Name of Faculty: Prof. Puneet Nema
Designation: Assistant Professor

Department: CSE

Subject: Data mining

Unit: IV

Topic: Association Rule Mining:-Introduction, Basic, The Task and a Naïve Algorithm, Apriori Algorithms, Improving the efficiency of the Apriori Algorithm, Apriori-Tid, Direct Hasing and Pruning(DHP),Dynamic Itemset Counting (DIC), Mining Frequent Patterns without Candidate Generation(FP-Growth),Performance Evaluation of Algorithms,..

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

Computer Science and Engineering,VIII-Semester

CS-8203 Data Mining

UNIT-IV

Topic Covered: Data Mining

Association Rule Mining:-Introduction, Basic, The Task and a Naïve Algorithm, Aprior Algorithms, Improving the efficiency of the Apriori Algorithm, Apriori-Tid, Direct Hasing and Pruning(DHP),Dynamic Itemset Counting (DIC), Mining Frequent Patterns without Candidate Generation(FP-Growth),Performance Evaluation of Algorithms,.

Introduction to Association Rule Mining :

Association rules are if/then statements that are meant to find frequent patterns, correlation, and association data sets present in a relational database or other data . Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.present. And, this is the reason why data mining has become such an important area of study.
We use techniques for a long process of research and product development. As this evolution was started when business data was first stored on computers.

Association rule mining, at a basic level, involves the use of models to analyze data for patterns, or co-occurrence, in a database . It identifies frequent if-then associations, which are called *association rules*.
An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data.
A consequent is an item found in combination with the antecedent.moves on to cover topics such as knowledge .

Example of Association Rule Mining:

Classic example of association rule mining refers to a relationship between diapers and beers. The example, which seems to be fictional claims that men who go to a store to buy diapers are also likely to buy beer. Data that would point to that might look like this.

A supermarket has 200,000 customer transactions. About 4,000 transactions, or about 2% of total transactions, include the purchase of diapers. About 5,500 transactions (2.75%) include the purchase of beer. Of those, about 3,500 transactions, 1.75%, include both the purchase of diapers and beer. Based on the percentages, that number should be much lower. However, the fact that about 87.5% of diaper purchases include the purchase of beer indicates a link between diapers and beer.

Example:

Milk -> Bread{Support = 2%, Confidence = 60%}

A **support** of 2% for Association rule means that 2% of all the transactions show that milk and bread .

And 60% of **confidence** means 60% of all the customers who buy milk also bought bread.

Association rule is considered interesting if it satisfies both minimum support and minimum confidence

Application of Association Rule Mining:

- Market based data analysis
- Catelog Design
- Cross Marketing.

Apriori Algorithms:

- Apriori algorithm, a classic algorithm, is useful in mining frequent itemsets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket.
- It has got this odd name because it uses 'prior' knowledge of frequent itemset properties. The credit for introducing this algorithm in 1994. We shall now explore the apriori algorithm implementation in detail.
- Three significant components comprise the apriori algorithm. They are as follows.
    - Support
    - Confidence
    - Lift

This example will make things easy to understand.

- As mentioned earlier, you need a big database. Let us suppose you have 2000 customer transactions in a supermarket. You have to find the Support, Confidence, and Lift for two items, say bread and jam. It is because people frequently bundle these two items together.
- Out of the 2000 transactions, 200 contain jam whereas 300 contain bread. These 300 transactions include a 100 that includes bread as well as jam. Using this data, we shall find out the support, confidence, and lift.

improved without altering the external behavior or code design.

Support:

Support is the default popularity of any item. You calculate the Support as a quotient of the division of the number of transactions containing that item by the total number of transactions. Hence, in our example,

Support (Jam) = (Transactions involving jam) / (Total Transactions)

= 200/2000 = 10%

Confidence:

In our example, Confidence is the likelihood that customers bought both bread and jam. Dividing the number of transactions that include both bread and jam by the total number of transactions will give the Confidence figure..

Confidence = (Transactions involving both bread and jam) / (Total Transactions involving jam).

= 100/200 = 50%

It implies that 50% of customers who bought jam bought bread as well.

Lift:

According to our example, Lift is the increase in the ratio of the sale of bread when you sell jam.

Lift = (Confidence (Jam – Bread)) / (Support (Jam))

=50 / 10 = 5.

It says that the likelihood of a customer buying both jam and bread together is 5 times more than the chance of purchasing jam alone. If the Lift value is less than 1, it entails that the customers are unlikely to buy both the items together. Greater the value, the better is the combination..

This refers to the summary of general characteristics or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such produc running SQL queries,

It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts

How does Aprioro Algorithms Work:
Example

Consider a supermarket scenario where the itemset is I = {Onion, Burger, Potato, Milk, Beer}. The database consists of six transactions where 1 represents the presence of the item and 0 the absence..

Transaction id= {t1,t2,t3,t4,t5,t6}
Onion= {1,0,0,1,1,1}
Potato= {1,1,0,1,1,1}
Berger= {1,1,0,0,1,1}
Milk= {0,1,1,1,0,1}
Peer= {0,0,1,0,1,1}

The Apriori Algorithm makes the following assumptions

Step 1:

Create a frequency table of all the items that occur in all the transactions. Now, prune the frequency table to include only those items having a threshold support level over 50%. We arrive at this frequency table .

Item= {onion,potato,berger,milk}
Frequency of Items= {4,5,4,4}

Step 2

Make pairs of items such as OP, OB, OM, PB, PM, BM. This frequency table is what you arrive at.

Item set	Frequency
OP	4
OB	3
OM	2
PB	4
PM	3
BM	2

Step 3:

Apply the same threshold support of 50% and consider the items that exceed 50% (in this case 3 and above).

Thus, you are left with OP, OB, PB, and PM

Step 4:

Look for a set of three items that the customers buy together. Thus we get this combination.

OP & OB gives OPB

PB & PM gives PBM

Step 5:

Item set	Frequency
OPB	4
PBM	3

Determine the frequency of these two itemsets. You get this frequency table.

- If you apply the threshold assumption, you can deduce that the set of three items frequently purchased by the customers is OPB.
- We have taken a simple example to explain the apriori algorithm in data mining. In reality, you have hundreds and thousands of such combinations.

Apriori Algorithms : Pros.

- Easy to understand and implement
- Can use on large itemsets

Apriori Algorithms : Cons

- At times, you need a large number of candidate rules. It can become computationally expensive.
- It is also an expensive method to calculate support because the calculation has to go through the entire database.

How to Improve the Efficiency of the Apriori Algorithm:

Use the following methods to improve the efficiency of the apriori algorithm.

1. **Transaction Reduction:**
   - A transaction not containing any frequent k-itemset becomes useless in subsequent scans.

2. **Hash-based Itemset Counting :**
   - Exclude the k-itemset whose corresponding hashing bucket count is less than the threshold is an infrequent itemset.
   - Algorithms are easier to maintain, test, and reduce error propagation and can be reused in other programs as well. Thus, functional independence is a good design feature which ensures databases quality.

Direct Hashing and Pruning:

Hashing & Pruning is very popular association rule mining technique to improve the performance of traditional Apriori algorithm. Hashing technique uses hash function to reduce the size of candidate item set.

Direct Hashing & Pruning (DHP),Perfect Hashing &Pruning (PHP) are the basic hashing algorithms. Many algorithms have been also proposed by researchers like Perfect Hashing Scheme (PHS), Sorting-Indexing andTrimming (SIT),HMFS etc.

THP arranges the item sets into vertical format and then hashed the transactions id (TID) of candidate-k item sets into hash table bucket corresponding to that item set.

Working of DHP algorithm:

Step1: Scan the database to count the support of candidate-(C1) item set and select the items : count>=min_sup to add into large item set(L1).

Step 2: Now make possible set of candidate-2 item set in each transaction of database (D2). Hash function is applied on each candidate-2 item set to find the corresponding bucket number.

Step 3: Scan database (D2) and hash each item set of transactions into corresponding hash bucket. Some item sets are hashed into same bucket this is called collision problem.

Step4: Select only that candidate-2 item setwhose corresponding bucket count>=min_sup.If there is no collision then adds into L2.

a. If there is no collision then add the selected item sets into L2.
b. Else one more scan of database is required to count the support of collided item sets and the item set

Step 5: Now make possible set of candidate-3 item set (D3) and repeat the same procedure until Ck=’

Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not.

Pros & Cons of DHP algorithm:

1. DHP uses simple hash function to reduce the size of candidate item set.
2. Size of hash table is small which requires less memory to store.
3. There is collision problem in DHP algorithm.
4. More database scans are required to count the support of collided item .

Pruning Methods:

In machine learning and data mining, pruning is a technique associated with decision trees. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood. This post will go over two techniques to help with overfitting - pre-pruning or early stopping and post-pruning with examples.

Pruning or post-pruning:

As the name implies, pruning involves cutting back the tree. After a tree has been built (and in the absence of early stopping discussed below) it may be overfitted. The CART algorithm will repeatedly partition data into smaller and smaller subsets until those final subsets are homogeneous in terms of the outcome variable. In practice this often means that the final subsets (known as the *leaves* of the tree) each consist of only one or a few data points. The tree has learned the data exactly, but a new data point that differs very slightly might not be predicted well.

Caterogy of Pruning:

Minimum error:

The tree is pruned back to the point where the cross-validated error is a minimum. *Cross-validation* is the process of building a tree with most of the data and then using the remaining part of the data to test the accuracy of the decision tree.

Smallest tree:

The tree is pruned back slightly further than the minimum error. Technically the pruning creates a decision tree with cross-validation error within 1 standard error of the minimum error. The smaller tree is more intelligible at the cost of a small increase in error.

Dynamic Itemset count :

A set of items together is called an itemset. If any itemset has k-items it is called a k-itemset. An itemset consists of two or more items. An itemset that occurs frequently is called a frequent itemset**.**

Thus frequent itemset mining is a data mining technique to identify the items that often occur together.

- **For Example**, Bread and butter, Laptop and Antivirus software, etc.

- A set of items is called frequent if it satisfies a minimum threshold value for support and confidence. Support shows transactions with items purchased together in a single transaction. Confidence shows transactions where the items are purchased one after the other.

- For frequent itemset mining method, we consider only those transactions which meet minimum threshold support and confidence requirements. Insights from these mining algorithms offer a lot of benefits, cost-cutting and improved competitive advantage., no direct coupling.

  - Frequent itemset or pattern mining is broadly used because of its wide applications in mining association rules, correlations and graph patterns constraint that is based on frequent patterns, sequential patterns, and many other data mining tasks..
  - Here a virtual mediated schema provides an interface that takes the query from the user,
  - Transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.
  - In this approach, the data only remains in the actual source databases. However, mediated schema contains several "adapters" or "wrappers" that can connect back to the source systems in order to bring the data to the front end.

Frequent Pattern Mining (FPM):

The frequent pattern mining algorithm is one of the most important techniques of data mining to discover relationships between different items in a dataset. These relationships are represented in the form of association rules. It helps to find the irregularities in data.

FPM has many applications in the field of data analysis, software bugs, cross-marketing, sale campaign analysis, market basket analysis, etc.

Frequent itemsets discovered through Apriori have many applications in data mining tasks. Tasks such as finding interesting patterns in the database, finding out sequence and Mining of association rules is the most important of them.

Association rules apply to supermarket transaction data, that is, to examine the customer behavior in terms of the purchased products. Association rules describe how often the items are purchased

Frequent Pattern Algorithms:

The probability that item I is not frequent is if:

- (I) < minimum support threshold, then I is not frequent.
- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

The steps followed in the Apriori Algorithm of data mining are:

- **Join Step**: This step generates (K+1) itemset from K-itemsets by joining each item with itself.

- **Prune Step**: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Steps In Apriori:

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database . This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

Step1:

In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

Step2:

Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned..

step3:

Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

Step4:

The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

Step 5:

The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

Step6:

Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does