Abstract- all

In the field of Knowledge Discovery in Databases the effectiveness of mining association rules is important. Association Rules is a technique of data mining, where there is identifying the relationship between one item to another. For mining association rules Apriori algorithm is dominant. The idea of Apriori algorithm is to find the frequent sets from our transactional database. Through the frequent sets, obtain association rules, these rules must satisfy minimum support threshold and minimum confidence threshold. This paper presents improved method for deciding optimum minimum support threshold and minimum confidence threshold for a dataset, pruning of rules based on a contingency table, the decision about whether to go for lift or confidence in order to get rid of uninteresting, misleading and confusing association rules and Chi-square test for rules in order to get optimum association rules

1. Introduction - Aditya, Shreyansh

In data mining, the Mining association rule is important. How to create frequent itemsets is important. The prime aspect is to better the mining algorithm that how to deduce itemsets candidate in order to produce frequent itemsets effectively [1].

Market Basket Analysis basically contains two factors: Frequent itemsets and Frequent sequential patterns. A frequent itemset is a collection of items that are often purchased together. In technical terms, frequent itemsets are the itemsets which satisfy the minimum support threshold defined by the practitioner. For instance, a grocery store has stored a transactional database, which tells us what commodities customers buy. If a few items, for example say, Bread and Milk, are repeated in most of the transactions, then the two commodities Bread and Milk form a frequent itemset.

In Frequent sequential patterns, to understand this pattern, assume that we visit a computer store to study the transactional patterns of customers that buy goods. It becomes evident from multiple sales analysis that many customers follow a certain pattern while purchasing computers. For example, A consumer buys a laptop and in addition to it, will pair it with a purchase of an antivirus software system. The database suggests that this pairing of commodities occurs many times. This frequent occurrence of patterns in buying is termed as sequential pattern.

Frequent Item-set pattern and Frequent sequential pattern are principal to understanding market basket analysis. The retail stored can arrange or position all the commodities on the shelf in a certain manner such that it can be easily identified by the customer which in turn will result in the customer buying the items together.

1.1. Interesting measures/Rule evaluation metrics: - Shubham

a. **Support** is the count of item or itemset occurred in transactions. It basically measures the frequency of item or itemset [20]

   Support(X)=count of transactions in having X / Total number of transactions

b. **Confidence** of association rule basically depicts the percentage value of frequently occurring rules among all the groups containing rules. Confidence value helps in concluding the quality of the rule. Confidence in a rule is calculated by dividing the probability of the items that occur together by the probability of the occurrence of the antecedent.[14][20]

If X and Y are two items then Confidence of (X => Y) = support of (X, Y) / support of (X)

c. **Lift** gives information about the degree to which the inclusion or occurrence of one item "lifts" or increases the probability of. Lift is used to find misleading rules. Misleading rules are the rules which satisfy the minimum support and minimum confidence
Suppose a rule R1, which has Bread and Butter, and which derives Diapers

R1= {Bread, Butter} => {Diapers}
Let confidence of rule R1 3/5
Let support of diapers be 7/10

So hence support is greater than confidence, so it depicts that it is more likely to observe Diapers than together with Bread, Butter, and Diapers. So, this rule is misleading.
Lift ({ Bread , Butter } => { Diapers }) = (confidence of R1) / (Support of Diapers).For lift lower bound is zero and upper bound is infinity i.e. [0 , infinity)[15][16][20]

<u>RELATION BETWEEN LIFT, CONFIDENCE, AND SUPPORT</u>

- If CONFIDENCE > SUPPORT then LIFT is positively correlated
- If CONFIDENCE < SUPPORT then LIFT is negatively correlated
- If CONFIDENCE = SUPPORT then LIFT is 1 that is independent

d. **Conviction** estimates the anticipated inaccuracy of the rule. That is how the bread occurs in a transaction where milk does not, it basically measures the firmness of the rule with respect to the complement of consequent

Conviction(A=> B) = (1- support of (X)) / (1 - confidence of (X => Y))

So, it can be interpreted that the ratio of the expected frequency that A occurs without B.
This is used to predict the frequency that the rule makes an incorrect prediction
Leverage is estimating the contrast between observed and expected joint probability AB supposing that A and B are independent leverage

e. **Leverage** (A−→ B) = P (AB) − P (A) · P (B) = rsup(AB) − rsup(A) · rsup(B) Jaccard : The Jaccard coefficient estimates the similarity between two sets.[19]

1.2. Association rules- Mohit

Association rules is a pattern that states that when an event occurs another even occurs with a certain probability. Association rules help to find correlation between independent data in a data repository. It finds a correlation between the items that are used together regularly.
Association rules find all sets of items that have support greater than minimum support. Later, it uses huge item-sets to produce desired rules that have confidence which is greater than minimum confidence. Market basket analysis is an example of association rule. Association rules use the support and confidence terminologies to find the correlation and rules that are produced for analyzing data for repeated patterns. Association rules should work with user-specified minimum support and user-specified minimum confidence.[13][18][20][21]

Let T1,T2,T3,T4,T5 be the transactions A,B,C,D,E be items

Transactions	Items
T1	A,B,C
T2	A,C,D
T3	B,C,D
T4	A,D,E
T5	B,C,E

RULES	SUPPORT	CONFIDENCE	LIFT
A => D	2/5	2/3	10/9
C => A	2/5	2/4	5/6
A => C	2/5	2/3	5/6
B and C => D	1/5	1/3	5/9

1.3. Other improvements in apriori: (Not implemented)

a. Multilevel

i. Multilevel association rules are used to discover the frequently arising patterns and convincingly powerful rule inference

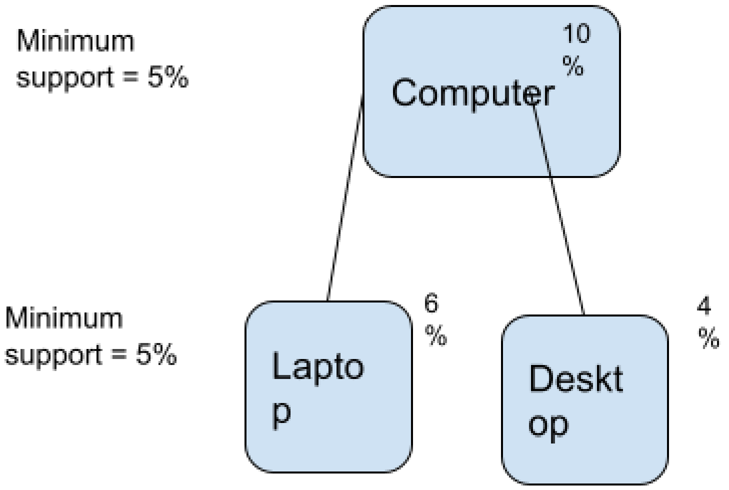
ii. It is a mining association rules from a multilevel structure

iii. The multilevel association rules can be executively mined utilizing concept hierarchies, which explains a sequence of mappings from a set of low level concepts to higher-level with more general concepts

iv. Here top down proceed is used ,where the support threshold changes from level to level ,High abstraction level nodes have high support as compared with lower nodes and lower abstraction level has low support

v. There are two approaches which can be used to implement this technique

1. Using Uniform minimum support for all levels



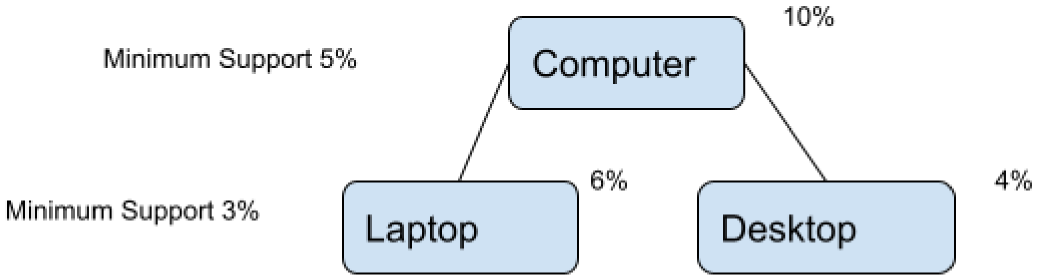
Here all abstraction levels are given the same minimum support value. But in this case problem will occur suppose in order to match higher abstraction level if consider the minimum support value high that is suppose the minimum support considered ne 8% , then only computer can be qualified and other lower abstraction level are nor qualified or if suppose the minimum support considered ne 2% that is very low then we will get unnecessary rules

2. **Reduced support**

a. Here as we move to lower abstraction level me deduce the minimum support value

b. Here every level of abstraction has its own minimum support value

c. The above example also depicts the same as we move to the lower level or minimum support value decreases.



2.

3. Multidimensional

   a. Multidimensional association rules are rules that are based on the table structured database where each item has distinct attributes which are not detected in association rules.

   b. Each item in the rule is paired with attributes and values. In a Multidimensional rule items in the rule refer to two or more dimensions or predicates. Hence it can be said that it has no repeated predicates.

c. Multidimensional association rules with no frequent predicates are called inter-dimensional association rules.

d. The main difference between association rules and multidimensional association rules is that each item in multidimensional association rules has unique attributes, but there is no such attribute in the items of association rules.

2. **Methodology**:- Mohit, Shreyansh, Shubham--

1. Apriori algorithm
    a. Loading the database. The file basically contains transactional data. Each row can be considered as a list of items bought by a customer.
    b. We are dividing our records in a 30:70 ratio that is for training and testing. We are dividing the file in order to validate our algorithm.
    c. Then we are creating a candidate set of unique items.
    d. After that, we are finding count as well as support of each and every item
2. Deciding minimum support for a dataset
    a. After this, we prune our candidate set by using minimum support value. We only keep the items whose support count is greater than or equal to minimum support value.
    b. After getting a list of items satisfying the condition each and every element present in the list of items is compared with every other element in the same list and sets are generated later. Support is calculated for these sets.
    c. The same is repeated for frequent itemsets of increasing length till there are no itemsets satisfying the minimum support condition.
3. Deciding minimum confidence for a dataset
    a. In order to prune association rules by minimum confidence, a loop from 0.1 to 1 with increment 0.1 is applied.
    b. We find the confidence value of each and every association rule and compare it with the minimum confidence.
    c. In this way we get association rules with corresponding values of confidence with respect to varying confidence,
    d. Among these values the confidence value giving optimum number of rules is selected as minimum confidence.
    e. Then by using this minimum confidence we are getting our association rules.
4. Pruning of rules based on a contingency table
    a. The contingency table is created for all the valid association rulesWe decide whether confidence is sufficient enough or lift is required for the association rule.
    b. Lift is required when the confidence is not sufficient enough to eliminate misleading rules,
    c. Lift prunes misleading rules
    d. This is done for both training and testing data

e. After finding association rules for both training and testing data and find the common rules.
f. This comparison is basically done in order to check the accuracy

5. Deciding minimum support for a dataset

a. Minimum support of the rule is the minimum number of transactions that encompass both the antecedent and consequent parts in order to certify to be part of frequent itemset
b. So in order to certify to be part of the frequent itemset, a minimum support value must be provided in order to prune the item set, So in order to deal with this issue of entering minimum support value we are finding central tendencies of the support values of items
c. And this value is used as minimum support in order to create our candidate set This will be done for both for training and testing data

6. Deciding minimum confidence

a. In order to prune item sets by considering minimum confidence a loop is running that starts with 0.1 and stops at 1 and increments by 0.1 and these values in the loop are the values of our minimum confidence.
b. Here we are using our list of pruned elements and our item sets created. we are running a loop from 1 to length of our pruned elements list, Here we are choosing item sets consisting elements more than one that is if our itemset contains more than one element, we find confidence value of each and every set and compare it with the minimum confidence and check whether it is valid or not
c. In this way we get item sets with corresponding values of confidence with respect to varying confidence from 0.1 to 1, then we find the median of all our confidence values and consider that value as our optimum minimum confidence.

7. Pruning of rules based on a contingency table

a. Here the focus is on how to eliminate strong association rules that can be uninteresting, misleading, and confusing.
b. Contingency table basically gives information about the itemsets and their relation with each other, that is the presence of one of them, presence of both or absence of both items in the transactions.
c. Of the 10,000 transactions analyzed, 6,000 of the customer transactions included Beer, while 7,500 included milk, and 4,000 included both Beer and Milk.
d. Let the minimum support be 30% and minimum confidence of 60% and it is found out that the rule forms strong association rule by satisfying minimum threshold support and confidence

e. But the rule is misleading as the probability of buying Milk is 75%, which is greater than Beer.

f. But Beer and Milk are negatively associated with each other. So we can say that confidence is not sufficient enough to find the optimum association rules

g. So the correlation analysis using lift is implemented, which will filter out strong misleading association rules.

h. Misleading rules would be pruned out of our contingency table

*/

5. The decision about whether to go for lift or confidence

i. By the example of beer and Milk,we can say that confidence is not sufficient enough to find the optimum association rules .So the correlation analysis using lift is implemented, which will filter out strong yet misleading association rules.

	Beer	Not Beer	Total
Milk	4000	3500	7500
Not Milk	2000	500	2500
Total	6000	4000	10000

j. The confidence value of Beer and Milk is 0.666 and confidence value of Beer and Not Milk is 0.333, As Beer and Milk forms strong association rules but practically they have no relation this is main drawback of confidence, There is need to select the rules which have proper relation between each other, in order to achieve it we find lift

k. The value of the lift of our itemset Beer and Milk is 0.89. Here the value is calculated by using the lift formula.

l. According to the condition, the value is less than 1, So we can say that there is a negative correlation between the Beer and Milk

m. As there is negative correlation means the rule is not valid,So the rule is of no use hence it will be dumped out of our contingency table.

n. So in these cases finding lift is important to know the relation between items of our itemset in the contingency table

6. Correlation Analysis: (Not implemented)

CHI-SQUARE TEST: Mohit

Supervised Machine Learning feature extraction is a very responsible task, as to determine the correct output from the model. For best results , selection of the right features is important. Chi-Square Test is one of the best techniques for feature selection. This test is perfect for "categorical variables".

For each feature in the dataset, the $\mathbf{X^2}$ is calculated and then ordered in descending order according to the $\mathbf{X^2}$ value. The higher the value of $\mathbf{X^2}$, the more dependent the output label is on the feature and higher the importance the feature has on determining the output.

The formula to calculate chi-squared value($\mathbf{X^2}$) is:

$\mathbf{X^2 = \Sigma(O - E)^2 / E}$

O = Observed Frequency

E = Expected Frequency

Categorical Variable:

Categorical variables are the qualitative variables which define the quality and characteristics of particular variables. These variables can be divided into finite groups.

For eg. The category "Literary Genres" in the list of literatures could contain the categorical variables such as, Fiction, Novel, Thriller, Mystery Satire, etc.

The Categorical variables can be divided into 2 types:

1. Nominal variable ; 2. Ordinal Variable

Nominal variable:

These are the variables which do not have any natural ordering to its categories.
For eg. Gender(Male, Female, Transgender), Marital Status(Married, Unmarried)

Ordinal variable:

These are the variables which have some sort of ordering to its categories.
For eg. User Satisfaction(Outstanding, Excellent, Very Good, Good, Average, Bad)

ASSUMPTIONS OF CHI-SQUARE TEST:

- Data must be in frequencies or counts. It must not be in percentages.
- Observations should be independent of each other.
- Categories should be mutually exclusive, meaning each subject should fit in one category only. For e.g. Recorded observations of people attending seminar on Saturday should not mix with people attending a seminar on Sunday.

TYPES OF CHI-SQUARE TEST:

Goodness of Fit :

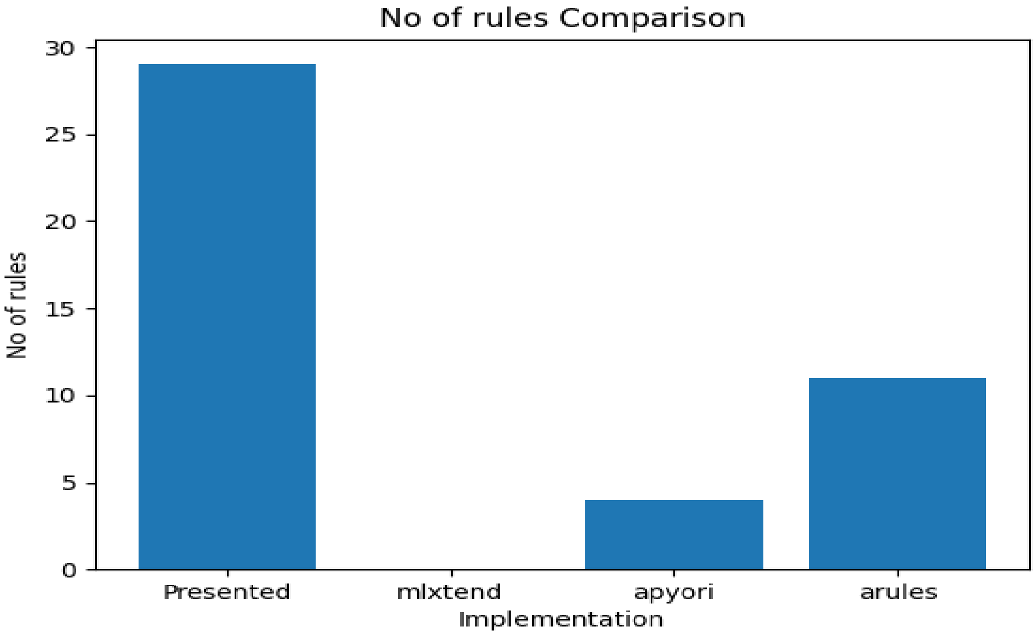
Used to find the significant difference between observed and expected observations.

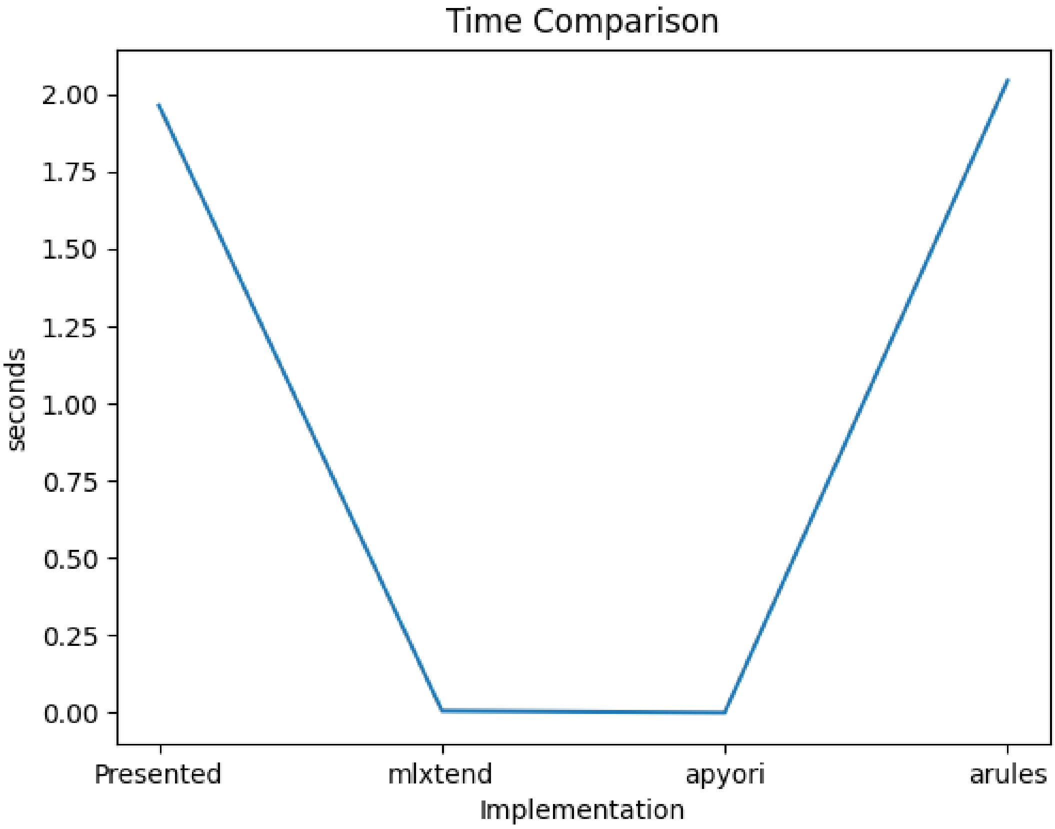
Test For Association/Independence:

This test is used when there is categorical data for two independent variables and we want to see if there is any relationship between the variables.

6. Results-

1. Graphs of time complexity analysis





2. Results of pruning

a. Minimum support

i. This is the first method for pruning,in which we prune our candidate set base on the threshold minimum support

ii. So only items which satisfy minimum support or itemsets generated out of items which satisfy minimum support, are selected and else are pruned

b. Minimum confidence

iii. In order to check whether itemsets generated have proper and valid association with each other we find confidence value of the rule and compare it with threshold minimum confidence value and get the valid rules, remaining are pruned

iv. Disadvantage of confidence measure is that it might misstate the significance of an association.

v. This is because it only reports how popular milk is, but not beer. If in general beer is also very popular ,there might be a big chance that a transaction contains both milk as well as beer, thus confidence measures are increased. To report for the base popularity of both constituent items, we use a third measure called lift.

c. Pruning of rules based on a contingency table

vi. Here the focus is on how even powerful association rules can be uninteresting ,misleading and confusing.Then how the support-confidence framework can be increased with additional interestingness measures based on statistical significance and correlation analysis

vii. As support and confidence measures are inadequate at filtering out uninteresting association rules. So a correlation measure can be used to augment the support-confidence framework for association rules. This leads to form correlation rules

d. Training and Test association rules

viii. In order to get optimum rules, we divide our dataset into two parts that are 70 as to 30 ratios and after generating final association rules, we compare all the rules and select common between them

ix. Other remaining rules are pruned

x. Rules after this process are our final optimum association rules