Comparative Analysis of Data extraction Association rule Algorithms

Dr. Monika Rathore
Asstt. Professor IIIM, Jaipur

Ms. Yogita Sharma
Asstt. Professor IIIM, Jaipur

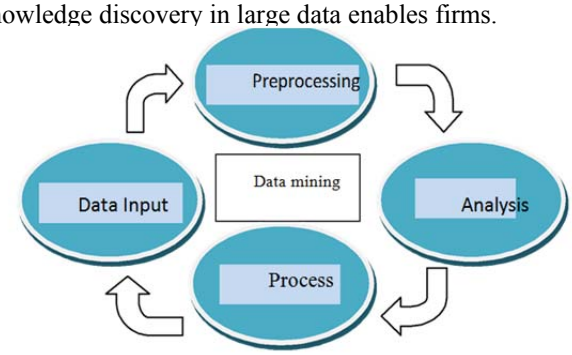
***Abstract*— In every field there is huge amounts of data is present but extracting useful data from that is difficult. For this reason the concept of data mining is developed. Many industries are dependent on data mining for making important decisions. Depending on the use of data mining result the tasks are being discovered. Data mining, also known as Knowledge Discovery in Databases (KDD), which consist of two types of languages approaches. There are so many techniques and tools available, which are helpful in overcoming the issues of data mining. Association and WEKA are used for the same in this. Market basket analysis is a very useful technique for finding out interrelated items in consumer shopping baskets. By this we can find out how different products in a store are related with each other. Association rules mining from transactional data gives us valuable information about how one product is dependent on other and how it affect its purchase. In such type of activity these information can be used for making decisions related to it. To find association rules we use two algorithms. (FP growth and Apriori algorithm).This paper contains the analysis of these algorithms (FP growth and Apriori algorithm) on different parameters. This paper contains the analysis on data collected from Jaipur Retail Store.**

***Keywords* — Apriori Algorithm, Association Rules, Data Mining, FP Growth, Frequent Item Sets, Market Basket Analysis**

I. INTRODUCTION

Data mining is the process of extracting hidden information or knowledge from large amount of data [1]. Data mining has been defined as the significant way of extracting previously unknown information from data. It is a powerful technology that can be best defined as the automated process of extracting useful knowledge and information including, patterns, associations [2].Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. It is the perception of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details [3]. Traditional data analysis methods usually involve manual work and perception of data that is slow, expensive and highly subjective. Useful information from the large databases has been extracted in the form of the association rules. There are many algorithms have been developed to extract the association rules from the large databases [4]. Apriori algorithm and fp-growth is the most popular algorithm to extract the association rules from the databases [5].

To implement the Apriori algorithm, there are many tools available in the market. WEKA is an open source software tool for implementing machine learning algorithms [6]. In data mining we can easily find out how different products in a retail shop assortment interrelate. Mining association rules from transactional data will provide us with valuable information about interrelated and co-purchases of products. Data Mining called as knowledge discovery in large data enables firms.



II. DATA MINING TASKS

The data mining tasks are of different types depending on the use of data mining result these tasks are classified as [7]:

1) **Exploratory Data Analysis**:
It is simply exploring the data without any clear ideas of what we are looking for. This technique is interactive and visual.

2) **Descriptive Modeling:**
It describe all the data, it includes models for overall probability distribution of the data.

3) **Predictive Modeling:**
This model permits the value of one variable to be predicted from the known values of other variables.

4) **Discovering Patterns and Rules:**
It concern with pattern detection, the main function is finding fraudulent behavior.

5) **Retrieval by Content:**
It is finding pattern i.e. to the pattern of interest from the data set.

III. THE KNOWLEDGE DISCOVERY PROCESS

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contains [8]

1) **Data Integration:** In this step data are collected and integrated from the different and multiple sources.

2) **Data Selection:**
In this step relevant data are retrieved from the database

3) **Data Cleaning:**
Data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data that's why the techniques are applied on them.

4) **Data Transformation:**
After cleaning, the data is transformed into correct form for mining
5) **Data Mining:**
At this stage data mining techniques are applied on data in order to find out the patterns
6) **Pattern Evaluation and Knowledge Presentation:**
This step the data patterns extracted from mining process are evaluated or tested.
7) **Decisions / Use of Discovered Knowledge:**
This step helps user to make use of the knowledge acquired to take better decisions

V. LANGUAGE APPROACHES

1) **Supervised Learning**
In supervised learning the variables under investigation can be divided into two groups: explanatory variables and dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables that is done in regression analysis. To proceed with directed data mining techniques the values of the dependent variable must be known.
2) **Unsupervised Learning:**
In unsupervised learning, all the variables are treated in same way; there is no distinction between dependent and explanatory variables. In contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The difference between unsupervised learning and supervised learning is the same that distinguishes analysis from cluster analysis. Supervised learning requires, target variable that should be well defined and sufficient number of its values are given. And in Unsupervised learning typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown

VI. ISSUES IN DATA MINING

Data mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real world databases. The main challenges to the data mining and the corresponding considerations in designing the algorithms are as follows:

- Massive datasets and high dimensionality.
- Over fitting and assessing the statistical significance.
- Understandability of patterns.
- Nonstandard incomplete data and data integration.
- Mixed changing and redundant data.

VII. DATA MINING TECHNIQUES

1) **Descriptive** approach includes models for overall probability distribution of the data, partitioning of whole data into
groups and models describing the relationships between the variables.
2) **Predictive** approach permits the value of one attribute/variable is too predicted from the known values of other attribute/variable.

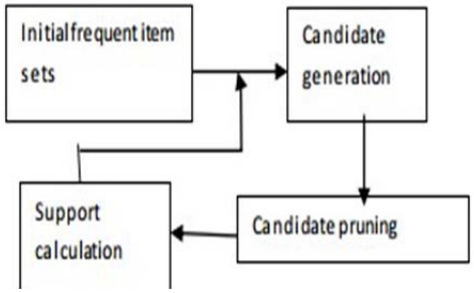
VII. TOOL USED FOR DATA MINING: WEKA(WAIKA TO ENVIORMENT FOR KNOWLEDGE ANALSIS)

Weka is an open source software tool. It is used to implementing machine learning algorithm. It is a collection of tools which is used for practice data mining technique. In this the data mining approaches are directly applied on the data sets. In this we can easily perform any of the technique for analyzing the data. There are following tools available in Weka.
1) **Explored** is used for explore and extract the dataset on which the operation has to be performed.
2) **Experimenter** is used for performing experiment on the dataset.
3) **Knowledge Flow** provides same functionality as Explorer but it is advance by taking also drag and drop interface.
4) **Simple GUI** provides simple command line interface that provide direct execution of Weka commands.
In Weka, basic implementation had done on the ARFF (Attribute-Relation File Format) files.
Firstly excel file is made then converted into ARFF file An ARFF file is an ASCII text [9]. ARFF files have two distinct sections.
1) **Header part** @RELATION, @ATTRIBUTE and @DATA declarations are case insensitive. This contains the name of the relation, a list of the attributes (the columns in the data), and their types.
2) **Data part** The @DATA declaration is a single line denoting the start of the data segment in the file.
ASSOCIATION RULE APPLIED
Association rules have if / then statement. They help to disclose the relation that is not directly related to each other in a relational database. Example- IF a customer buys a dozen eggs we can ensure that he also purchase milk.80%. Association will has 2 parts
Antecedent(if)
Consequent items (then)
An antecedent is an item which found in the data .A consequent is an item which is the combination of the antecedents. Using the criteria support in confidence to identify most important relationship. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. In the business analysis these information is very useful for understanding customers buying behavior. in the form of interdependency and co- occurrence of products.

IX. MARKET BASKET APPROACH

A. Apriori Algorithm

The Apriori algorithm is divided into 3 sections as-



Initial frequent item sets are fed into the system, and candidate generation, candidate pruning, and candidate support is executed in turn. The support information is fed back into the candidate generator and the cycle continues until the final candidate set is determined.

Table 1

ID	Item Sets
1	{ Bread, Butter}
2	{ Jam, Dairy, Canned Foods}
3	{ Jam, Breads, Butter, Paper_goods}
4	{Canned Foods, Dairy, Bread, Butter}

Consider in Table 1, the following rule can be extracted from the database is shown {Bread} → {Butter}

Binary database

(A)

ID	Bread	Butter	Jam	Juices	Paper_goods	Dairy	Canned Foods
1	1	1	0	1	0	1	1
2	0	0	1	1	1	0	1
3	1	0	1	0	1	0	1
4	1	1	1	0	1	0	0
5	0	1	0	1	1	1	0

Vertical Database

(B)

	Bread	Butter	Jam	Juices	Paper_goods	Dairy	Canned Foods
Item set	1	1	2	1	2	1	1
	3	4	3	2	3	5	2
	4	5	4	5	4		3
					5		

Transaction Database

(C)

ID	Item Set
	Bread,Butter,Juices, Dairy products, Canned Foods
2	Jam,Juices,Paper Goods,Canned Goods
	Bread,Jam,Paper Goods,Canned goods
4	Bread,Butter,Jam,Paper Goods
5	Butter,Juices,paper Goods,Dairy products

The support of rule X and Y are called antecedent (LHS: left hand side) and consequent (RHS: right hand side) of the rule. Consider in table 1, the set of items is I = {Breads, Butter, Juices, Dairy, and Canned Foods}. {Breads, Butter}{Juices} meaning

That if bread and butter is bought, customers also buy Juices.[10] The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X.

Support:

$$\text{Support}(X \rightarrow Y) = \frac{\text{Probability}(X \cup Y)}{\text{Total number of transactions}}$$

Lift:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Probability}(X \cap Y)}{\text{Probability}(X)\text{Probability}(Y)}$$

Confidence

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Probability}(X \cap Y)}{\text{Number of transaction}(X)}$$

Consider from table1: If the rule {Jam, Bread} → {Butter}. Therefore, the support count for {Jam, Bread, Butter} is 3 and the total number of transactions is 5, the rule's support is 3/5 = 0.6.While the rule's confidence is obtained by dividing the support count for {Jam, Bread, Butter} by the support count for {Bread, Butter}. Since there 3 transactions that contain Breads and Butter, the confidence for this rule is 2/3 = 0.67. The algorithm is as follows:

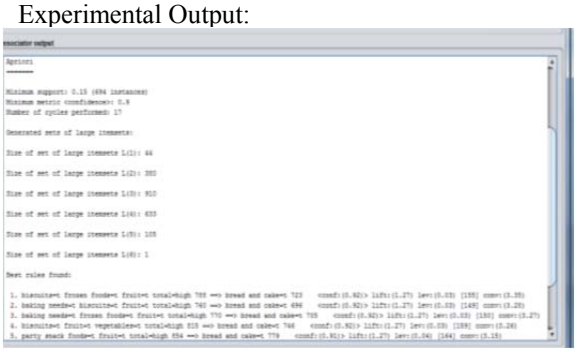
1) Join Step: Ck is generated by joining Lk-1 with itself.
2) Prune Step: Any (k-1) item set that is not frequent can't be a subset of a frequent k-item set.
3) Pseudo-Code:
4) Ck : Candidate item set of size k
5) Lk : Frequent item set of size k
6) L1 = {frequent items}
7) For(k=1;Lk!= ϕ;k++) do begin
8) Ck+1=candidates generated from Lk;
9) For each transaction t in database do
10) Increment the count of all candidates in Ck+1 that are
11) Contained in t
12) Lk+1=candidates in Ck+1with min_support
13) End
14) Return ∪k Lk;

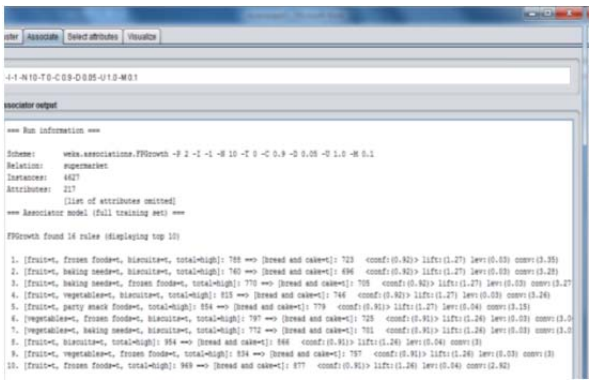
2. Processor for the Experiment:

a. Open the file in preprocess segment: When the chosen ARFF file is open in the Preprocess section, there were two boxes :
   Left most box was the 'Attributes'- which showed the attributes name of that particular ARFF file which was currently opened and provide the checkboxes to select the attributes and there was a button to remove the checks.
   Right side box was 'selected attributes'- which upper portion showed the value items of the selected attributes with the description of their labels and their counts and other information. And the lower portion showed the graphical representation of the attributes with itself.

b. Open the file in Associate segment: When the chosen ARFF file is open in the Associate section, there were an output box appeared as :

3. Experimental Output:





CONCLUSION

The FP-growth is algorithm which overcomes the major problems association rule with Apriori algorithm. It follows a divide and conquers strategy. The original dataset is transformed into a tree well known as FP-tree, which holds all the information regarding frequent items. All frequent item sets can be mined from the tree directly via the FP-growth algorithm. The FP-growth algorithm to determine the frequent item sets and the create association rules algorithm to generate association rules based on the frequent item sets discovered. This algorithm calculates all frequent item sets, building a FP-tree structure from database transactions. A major advantage of FP-growth algorithm compared to Apriori algorithm is that it uses only 2 data scans and is therefore often applicable even on large data sets.

// Initial Call: R←FP-tree(D), P← Æ , F← Æ
1) FP-growth (R, P, F, minsup):
2) Remove infrequent items from R
3) if IsPATH (R) then
4) foreach Y Í R do
5) X←P È Y
6) Sup(X) ←minxєY{cnt(x)}
7) F←FU{(X, sup(X))}
8) else
9) foreach i Î R in increasing order of sup(i) do
10) X←PU{i}
11) sup(X) ←sup(i)
12) F←FU {(X, sup(X))}
13) RX← Æ //projected FP-tree for X
14) foreach path Î PATH FROM ROOT(i) do
15) cnt(i) ← count of i in path
16) Insert path, excluding I, into FP-tree RX with count cnt(i)
17) If Rx ≠ Æ then FP-growth (Rx, X, F, minsup)

REFERENCES

[1] S. Prakash Kumar and 2K.S. Ramaswami, Journal of Computer Science 7 (11): 1652-1658, 2011 ISSN 1549-3636 © 2011 Science Publications

[2] Ghadeer S. Abu-Oda and Alaa M. El-Halees, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015 [3] Ms. Laxmi S. Patil , Prof. M. S. Bewoor and Dr. S. H.Patil, Vol. 2, Issue 3, May-Jun 2012

[4] Ajay Kumar Shrivastava, R. N. Panda, Vol. 1, Issue 1, January 2014

[5] Ajay Kumar Shrivastava, R. N. Panda, Vol. 1, Issue 1, January 2014

[6] MACHINE LEARNING GROUP , http://www.cs.waikato.ac.nz/ml/index.html

[7] APPLICATION OF DATA MINING ,http://shodhganga.inflibnet.ac.in/bitstream/10603/24256/10/10_chapter5.pdf

[8] Ajay Kumar Shrivastava, KIET International Journal of Intelligent Computing and Informatics, Vol. 1, Issue 1, January 2014

[9] Kavitha Venkatachari , MARKET BASKET ANALYSIS