International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 5 84 - 88

Development of Association Rule Mining with Efficient Positive and Negative Rules

Kavita.S.Yadav
3rd and 4th Sem M. Tech CSE,
Vidarbha Institute of Technology, Nagpur,India
kyadav63@gmail.com

Prof. Pravin Kulurkar
Assistant Professor,
Vidarbha Institute of Technology, Nagpur, India
pravinkulurkar@gmail.com

Abstract:- Association rule mining (ARM) is one of the most researched areas of data mining and recently from the database community it has received much attention. In the marketing and retail communities, they are proven to be quite useful in the other more diverse fields. On this area some of the previous research is done, the concept behind association rules are provided at the beginning followed by an overview to some research. The advantages and limitations are concluded with an inference. There are several algorithms, in frequent pattern mining. The classical and most famous algorithm is Apriori. To find frequent item sets and association between different items sets is the objective of using Apriori algorithm, i.e. association rule. In this paper author considers data (Online Seller transaction data) and tries to obtain the results using weak a data mining tool. To find out best combination, association rule algorithm are used of different attributes in any data.

Keywords: *ARM, Apriori, Transaction*

*****

INTRODUCTION

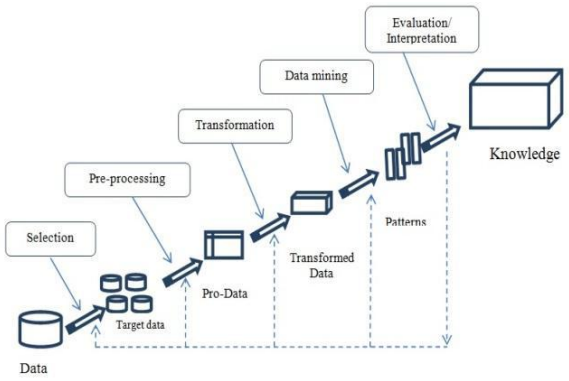
The discovery of hidden information is data mining which is found in databases and can be viewed as a step in the knowledge discovery process. The functions of data mining include classification, clustering, prediction and link analysis. The data mining applications which is important is that of mining association rules. Among a set of items in a database, association rules are used to identify relationships which was first introduced in 1993[Agrawal 1993]. These relationships are not based on inherent properties of data themselves, but rather based on co-occurrence of the data items.

For better understands let's take an example – A grocery store has weekly specials for which advertising supplements are created and distributed in the local newspaper. When an item such as peanut butter has to go on sale, management determines which are the companion items which are frequently purchased with peanut butter, they find that bread is purchased with peanut butter 30% of the time and also jelly is purchased with but with a slight more percentage that is 40% of the time. Based on these association, the special displays of butter is placed near jelly and bread to increase the sale of all the three items. But to increase the profits the jelly and bread are not on sale. These actions are aimed at increasing overall sales by taking advantage of the frequency with which these items are purchased together.

There are two association rules mentioned in the example stated above. The first rule states that when peanut butter is purchased bread is also purchased about 30% of the time from all the transactions. The second rule states that 40% of the time from the total transactions jelly is also purchased. Association rules are often used by retail stores to analyze market and the buying needs of the people using their transactional data. The invented association rules can be used by owners to increase the effectiveness of the sale and to reduce the cost associated with advertising, marketing, inventory and stock location on the floor. For other applications, association rules are also used such as prediction of failure in telecommunication networks by identifying what even occur before a failure.

Analyzers treat data mining that is finding the rules as the essential process of knowledge discovery in database as it can help in a variety of ways. It is also known as the extraction of information, pattern analysis, and data archaeology, information harvesting and business intelligence. The discovery of associations leads by frequent item set mining and correlations among items in the large transactional or relational databases. The aged algorithms for mining association rules built on binary attributes databases. By means of decreasing the times of database searching, an efficient algorithm should reduce the I/O operation of the process of mining.

The frequent patterns among item sets are discovered by association rule mining. Its aim is to extract interesting association, frequent patterns and correlations among sets of items in the data repositories. For example, 80% of the customers in the laptop store in India, who buy laptop computer also buy data card for internet and pen drive for data portability.



LITERATURE REVIEW:

In this section we will discuss some research papers which had been previously undertaken in the field of association rule mining is frequent pattern mining. Notion of association rules is introduced by Agrawal et al. Nahar J.et al. used the concept of frequent pattern in cancer prevention for which she used Apriori, Predictive Apriori, and Tertius algorithms. The mining of frequent item sets using Apriori algorithm with example was addressed by Jogi.Suresh and T.Ramanjaneyulu. Sunitha B. Aher and Lobo performed the comparative study of association rule algorithm for course recommender in system in E-learning.

Let's take a look to their work

Association Rules: Statements that find relationship between data in any database are known as association rule. Association rule has two parts Antecedent and Consequent. The item found in database is the Antecedent and Consequent is the item that is found in combination with the first. Association rules are generated during searching for frequent patterns in the system or the transaction.

The problem of finding association rules is divided into two sub problems

1. Find frequent item sets.
2. Find association rules from these item sets.

Association rule for calculating important relationships uses the criteria of Support and in detail in the following sections:

- Support (s): It is coined as an indication of item how frequently it occurs in the database. For a rule A=>B, its support is the percentage of transaction in database that contain union of B that is the similarity.
- Confidence (c): the number of times the statements found to be true in the dataset is termed as confidence. Confidence of the rule given above is the percentage of transaction in database containing A that also contain B.
- Lift: The calculation of lift is explained as follows:
  Lift(A=>B)=Supp(AUB)/Supp(B)*Supp(A)
- Conviction: it is very much similar to lift, but it measures the effect of the right-hand-side not being true which is different from lift and it also inverts the ratio.to calculate conviction following formula is used: Conviction= P (L). P(not R)/P(L,R)

Apriori:

The Apriori algorithm developed by [Agrawal 1994] is a great achievement in the history of mining association rule [Cheung 1996c]. The technique used by this property is that any sub set of a large item set must be a large item set. These common item sets are extended with other individual items in the transaction to generate candidate item sets. When the calculation of superset of large and small set is done the result obtained is again a small set. Due to generation of multiple small set large memory is used therefore it is an important issue addressed. The algorithm also generates the candidate item sets. These sets are created by joining the large item sets of the previous pass and deleting those sub sets which are small in the previous pass without considering the transactions in the database which is important part of the systems. The number of candidate large item sets will decrease if only large sets are consider for processing.

Algorithm:

The apriori_gen() function has two steps. The first step, Lk-1 is joined with itself to obtain Ck for the algorithm. Once the Ck is found the second step is implemented, where apriori_gen() deletes all item sets from the join result, which have some (k-1)-sub set that is not in Lk-1. Once the subset is found it returns the remaining large k-item sets.

Advantages of Apriori:

This algorithm in not complex and can be easily implemented and achieve good results. The data base is defined as D which contain the user transaction information and the supply is the min_supply() function is used. From the previously taken steps to produce the frequent item sets, algorithm uses the information.

Limitations of Apriori:

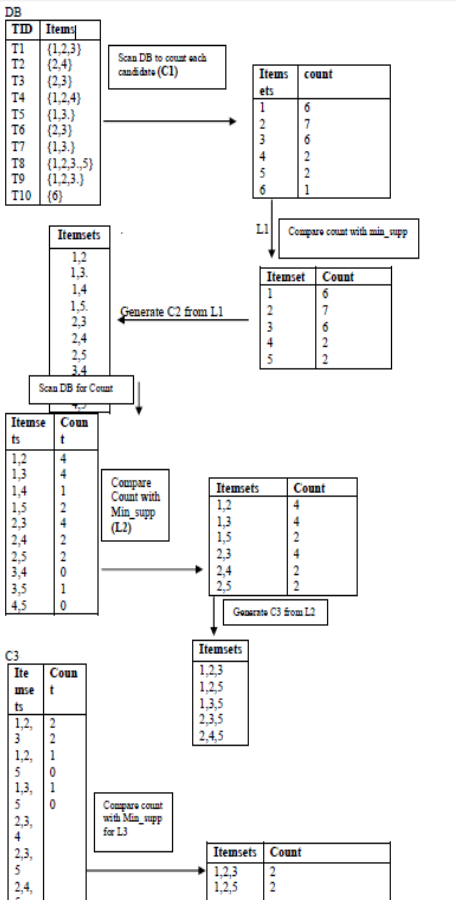
Large number of scans of dataset is required by Apriori. All items in Apriori are treated equally by using the presence and absence of items. Only the presence and the absence of an item is explained in transactional database. Here,

minimum threshold used is uniform. Whereas, other methods can address the problem of frequent pattern mining with non-uniform minimum support threshold. Apriori algorithm produce large number of candidate item sets, in case of the large dataset. For searching frequent item sets, algorithm scan database repeatedly, so more time and resources are required in large number of scans so it is inefficient in large datasets.

Method: apriori_gen() [Agrawal1994]

Input: set of all large (k-1)-item sets Lk-1
Output: A super set of the set of all large k-item sets
//Join step
Ii = Items I insert into Ck
Select p.I1, p.I2, ……. , p.Ik-1, q .Ik-1
From Lk-1 is p, Lk-1 is q
Where p.I1 = q.I1 and …… and p.Ik-2 = q.I k-2 and p.Ik-1 < q.Ik-1.
//pruning step
forall item sets c∈Ck do
forall (k-1)-sub sets s of c do
If (s∉Lk-1) then delete c from Ck

Generating Candidate Itemsets and Frequent Itemsets (Min_supp=2 (20%))



PROPOSED WORK

Association Rule Mining:

Association Rule Mining (ARM) major work is to discovers the frequent patterns among the item set. The work includes extracting the interesting associations, frequent pattern and correlation among set of items in the repository. To get a better understanding let's take an example, in any of the laptop store in India 80% customers of the laptop buys key guard with it. Here the relation is the mined data.

Agrawal specified the formal statement of the ARM. Let I = I1, I2, … , Im be a set of m different attributes, T be the transaction that comprises a set of items such that T subset of I, D be a database with different transactions Ts. The rule is an insinuation in the form of X →Y where both X and Y are the subsets of I and X ∩ Y = θ. X is termed as antecedent whereas Y is termed as consequent. The rule simply states X implies Y.

The basic support measures of the association rule is the Support and Confidence. Since the size of database is very large. The user is only concerned with the recent items that too frequently bought, to make this happen the user will pre-define the support and confidence which will act as a filter to the data, which are termed as the minimum confidence and minimum support. Both support and the confidence are explained as follows:

Support: it is defined as the proportion of set containing X u Y of the overall database. The amount of each item is the augmented by one when the item is crossed during the time of scanning the database for the entries. The support can be calculated by the following formula:

$$\text{Support (XY)} = \frac{\text{Support sum of XY}}{\text{Overall records in the database D}}$$

Confidence: it is calculated as the number of transactions that contain X u Y to the overall records that contain X, if the ratio performs better than the threshold of confidence can be generated. The equation is given as follows:

$$Confidence(X/Y) = \frac{Support\ (XY)}{Support\ (X)}$$

For ARM the confidence is the degree of strength, if the confidence of the rule is 80% then it implies that 80% of the transactions going have both X and Y in them. Likewise to confirm the interestingness of the rules specified minimum confidence is also predefined by the user of the system. Association Rule mining is to discover the rules to fulfil the demands made by the support and confidence. The problem is then divided into two, the first problem is to find the item

set, and the second is to generate association rules with minimum confidence. If one of the largeitem sets is Lk, Lk= {I1, I2… Ik-1, Ik}, then association rules are generated with those item sets. Checking the confidence with the rule {I1, I2… Ik-1} □ {Ik}, it can be decided for interestingness. But when we delete the last items the other rules are created placing it to the consequent. The confidence of the new rules are checked, this process is iterated till the antecedents become empty.

POSITIVE AND NEGATIVE ASSOCIATION RULES:

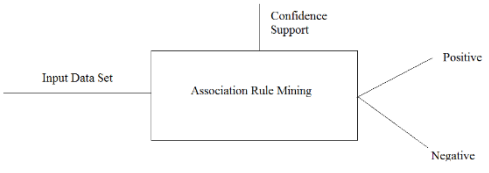
The most common framework in the ARM generation is the support and the confidence, although these two parameters allow the pruning of many associations that are discovered in data but there may be cases where multiple uninteresting rules may be produced in this system we consider another framework that adds to the support and confidence some measures based on correlation analysis.

Positive Association Rules: let's assume an example of an online store which sells daily need products. When a user goes to the site and orders for bread it is a great possibility that the user will buys butter with it. So to mine the data of a particular day that how many people buys bread and butter both with the total transaction. Let the X be the number of users who bought both bread and butter and Y be the total number of users so the positive rule is applied as X, Y → (12, 20). It means that 12 users have bought both the products.

Negative Association Rules: to understand the negative rule mining the same example goes by if in the online store the user is buying something but in this case say with the bread the user is buying almonds not the butter. So from the transactions number of user not buying the combination is the negative rule in the system. Let the X be the number of users who bought both bread and butter and Y be the total number of users so the positive rule is applied as X, Y → (5, 20). It means that 5 users have not bought both the products or we can say bought some other combination.

IMPLEMENTATION:

In this section the proposed system is explored and implemented. The block diagram of the implementation is as follows:



As shown in the figure the module has multiple parts which are: the input dataset, the Association rule mining, the confidence and support input and lastly the positive and negative results from the system. All the modules are explained as follows:

Input Data Set: This is the input of the proposed system. These are text files of the tractions on the store. The dataset majorly used is of Amazon and Flipkart. The dataset contains 1000, 2000, 5000 and 10000 transaction details that is 10000 users' details who bought products from the online store. From these files the rules are calculated. It contains the user with the products combination bought to process.

Confidence and support: This is another important input of the system. Since the confidence and the support are the major filters which helps in filtering the excess data from the results. For example we just wanted to check the number of user buying the cell phones with glass and cellphones without glass then these filters are required.

Association Rule Mining: This section is the brain of the system, here the inputted confidence and the support is applied upon the inputted transactional files and the positive and the negative rules are found out. The algorithm used to do the same is as follows:

Algorithm Positive and Negative Association Rules Generation
Input TD, $minsupp$, $minconf$, and ρ_{min}, respectively Transactional Database, minimum support, minimum confidence, and correlation threshold.
Output AR: Positive and Negative Association Rules.
Method:
(0) **if** ρ_{min} is undefined **then** ρ_{min} = 0.5
(1) $positiveAR \leftarrow \emptyset$; $negativeAR \leftarrow \emptyset$ /*positive and negative AR sets*/
(2) scan the database and find the set of frequent 1-itemsets (F_1)
(3) **for** ($k = 2, F_{k-1} \neq \emptyset, k++$){
(4) $C_k = F_{k-1} \bowtie F_1$
(5) **foreach** $i \in C_k$ {
(6) s=support(TD,i) /*support of item i is computed*/
(7) **if** s$\geq minsupp$ **then**
(8) $F_k \leftarrow F_k \cup \{i\}$ /*item i is added to F_k*/
(9) **foreach** X,Y ($i = X \cup Y$) {
(10) ρ=correlation(X,Y) /*correlation btw X and Y is computed*/
(11) **if** $\rho \geq \rho_{min}$ **then**
(12) **if** s$\geq minsupp$ **then**
(13) **if** $confidence(X \rightarrow Y) \geq minconf$ **then**
(14) $positiveAR \leftarrow positiveAR \cup \{X \rightarrow Y\}$
(15) **else if** $confidence(\neg X \rightarrow \neg Y) \geq minconf$ and $supp(\neg X \neg Y) \geq minsupp$ **then**
(16) $negativeAR \leftarrow negativeAR \cup \{\neg X \rightarrow \neg Y\}$
(17) **if** $\rho \leq -\rho_{min}$ **then** /*$\rho < 0$ and $|\rho| \geq \rho_{min}$ */
(18) **if** $confidence(X \rightarrow \neg Y) \geq minconf$ **then**
(19) $negativeAR \leftarrow negativeAR \cup \{X \rightarrow \neg Y\}$
(20) **if** $confidence(\neg X \rightarrow Y) \geq minconf$ **then**
(21) $negativeAR \leftarrow negativeAR \cup \{\neg X \rightarrow Y\}$
(22) }
(23) }
(24) }
(25) $AR \leftarrow positiveAR \cup negativeAR$
(26) **if** $AR = \emptyset$ **then** {
(27) $\rho_{min} = \rho_{min} - 0.1$
(28) **if** $\rho_{min} \geq 0$ **then go to** step (3)
(29) }
(30) **return** AR

For a better understanding let's assume that the data analyst has to find out the number of buyers buying cellphones with its Glass from a number of transaction, the need is to order the number of glasses according to the sale rate so if the sale of glass is more the order will automatically increase. So the analyst will first take the data which contains the recent

transactions as the old transactions are of no use to him. When the analyst enters the files the confidence and support is set so that the filtering is done.

Once the filter is done the outputs are out. As when to take the result of glass and cell phones the analyst will give the input files, once given the support and confidence are set such that the cell phones data is only processed and the other transactions are discarded from the system. From that the positive output rules are the once where the cell phone and the glass is bought together and the negative rule would be the remaining transactions which will helps the analyst in finding the solution. Let the X be the number of users who bought both bread and butter and Y be the total number of users so the positive rule is applied as X, Y → (12, 20). It means that 12 users have bought both the products.

CONCLUSION :

MOPNAR gives good positive and negative rules, and is suited for large scale data mining numerical applications

REFERENCES:

[1] T. Karthikeyan and N. Ravikumar, A Survey on Association Rule Mining in International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

[2] Amit A. Nanavati, Krishna, I. and ChitrapuraSachindra Joshi Raghu Krishnapuram. Mining Generalised Disjunctive Association rules. ACM 1-581 13-436-3/01/0011, 2001.

[3] Amitabha Das, Wee Keong Ng and Yew KwongWoon. Rapid Association Rule Mining. ACM, 1-58113-436/01/0011, 2001.

[4] Bhatnagar, S. Algorithm for finding association rules in distributed databases. 2nd IEEE International Conference on Parallel Distributed and Grid Computing, 915-920, Solan, 6-8 Dec, 2012.

[5] Gosain, A. and Bhugra, M. A comprehensive survey of association rules on quantitative data in data mining. IEEE Conference on Information & Communication Technologies, 1003-1008, JeJu Island, 11-12 Apr 2013.

[6] Maria-LuizaAntonieOsmar R. Za¨ıane, Mining Positive and Negative Association Rules: An Approach for Confined Rules.

[7] Antonie, M.L., Za¨ıane, O.: Mining positive and negative association rules: An approach for confined rules. Technical Report TR04-07, Dept. of Computing Science, University of Alberta (2004)