K.T.S.P.Madal's

Hutatma Rajguru Mahavidyalaya, Rajgurunagar

Tal-Khed, Dist.-Pune 410505.

TY.BSc (Computer Science)

Semester-VI

Subject- Data Analytics

According to new CBCS syllabus w.e.f.2019-2020

Prof.S.V.Patole

Department of Computer Science

Hutatma Rajguru Mahavidyalaya,

Rajgurunagar.

Chapter 3. Mining Frequent Patterns, Associations and Correlations

Overview of Data Mining:

The data mining tutorial provides basic and advanced concepts of data mining. Our data mining tutorial is designed for learners and experts.

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.

- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

What Kind of Patterns can be Mined?

Different types of data can be mined in data mining. However, the data should have a pattern to get helpful information.

Descriptive patterns

It deals with the general characteristics and converts them into relevant and helpful information.

Descriptive patterns can be divided into the following patterns:

- **Class/concept description**: Data entries are associated with labels or classes. For instance, in a library, the classes of items for borrowed items include books and research journals, and customers' concepts include registered members and not registered members. These types of descriptions are class or concept descriptions.

- **Frequent patterns**: These are data points that occur more often in the dataset. There are many kinds of recurring patterns, such as frequent items, frequent subsequence, and frequent sub-structure.
- **Associations**: It shows the relationships between data and pre-defined association rules. For instance, a shopkeeper makes an association rule that 70% of the time, when a football is sold, a kit is bought alongside. These two items can be combined together to make an association.
- **Correlations**: This is performed to find the statistical correlations between two data points to find if they have positive, negative, or no effect.

Class/Concept Description:

characterization and discrimination in data mining

Characterization: data characterization is a method useful for the derivations of descriptions of data classes or concepts. Basically, characterization is a summary of the general features (characteristics) of a target class (the data of the class being studied). OLAP roll-ups performed on data cubes are useful for summarization. Ex: a data mining system can be used, by city planners, to create descriptions summarizing residents that moved from urban to suburban areas in a ten-year period. The result could a general profile of residents, such as their age, income, etc.

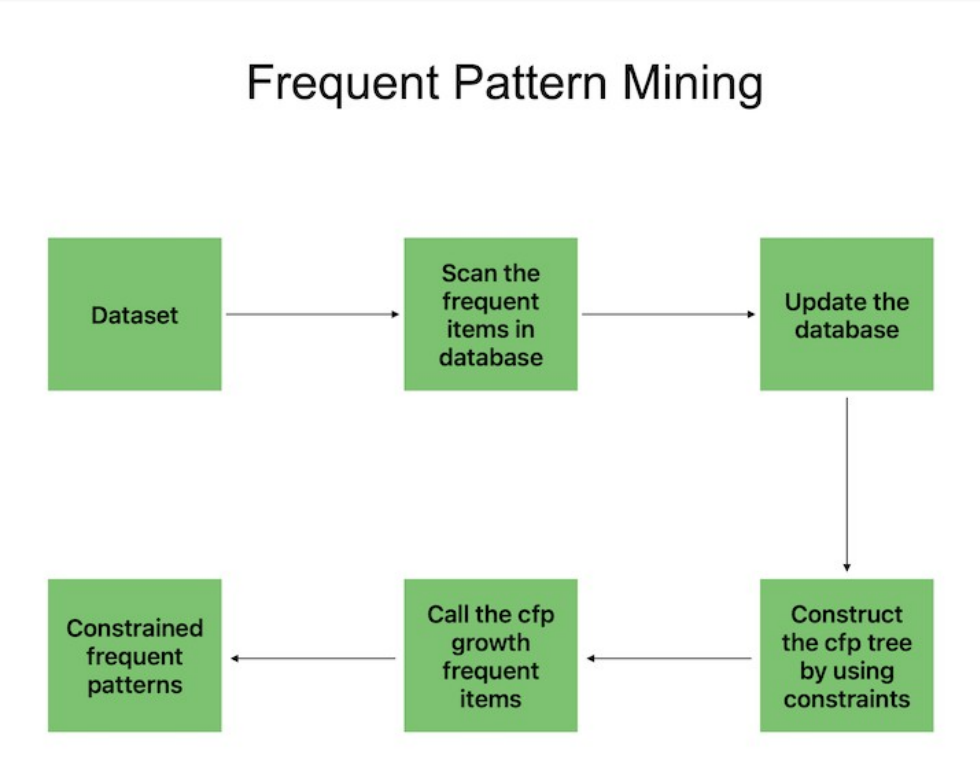
***Discrimination*:** data discrimination is a method useful for the derivation of descriptions of data classes or concepts. Basically, discrimination is used to compare general features of target class data objects with the general features of objects from one set of contrasting classes. The target and contrasting classes are user-specified; corresponding data objects are attached thru database queries. The methods are similar to characterization. Ex: Sales of brand A car go up 20% in a year vs. sales of brand B which have gone down 30% -- what are the general features of these cars? Could any one (or more than one) of these features have contributed to the difference in sales.

Mining Frequent Pattern:

Finding recurrent patterns or item sets in huge datasets is the goal of frequent pattern mining, a crucial data mining approach. It looks for groups of objects that regularly appear together in order to expose

underlying relationships and interdependence. Market basket analysis, web usage mining, and bioinformatics are a few areas where this method is important.

The patterns that fulfill the required support criteria are successfully identified through this iterative approach.



Market basket analysis

This is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies, and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Market basket analysis mainly works with the ASSOCIATION RULE {IF} -> {THEN}.

- **IF** means **Antecedent:** An antecedent is an item found within the data
- **THEN** means **Consequent:** A consequent is an item found in combination with the antecedent.

SUPPORT: It is been calculated with the number of transactions divided by the total number of transactions made,

support(pen) = transactions related to pen/total transactions

i.e support -> 500/5000=10 percent

CONFIDENCE: It is been calculated for whether the product sales are popular on individual sales or through combined sales. That is calculated with combined transactions/individual transactions.

Confidence = combine transactions/individual transactions

i.e confidence-> 1000/500=20 percent

LIFT: Lift is calculated for knowing the ratio for the sales.

Lift-> 20/10=2

Frequent Itemset, Closed Itemset and Association Rules:

1. Frequent Itemset

A frequent item set is a set of items that occur together frequently in a dataset. The frequency of an item set is measured by the support count, which is the number of transactions or records in the dataset that contain the item set.

2. Closed Itemset

A frequent itemset is *closed*, when no (immediate) superset has the same support.

Note: Every maximal frequent itemset is closed, but not every closed itemset is maximal.

If an itemset is not closed, we can look at the next larger closed itemset instead – support does not change.

we can recover all frequent itemset *and their support* from the closed frequent itemset.

3. Association Rules

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis. Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

Frequent Itemset Mining Method:

Frequent item sets, also known as association rules, are a fundamental concept in association rule mining, which is a technique used in data mining to discover relationships between items in a dataset. The goal of association rule mining is to identify relationships between items in a dataset that occur frequently together.

Apriori Algorithm:

Apriori algorithm is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.
To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

Apriori Property

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that

Before we start understanding the algorithm, go through some definitions which are explained in my previous post.

Consider the following dataset and we will find frequent itemsets and generate association rules for them.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2
minimum confidence is 60%

Step-1: K=1
(I) Create a table containing support count of each item present in dataset – Called **C1(candidate set)**

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

(II) compare candidate set item's support count with minimum support count(here min_support=2 if support_count of candidate set items is less than min_support then remove those items). This gives us itemset L1.

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Step-2: K=2

- Generate candidate set C2 using L1 (this is called join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common.
- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.(Example subset of{I1, I2} are {I1}, {I2} they are frequent.Check for each itemset)

- Now find support count of these itemsets by searching in dataset.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

(II) compare candidate (C2) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L2.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Step-3:

- Generate candidate set C3 using L2 (join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common. So here, for L2, first element should match.
  So itemset generated by joining L2 is {I1, I2, I3}{I1, I2, I5} {I1, I3, i5}{I2, I3, I4}{I2, I4, I5}{I2, I3, I5}
- Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.(Here subset of {I1, I2, I3} are {I1, I2},{I2, I3},{I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)
- find support count of these remaining itemset by searching in dataset.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

(II) Compare candidate (C3) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L3.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Step-4:

- Generate candidate set C4 using L3 (join step). Condition of joining L_{k-1} and L_{k-1} (K=4) is that, they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.
- Check all subsets of these itemsets are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So no itemset in C4
- We stop here because no frequent itemsets are found further

Generating association rules from frequent itemsets

Association Rules find all sets of items (itemsets) that have ***support*** greater than the minimum su
then using the large itemsets to generate the desired rules that have ***confidence*** greater than the m
confidence. The ***lift*** of a rule is the ratio of the observed support to that expected if X and Y were
independent. A typical and widely used example of association rules application is market baske
analysis.

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

AIS Algorithm

1. Candidate itemsets are generated and counted on-the-fly as the database is scanned.
2. For each transaction, it is determined which of the large itemsets of the previous pass are c in this transaction.
3. New candidate itemsets are generated by extending these large itemsets with other items in transaction.

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

The disadvantage of the AIS algorithm is that it results in unnecessarily generating and counting t candidate itemsets that turn out to be small.

Frequent Pattern Growth (FP-Growth) Algorithm:

The two primary drawbacks of the Apriori Algorithm are:

1. At each step, candidate sets have to be built.
2. To build the candidate sets, the algorithm has to repeatedly scan the database.

These two properties inevitably make the algorithm slower. To overcome these redundant steps, association-rule mining algorithm was developed named Frequent Pattern Growth Algorithm. It overcomes the disadvantages of the Apriori algorithm by storing all the transactions in a Trie Da Structure. Consider the following data:-

The above-given data is a hypothetical dataset of transactions with each letter representing an it frequency of each individual item is computed:-

Let the minimum support be 3. A **Frequent Pattern set** is built which will contain all the eleme whose frequency is greater than or equal to the minimum support. These elements are stored in descending order of their respective frequencies. After insertion of the relevant items, the set L like this:-

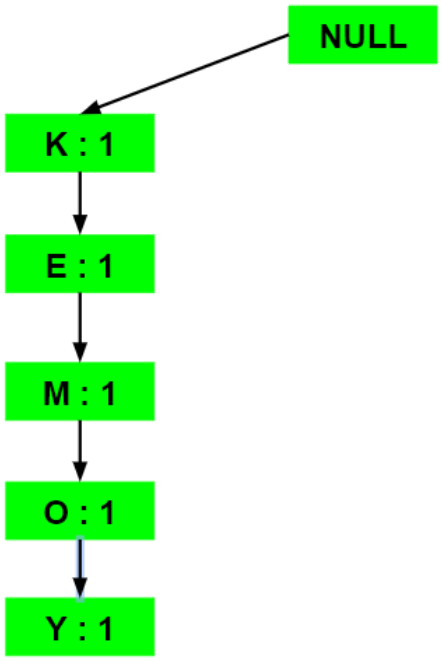
L = {K : 5, E : 4, M : 3, O : 4, Y : 3}

Now, for each transaction, the respective **Ordered-Item set** is built. It is done by iterating the Frequent Pattern set and checking if the current item is contained in the transaction in question. If the current item is contained, the item is inserted in the Ordered-Item set for the current transaction. The following table is built for all the transactions:

Now, all the Ordered-Item sets are inserted into a Trie Data Structure.

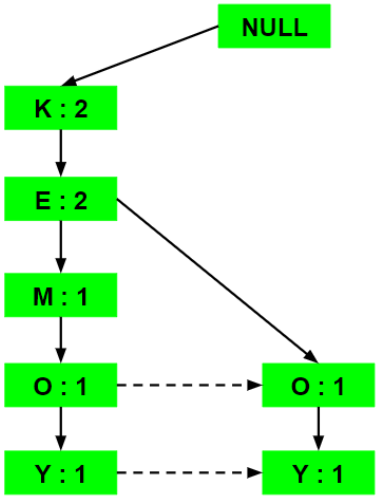
a) **Inserting the set {K, E, M, O, Y}:**

Here, all the items are simply linked one after the other in the order of occurrence in the set and initialize the support count for each item as 1.



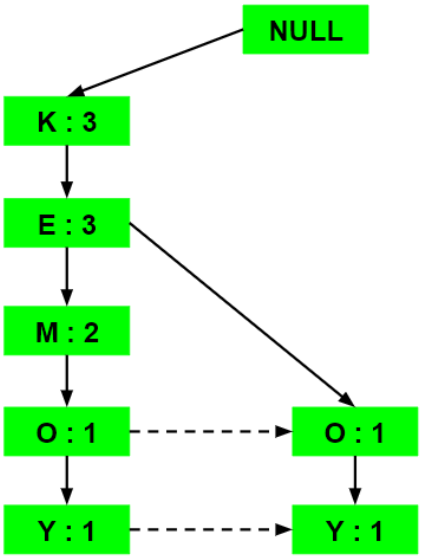
b) **Inserting the set {K, E, O, Y}:**

Till the insertion of the elements K and E, simply the support count is increased by 1. On inserting O we can see that there is no direct link between E and O, therefore a new node for the item O is initialized with the support count as 1 and item E is linked to this new node. On inserting Y, we first initialize a new node for the item Y with support count as 1 and link the new node of O with the new node of Y.



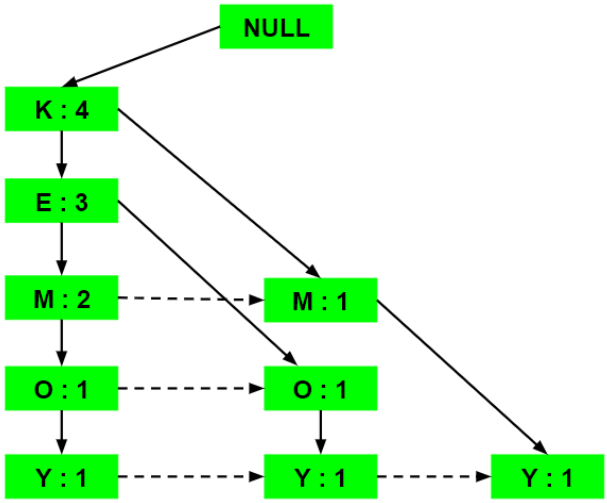
c) **Inserting the set {K, E, M}:**

Here simply the support count of each element is increased by 1.



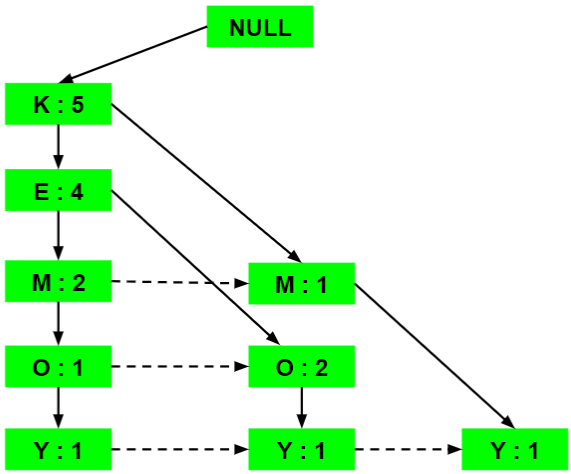
d) **Inserting the set {K, M, Y}:**

Similar to step b), first the support count of K is increased, then new nodes for M and Y are init
and linked accordingly.



e) **Inserting the set {K, E, O}:**

Here simply the support counts of the respective elements are increased. Note that the support c
the new node of item O is increased.



Now, for each item, the **Conditional Pattern Base** is computed which is path labels of all the p which lead to any node of the given item in the frequent-pattern tree. Note that the items in the b table are arranged in the ascending order of their frequencies.

Items	Conditional Pattern Base
Y	{{K,E,M,O : 1}, {K,E,O : 1}, {K,M : 1}}
O	{{K,E,M : 1}, {K,E : 2}}
M	{{K,E : 2}, {K : 1}}
E	{K : 4}
K	

Now for each item, the **Conditional Frequent Pattern Tree is built.** It is done by taking the se elements that is common in all the paths in the Conditional Pattern Base of that item and calcula support count by summing the support counts of all the paths in the Conditional Pattern Base.

Items	Conditional Pattern Base	Conditional Frequent Pattern Tree
Y	{{K,E,M,O : 1}, {K,E,O : 1}, {K,M : 1}}	{K : 3}
O	{{K,E,M : 1}, {K,E : 2}}	{K,E : 3}
M	{{K,E : 2}, {K : 1}}	{K : 3}
E	{K : 4}	{K : 4}
K		

From the Conditional Frequent Pattern tree, the **Frequent Pattern rules** are generated by pairin items of the Conditional Frequent Pattern Tree set to the corresponding to the item as given in th

table.

Items	Frequent Pattern Generated
Y	{<K,Y : 3>}
O	{<K,O : 3>, <E,O : 3>, <E,K,O : 3>}
M	{<K,M : 3>}
E	{<E,K : 4>}
K	

For each row, two types of association rules can be inferred for example for the first row which the element, the rules K -> Y and Y -> K can be inferred. To determine the valid rule, the confi both the rules is calculated and the one with confidence greater than or equal to the minimum confidence value is retained.