is the probability of purchasing an itemset Y in a randomly chosen transaction T depend on the purchasing of an itemset X.

D.Market Basket Analysis

A broadly-used example of association rule mining is market basket analysis. In market basket databases consist of a large no. of records and in each record all items bought by a customer on a single purchase transaction are listed. Managers would be paying attention to know that which groups of items are constantly purchased together. This data is used by them to adjust store layouts (placing items optimally with respect to each other), to cross-sell, to promotions, to catalogue design and to identify customer segments based on buying patterns. For example, suppose a shop database has 200,000 point- of-sale transactions, out of which 4,0000 include both items A and B and 1600 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 1600 transactions (alternatively 0.8% = 1600/200,000) and a confidence of 4% (=1600/4,0000).

The probability of a randomly selected transaction from the database will contain all items in the antecedent and the consequent is known as support, whereas the conditional probability of a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent is known as confidence. Now a day's products are coming with bar codes. A large amount of sales data is produced by the software supporting these barcode based purchasing/ordering system which is typically captured in “baskets”. Commercial organizations are interested in discovering “association rules” that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This “market basket analysis” result can be used to suggest combinations of products for special promotions or sales.

E.Credit card business CRM

Customer Relationship Management (CRM), through which, banks expect to identify the preference of different customer groups, products and services adapted to their liking to enhance the cohesion between credit card customers and the bank, has become a topic of great interest. The collective application of association rule techniques reinforces the knowledge management process and allows marketing personnel to know their customers well to provide better quality services. The idea behind this is to discover changes from two datasets and generate rules from each dataset to carry out rule matching.

II. RELATED WORK

In Apriori algorithm, in spite of being simple, has some limitation [1]. But the major advantages of FP-Growth algorithm is that it uses compact data structure and eliminates repeated database scan FP-growth is an order of magnitude faster than other association mining algorithms and is also faster than tree Researching. According to [2] availability of determine "Which groups or sets of items are customer's likely to quality services is vital for the well-being of the economy. Market basket circles are very seriously covering all important aspects of the services analysis may be performed on the retail data of customer transactions. According to [3] in situations with a large number of frequent patterns, long patterns, or quite low minimum support thresholds, an Apriori like algorithm may suffer from the following two nontrivial costs: – It is costly to handle a huge number of candidate sets. – It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. This is the inherent cost of candidate generation, no matter what implementation technique is applied.

III. PROPOSED ALGORITHM

A. *Apriori Algorithm*

Apriori is a popular data mining algorithm introduced by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for association rules. The basic algorithm steps are as follows:

Pass 1:

1. Generate the candidate item sets in CItem
2. Save the frequent item sets in FItem

Pass k

- ❖ Generate the candidate item sets in CItem_k from the frequent item sets in FItem_k-1
- ❖ Join FItem_k -1 p with FItem_k -1q, as follows:
    - ➢ Insert into CItem_k
    - ➢ Select p.item1, p.item2, p.itemk-1, q.itemk-1 from FItem_k -1 p, FItem_k -1q
    - ➢ Where p.item1 = q.item1, p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1
- ❖ Generate all (k-1)-subsets from the candidate item sets in CItem_k
- ❖ Prune all candidate item sets from CItem_k where some (k-1)-subset of the candidate item set is not in the
    - ▪ Frequent item set FItem_k-1

2. Scan the transaction database to determine the support for each candidate itemset in CItem_k
3. Save the frequent item sets in FItem_k

B. Advantages of Apriori Algorithm

1. Uses large item set property
2. Easily parallelized
3. Easy to implement
4. The Apriori algorithm implements level-wise search using frequent item property

C. Disadvantages of Apriori Algorithm

1. There is too much database scanning to calculate frequent item (reduce performance)
2. It assumes that transaction database is memory resident
3. Generation of candidate item sets is expensive (in both space and time)
4. Subset checking (computationally expensive)

- Multiple Database scans (I/O)

C.FP-Growth Algorithm

The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.When the database is large; it is sometimes unrealistic to construct a main memory based FP- tree. An interesting alternative is to first partition the database into a set of projected databases, and then construct an FP-tree and mine it in each projected database. Such a process can be recursively applied to any projected database if its FP-tree still cannot fit in main memory. A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm. It is also faster than a Tree-Projection algorithm, which recursively projects a database into a tree of projected databases.

- ➢ Step 1: FP Tree Construction: FP-Tree is constructed using 2 passes over the data-set.
- ➢ Pass 1:
- ❖ Scan data and find support for each item.
- ❖ Discard infrequent items.
- ❖ Sort frequent items in decreasing order based on their support.
- ❖ Use this order when building the FP-Tree, so common prefixes can be shared.
- ➢ Pass 2:
- ➢ Nodes correspond to items and have a counter.
- ❖ FP-Growth reads 1 transaction at a time and maps it to a path.
- ❖ Fixed order is used, so paths can overlap when transactions share items (when items have the same prefix).
    - ➢ In this case, counters are incremented
- ❖ Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)
    - ➢ The more the paths overlap, the higher the compression. FP-tree may fit in memory.
- ❖ Frequent item sets are extracted from the FP-Tree.

➢ Step 2: Frequent Item set Generation

❖ Each prefix path sub-tree is processed recursively to extract the frequent itemsets. Solutions are then merged.

➢ E.g. the prefix path sub-tree for e will be used to extract frequent itemsets ending in e, then in de, ce, be and ae, then in cde, bde, cde, etc.

➢ Divide and conquer approach

D. Advantages of FP- Growth Algorithm

1. Only 2 passes over data-set
2. “Compresses” data-set
3. No candidate generation
4. Much faster than Apriori

E. Disadvantages of FP- Growth Algorithm

1. FP-Tree may not fit in memory!!
2. FP-Tree is expensive to build

3 Trade-off: takes time to build, but once it is built, frequent itemsets are read off easily.

4 Time is wasted (especially if support threshold is high), as the only pruning that can be done is on single items.

5 Support can only be calculated once the entire data-set is added to the FP-Tree.

F. Simulated Annealing

Simulated annealing is a method for finding a good (not necessarily perfect) solution to an optimization problem. The travelling salesman problem is a good example: the salesman is looking to visit a set of cities in the order that minimizes the total number of miles he travels. As the number of cities gets large, it becomes too computationally intensive to check every possible itinerary. Broadly, an optimization algorithm searches for the best solution by generating a random initial solution and "exploring" the area nearby. If a neighbouring solution is better than the current one, then it moves to it. If not, then the algorithm stays put. Simulated annealing injects just the right amount of randomness into things to escape local maxima early in the process without getting off course late in the game, when a solution is nearby. This makes it pretty good at tracking down a decent answer, no matter its starting point. Also Simulated Annealing is easy to implement. Basic algorithm is as follows:

1. First, generate a random solution
2. Calculate its cost using some cost function you've defined
3. Generate a random neighbouring solution
4. Calculate the new solution's cost
5. Compare them:

a. If cnew < cold: move to the new solution

b. If cnew > cold: maybe move to the new solution

6. Repeat steps 3-5 above until an acceptable solution is found or you reach some maximum number of iterations.

IV. CONCLUSION AND FUTURE WORK

The paper majorly focuses on association rule mining techniques and its optimization using Simulated Annealing technique. Also different methods of association rule mining are also discussed. In this paper two classical mining algorithms- Apriori algorithm and FP- Growth algorithm are discussed. The applications area of mining algorithms is also identified. Our future work is to extract knowledge from multidimensional data in efficient manner.

REFERENCES

1. Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms, B.Santhosh Kumar and K.V.Rukmani,Int. J. of Advanced Networking and Applications,2010.

2. Ankur Mehay, Dr. Kawaljeet Singh, and Dr. Neeraj Sharma, “AnalyzeMarket Basket Data using FP-growth and Apriori Algorithm,” International Journal on Recent and Innovation Trends in Computing and Communication, 2013.
3. Jiawei Han, Jian Pei, Yiwen Yin And Runying Mao ,”Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach”,Data Mining and Knowledge Discovery, 8, 53–87, 2004
4. Rahul Mishra and Abha choubey, “Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012.
5. N. Naga Saranya, and M. Hemalatha, "Estimation of Evolutionary Optimization Algorithm for Association Rule using Spatial Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 51– No.3, August 2012.