Volume 4, Issue 2, April 2017 ISSN: 2348-8565 (Online)
IJMERT
International Journal of Modern Engineering and Research Technology
Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

A Review of Apriori with Association Rule Mining

Vivek Tiwari
M. Tech. Research Scholar
Shri Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: ruchitatiwari21@gmail.com

Anupam Choudhary
Lecturer,
Kalaniketan Polytechnic College
Jabalpur (M.P.), [INDIA]
Email:chowdharyanupam7@yahoo.com

Sapna Choudhary
Assistant Professor,
Department of Computer Science and Engineering.
Shree Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: choudharysapnajain@gmail.com

ABSTRACT

Data mining is a robotized advancement that usages confounded estimations to find associations in far reaching data bases Extensive improvement of data gives the motivation to find imperative cases among the colossal data. Progressive case gives us captivating associations between different things in successive database. Association Rules Mining (ARM) is a part of DM research space and develop various masters energy to plot a high capable count to mine connection rules from trade database. Association Rule Mining accept a basic part amid the time spent burrowing data for unending case organizing. It is a comprehensive methodology which uses to refine the mining systems. In programming building and data mining, Apriori is a praiseworthy figuring for learning alliance rules Apriori estimation has been essential count in connection lead mining. Apriori computation - an affirmation of ceaseless illustration organizing in perspective of support and sureness measures made heavenly results in various fields. Basic idea of this count is to find important cases between different course of action of data. It is an essential estimation yet having various drawbacks. Many explores have been proficient for the change of this computation. This paper shows a whole review on couple of good upgraded philosophies of Apriori computation. This will be genuinely to a great degree steady for the best in class researchers to find some new considerations from these approaches. The paper underneath layouts the fundamental arrangement of alliance principles close by the mining association estimations. The calculations incorporate the most essential Apriori calculation alongside different calculations, for example, AprioriTid, AprioriHybrid.

Keywords:— *Data Mining, Association Rule Mining, Apriori, Support, Confidence.*

I. INTRODUCTION

Due to growth of the data volume in the last decade, a set of different techniques for deletion of repetitive data and conversion of data to more usable forms has been proposed under the name of Data Mining. There exist a set of different techniques concerning the Data Mining, Some of which are decision trees, associative rules and Data Clustering. There are two categories of data mining namely, Descriptive data mining and Predictive data mining.

Descriptive data mining is used for carrying out summarizations or generalizations. Wherein, for finding out the inference or predictions, Predictive data mining is used.

II. Basic Data Mining Tasks

Data Mining Techniques

Techniques : Classification

It is often referred as "supervised learning". It has a predefined set of groups or models based on that values we predict.

Example:

Airport security maintains a set of metrics and tries to predict the terrorist.

Techniques : Regression

The regression using known data formats like linear or logistic assumes the future data format will fall into the data structure. It then tries to predict the value by applying some arithmetical algorithms on the dataset.

Example

Investing on Pension fund. Calculating your annual income and trying to predict what you need after you retire. Then based on the present income and needed income makes investment decision. The Prediction is done by simple regression formula to revise every year.

Techniques : Time Series Analysis

With time series analysis, every attribute value is determined by the different time intervals.

Example:

Buying a company stock. Take X, Y, Z companies' month by month performance and try to predict their next one year performance and based on the performance, you buy stocks.

Techniques : Prediction

Prediction is related with time series but not time bound. It is used to predict values based on past data and current data.

Example:

Water flow of a river will be calculated by various monitors at different levels and different time intervals. It then uses that information to predict the water flow of future.

Techniques : Clustering

It is widely called as unsupervised learning. It is similar to classification except it won't have any predefined groups. Instead the data itself defines the group.

Example:

Consider a super market has buying details like age, job and purchase amount we can group by age against percentage as well job against percentage to make meaningful business decisions to target the specific user group.

Techniques : Summarization

Summarization is associating the sample subset with small description or snippet.

Example:

Display the data as a graph and calculate mean, median, mode etc.

Techniques : Association rules

It is also called as linked analysis. It is all about under covering relationship among data.

Example:

Amazon "People bought this also bought this" model.

Techniques : Sequence discovery

Sequence discovery is about finding sequence of an activity.

Example:

In a shop, people may often buy toothpaste after toothbrush. It is all about what sequence user buying the product and based on the shop owner can arrange the items nearby each other.

II. Association Rule Mining

Association rule mining, a standout amongst the most critical and all around inquired about systems of information mining. It means to remove intriguing relationships, visit examples, affiliations or easygoing structures among sets of things in the exchange databases or other information storehouses. Affiliation lead mining is to discover affiliation decides that fulfill the predefined least support and certainty from a given database. Affiliation principles are broadly utilized as a part of different regions, for example, media transmission systems, market and hazard administration, stock control and so on.

The issue is generally disintegrated into two sub issues. One is to discover those thing sets whose events surpass a predefined edge in the database; those thing sets are called incessant or substantial thing sets. At that point other control are produced by erasing the last things in the predecessor and embeddings it to the subsequent, promote the confidences of the new standards are checked to decide the intriguing quality of them. The main sub-issue can be additionally separated into two sub-issues: competitor huge thing sets era handle and incessant thing sets era prepare. We call those thing sets whose support surpass the bolster limit as expansive or continuous thing sets, those thing sets that are normal or have the plan to be extensive or successive are called hopeful itemsets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "non redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength. An association rule is an *implication* or *if-then- rule* which is supported by data. The motivation given in for the development of association rules is *market basket analysis* which deals with the contents of point-of sale transactions of large retailers [5]. A typical association rule resulting from such a study could be 90 percent of all customers who buy bread and butter also buy milk". Insights into customer behavior may also be obtained through client overviews, yet the examination of the value-based information has the upside of being considerably less expensive and covering every present client. Contrasted with client overviews, the examination of value-based information has some serious confinements, be that as it may. For instance, purpose of-offer information ordinarily does not contain any data about individual interests, age and control of clients. In any case, advertise bushel examination can give new bits of knowledge into client conduct and has prompted higher benefits through better client relations, client maintenance, better item positions, item advancement and fraud detection.

III. Concept of Association Mining

Item : It is a field of the transaction database.

Transaction **:**It is corresponding to a record of the database. Transaction usually is marked as small letter t to mark item i. ti={i1,i2,…,ip}. Each transaction has an only identifier called TID. The whole set of transaction ti constitutes a database D. D={t1,t2,…,tn}

Support**:** The support of association rule X-> Y in transaction database is a ratio. The ratio is between the count of item set which contains X and Y, and the count of all of item set. That marks support(X->Y). That is the percent of the item set containing X and Y at the same time in the transaction database.

Confidence: It is the ratio between the count of transaction containing X and Y and the count of transaction containing X. That is marked as confidence (X->Y). Confidence is the percent of the transaction sets containing X and Y at the same time in the transaction database.

Frequent Item Set**:** The item set, whose support is not lower than the minimum support (Min Sup).

Strong rule and Weak rule**:** If support(X -> Y) >=MinSupport and Confidence (X->Y) >=MinConf, then mark association rule X-> Y as strong rule, otherwise mark it as a weak rule.

IV. Searching Frequent Itemset

Frequent patterns, such as frequent item sets, substructures, sequences term-sets, phrase -sets, and sub graphs, generally exist in real-world databases. Identifying frequent item sets is one of the most important issues faced by the knowledge discovery and data mining community. Frequent item set mining plays an important role in several data mining fields as association rules [6] warehousing, correlations, clustering of high-dimensional biological data, and classification. Given a data set d that contains k items, the number of item sets that could be generated is 2k - 1, excluding the empty set[6]. In order to searching the frequent item sets, the support of each item sets must be computed by scanning each transaction in the dataset. A brute force approach for doing this will be computationally expensive due to the exponential number of item sets whose support counts must be determined. There have been a lot of excellent algorithms developed for extracting frequent item sets in very large databases. The efficiency of algorithm is linked to the size of the database which is amenable to be treated. There are two typical strategies adopted by these algorithms: the first is an effective pruning strategy to reduce the combinational search space of candidate item sets (Apriori techniques). The second strategy is to use a compressed data representation to facilitate in-core processing of the item sets (FP-tree techniques)

V. Apriori Algorithm

In 1994 Agrawal etc. put forward famous Apriori algorithm according to the property of association rule: the sub sets of the frequent item set is also frequent item set, the supersets of non-frequent item set is also non- frequent item set. The algorithm each time makes use of k-frequent item set carrying on conjunction to get k+1 candidate itemset.

The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-item sets are used to explore (k+1)-item sets. The working of Apriori algorithm is fairly depends upon the Apriori property which states that" All nonempty subsets of a frequent item sets must be frequent" [5]. It also described the anti-monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test.Therefore if the one set is infrequent then all its supersets are also frequent and vice versa. This property is used to prune the infrequent candidate elements. In the beginning, the set of frequent 1-itemsets is found. The set of that contains one item, which satisfy the support threshold, is denoted by. In each subsequent pass, we begin with a seed set of item sets found to be large in the previous pass. This seed set is used for generating new potentially large item sets, called candidate item sets, and count the actual support for these candidate item sets during the pass over the data. At the end of the pass, we determine which of the candidate item sets are actually large (frequent), and they become the seed for the next pass. Therefore, is used to find, the set of frequent 2- itemsets, which is used to find, and so on, until no more frequent k-item sets can be found. The feature first invented by in Apriori algorithm is used by the many algorithms for frequent pattern generation.

The basic steps to mine the frequent elements are as follows:

1. Generate and test: In this first find the 1-itemset frequent elements by scanning the database and removing all those elements from which cannot satisfy the minimum supportcriteria.

2. Join step: To attain the next level elements join the previous frequent elements by self join i.e. known as Cartesian product of. i.e. This step generates new candidate k-item sets based on joining with itself which is found in the previous iteration. Let denote candidate k-item set and be the frequent k-itemset.

3. Prune step: is the superset of so members of may or may not be frequent but all frequent item sets are included in thus prunes the to find frequent item sets with the help of Apriori property. i.e. This step eliminates some of the candidate k-item sets using the Apriori property A scan of the database.

To illustrate this, suppose n frequent 1-itemsets and minimum support is 1 then according to Apriori will generate and so on. The total number of candidates generated is greater than Therefore suppose there are 1000 elements then 1499500 candidate are produced in 2 itemset frequent and 166167000 are produced in 3-itemset frequent.

It is no doubt that Apriori algorithm successfully finds the frequent elements from the database. But as the dimensionality of the database increase with the number of items then:

More search space is needed and I/O cost willincrease. Number of database scan is increased thus candidate generation willincrease results in increase in computational cost. Therefore many variations have been takes place in the Apriori algorithm to minimize the above limitations arises due to increase in size of database. These subsequently proposed algorithms adopt similar database scan level by level as in Apriori algorithm, while the methods of candidate generation and pruning, support counting and candidate representation maydiffer.

The algorithms improve the Apriori algorithms by:

1. Reduce passes of transaction database scans
2. Shrink number of candidates
3. Facilitate support counting of candidates

Pseudocode for Apriori algorithm

Join Step : Ck is generated by Lk-1 with itself

Prune Step: Any (k-1)- itemset that is not frequent cannot be subset of a frequent k-itemset

C_k: Candidate itemset of set k

L_k: frequent itemset of size k

L_1 = {frequent items};

for {k=1; L_1 !=0; k++) do begin

C_{k+1} = candidate generated from L_k;

for eachtransaction tin database do

increment the count of all candidates in C_{k+1} with min_support

end

return $U_k L_k$

Apriori-tid Algorithm:

Similar to the Apriori algorithm, the AprioriTid algorithm also uses the apriori-gen function to determine the candidate itemsets but the difference is that the database is not used for counting support after the first pass. Instead, set of candidate itemsets is used for this purpose of k>1. In case a transaction does not have any candidate k- itemset then the set of candidate itemsets would not have any entry for that transaction which will eventually decrease the number of transaction in the set containing the candidate itemsets as compared to the database. As value of k increases each entry will be smaller than the corresponding transactions because the number of candidates in the transactions will decrease. Apriori performs better than AprioriTid in the initial passes but in the later passes AprioriTid has better performance than Apriori.

Apriori –Hybrid:

Based on a different concept, this algorithm supports the idea that it is not necessary to use the same algorithm in all passes over data. As mentioned in [7], Apriori demonstrates better performance in earlier passes, and Apiori-TID outperforms Apriori in later passes. Based on the experimental observations, the Apriori-Hybrid technique was developed which uses Apriori in the initial passes and switches to Apriori-TID when it expects that the set *Ck* at the end of the pass will fit in memory. Therefore, an estimation of *Ck* at the end of each pass is necessary. Also, certain cost is involved in switching from Apriori to Apriori-TID. The performance of this technique was also evaluated by conducting experiments for large datasets. It was observed that Apriori-Hybrid performs better than Apriori except in the case when the switching occurs at the very end of the passes [5].

VI. Review on Various Improvements of Apriori Algorithm

Several improved algorithms have been proposed to conquer drawbacks of Apriori algorithm in several ways. Here presents six different approaches that face the common drawback.

6.1 Intersection and Record filter approach

6.1.1 Enlightenment

To present proposed algorithm, Goswami D.N., Chaturvedi Anshu and Raghuvanshi C.S.[4] has given Record filter and Intersection approach. In Record filter approach, count the support of candidate set only in the transaction record whose length is greater than or equal to the length of candidate set, because candidate set of length k, can not exist in the transaction record of length k-1, it may exist only in the transaction of length greater than or equal to k. In Intersection approach, to calculate the support, count the common transaction that contains in each element‟ s of candidate set. This approach requires very less time as compared to classical Apriori. In Proposed Algorithm, set theory concept of intersection is used with the record filter approach. In proposed algorithm, to calculate the support, count the common transaction that contains in each element‟ s of candidate set. In this approach, constraints are applied that will consider only those transaction that contain at least k items.

6.1.2. Disadvantage

Memory optimization is done but still it needs much optimization.

6.3. Improvement based on set Size frequency

6.3.1 Enlightenment

To eradicate non noteworthy hopeful keys the changed calculation presents issues, for example, set size and set size recurrence. These issues can decrease applicant enters in a more effective manner. The enhanced calculation for Apriori[5] takes for the set size which is the quantity of things per exchange and set size recurrence which is the quantity of exchanges that have in any event set size things. At first database is given with set size and second database is of set size recurrence of the underlying database. Evacuate things with recurrence not as much as the base bolster esteem at first and decide starting set size to get the most astounding set size whose recurrence is more noteworthy than or equivalent to least support of set size. Set size which are not more prominent than or equivalent to min set size support are dispensed with.

6.3.2 Disadvantage

Ideal starting size of combination size for pruning candidate keys is not given.

6.4 Improvement by reducing candidate set and memory utilization

6.4.1 Enlightenment

This algorithm[6] introduces a more efficient way to achieve the pruning operation. The algorithm only needs to search Lk-1 one time to complete the deletion and the remaining of each element X in Ck. The idea of the algorithm is as follows. Ik is a k-dimensional itemset. If the number of (k-1)-dimensional subsets of all (k-1)- dimensional frequent itemset Lk-1, which contains Ik, is less than k, then Ik is not a k-dimensional frequent itemset. So the improved algorithm only needs to match up the count of each element of Lk-1with the count of each element (X) of Ck (each element X has a count). If the count of the element X equals to k, then keep X. Otherwise X must be deleted. I/O speed can be deduced by cutting down unnecessary transaction records. The item that not appears in Lk-1 will no longer appear in Lk. So we can revise these items to null in the transaction database. Then we can pay no attention to these data information in any search work to D. At the same time, delete the transaction records (T) of which the number of valid data is less than k so as to deduce the database.[4] Then the candidate set Ck will be generated by latest D. The deletion of D will greatly reduce the number of transaction records which will effectively increase the speed of the implementation of the algorithm. Ultimately this will increase efficiency and I/O speed of algorithm.

6.5 Algorithm based on Tradelist

6.5.1 Enlightenment

This algorithm scans the database at the start only once and then makes the undirected item set graph.[7] From this graph by considering minimum support it finds the frequent item set and by considering the minimum confidence it generates the association rule. If database and minimum support is changed, the new algorithm finds the new frequent items by scanning undirected item set graph. That is why it" s executing efficiency is improved conspicuously compared to classical algorithm. It makes each item as a node(V) and at the same time it makes the supporting trade list for each node. Supporting trade list is a binary group T=

{Tid, Itemset} (where Tid is transaction id and Itemset is trade item set). So the side between nodes can be accomplished by corresponding trade list operation. The algorithm does the intersection of two nodes with supporting trade list. For search strategy select a node Vi from node set V. If the number of times Vi appears in the database is not less than the minimum support minsup, then {Vi} will belong to the frequent 1-item set. If count of node Vi adjacent to node Vj‟ s side is not less than support S, then { Vi, Vj } will belong to the item in frequent 2-iterm set. When there are three nodes in undirected item set graph and count of each side of the node is not less than minimum support minsup, these three nodes { Vk, Vm, Vn} will belong to frequent 3-item set. When there more than three nodes in undirected item sets graph then count of each side of the node should not be less than minimum support minsup and all the subset of these n nodes should be frequent. Subsequently nodes are added to k-item set. Main advantage of this approach is scanning of database is done once and after that the graph to find frequent itemset.

6.6 Algorithm based on frequency ofitems

6.6.1 Enlightenment

In this paper, Mamta Dhanda recommends an earth shattering and mindful approach for the mining of fascinating affiliation designs from exchange database.[8] First, visit examples are found from the value-based database utilizing the Apriori calculation. From the successive examples mined, this approach removes novel intriguing affiliation designs with accentuation on criticalness, quantity, profit and certainty. To beat the shortcoming of the conventional affiliation rules mining, Weighted affiliation administer mining[6] have been proposed. Weighted affiliation govern mining considers both the recurrence and hugeness of itemsets. It is helpful in distinguishing the most valuable and high offering things which contribute more to the company‟ s benefit. This approach proposes an effective thought in view of principally weight element and utility for mining of high utility examples. At first, the proposed approach makes utilization of the classical Apriori calculation to produce an arrangement of affiliation principles from a database. Right off the bat it utilizes ascribes to get visit thing set. These properties resemble benefit proportion figuring utilizing Q-calculate. Q – Factor = P/ΣPi (1) Thanit gives Transactional database where each item‟ s recurrence is checked in every exchange. From that pruning is finished with minsup and confedence. At long last figuring of Weighting –factor is done in view of recurrence of itemset and Q-calculate. n PW = Σ recurrence * Q – Factor (2) i=1 Finally proficient successive example is chosen in light of min PW-consider.

6.6.2 Disadvantage

Initially classical algorithm is used. To improve efficiency some improvement can be done on pruning for faster execution.

6. 7 Utilization of Attributes

6.7.1. Enlightenment

In this approach [10] using Tanagra Tool frequent item set is found by applying Apriori algorithm on database. Main problem of finding all association rules that satisfy minimum support and confidence thresholds given by users. Work illustrates that Association rule mining has several problems that it only tells whether item is present in database or absent, it treats all present or absent items equally, it does not consider importance of item to user/business perspective and it fails to associate output i.e. frequent items with user and business objectives. These disadvantages can be removed by using attributes like profit, quantity, frequency of items which will give important information to user and business.

6.2 Disadvantage

Various attributes like frequency, weight can be associated with frequent item set which can provide more information for business and user point of view, which is not done here.

VII. CONCLUSION

It is very essential to have an information mining calculation with high effectiveness since exchange database for the most part are substantial. Different calculations have been proposed for mining affiliation administer yet in each calculation there establishes a typical disadvantage of different outputs over the database. The point of this paper is to present affiliation control with apriori approach in different frame. In the wake of doing study of above calculations conclusion can be given by this paper is that generally in enhanced Apriori calculations, point is to create less hopeful sets but get every single incessant thing. In the approach of Intersection and Record channel, crossing point is utilized with the record channel approach where to figure the support, check the normal exchange that contains in every components of candidateset. In this approach, just those exchanges are viewed as that contain at any rate k things. In other approach set size and set size recurrence are considered.

As we know that each and everything has its required advantages and disadvantages So the Apriori algorithms also has its advantages and its various uses are given by

1. Initial information: transactional database D and user-defined numeric minimum support threshold min_sup
2. Algorithm uses knowledge from previous iteration phase to produce frequent itemsets.
3. This is reflected in the Latin origin of the name that means" from what comes before".
4. The various limitations of the Apriori algorithm are given by
5. Needs several iterations of the data.
6. Uses a uniform minimum support threshold.
7. Difficulties to find rarely occurring events.
8. Alternative methods (other than Apriori) can address this by using a non-uniform minimum support threshold.
9. Some competing alternative approaches focus on partition and sampling.

REFERENCES:

[1] Ben franklin, Genealogical data Mining, 2015.

[2] Maragatham G, Lakshmi M, "A Recent Review On Association Rule ining", Computer Science & Engg Department, Sathyabama University, Chennai, Tamil Nadu, India.

[3] Ramakrishnan Srikant,"Fast Algorithms for Mining Association Rules and Sequential Patterns", Ph.D Dissertation, 2014, University of Wisconsin, Madison.

[4] Qiankun Zhao,Sourav S. Bhowmick, "Association Rule Mining: A Survey", Nanyang Technological University, Singapore.

[5] Wael A. AlZoubi, Azuraliza Abu Bakar, Khairuddin Omar, "Scalable and Efficient Method for Mining Association Rules", Proc. Of IEEE, vol. 1,no.8, 2015, pp. 36-41.

[6] Lijuan Zhou, Shuang Li, Mingsheng Xu "Research on Algorithm of Association Rules in Distributed Database System", Proc. Of IEEE, vol. 3, no. 9, 2014, pp. 216-219

[7] Agrawal, R., Imielinski, T., and Swami, A. N. 2014. "Mining Association Rules Between Sets of Items in Large Databases".

[8] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, May 2014, pp. 207-216.

[9] Mamta Dhanda, "An Approach To Extract Efficient Frequent Patterns From Transactional Database", In: International Journal of Engineering Science and Technology (IJEST), Vol.3 No.7 July 2015, ISSN:0975-5462

[10] Andrew Kusiak, Association Rules "The Apriori Algorithm [Online], Available": http://www.engineering. uiowa. edu/~comp /Public Apriori.pdf

[11] Mamta Dhanda, Sonali Guglani, Gaurav Gupta, "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes", In: International Journal of Computer Science and Technology Vol 2, Issue 3, September 2013, ISSN:0976-8491

[12] David Wai-Lok Cheung, Vincent T. Ng, Ada Wai-Chee Fu, and Yongjian Fu, "Efficient Mining of Association Rules in Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No.6, December 2015.

[13] Kotsiantis, Dimitris Kanellopoulos," Association Rules Mining: A Recent Overview", Educational Software Development Laboratory Department of Mathematics, University of Patras, Greece.

* * * * *