International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 01 || January 2017 || PP. 15-20

MARKET BASKET ANALYSIS FOR DATA MINING: concepts and techniques

P. ARPITHA
(MCA ,MTECH,(PHD))
Associate professor ,HOD(MCA)
Aurora PG College
Moosarambagh

Abstract: Data mining (DM), also called Knowledge-Discovery in Databases (KDD), is the process of automatically searching large volumes of data for patterns using specific DM technique. The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets. The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Association rule mining represents a data mining technique and its goal is to find interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits. For example, the discovery of interesting association relationships among huge amounts of business transaction records can help catalog design, cross marketing, loss leader analysis, and other business decision making processes. If/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" using confidence and support factors.
Key words: data mining, association rule, market basket analysis, knowledge discovery, support and confidence factors.

Introduction:

Many business enterprises accumulate large quantities of data from their day- to – day operations. For example, huge amounts of customers purchase data are collected daily at the checkout counters of grocery stores. **Market Basket Analysis** (Association Analysis) is a mathematical modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy another group of items. It is used to analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data.

Literature survey:

1. Data Mining

It is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data.

The information or knowledge extracted so can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Applications :
Data mining is highly useful in the following domains:

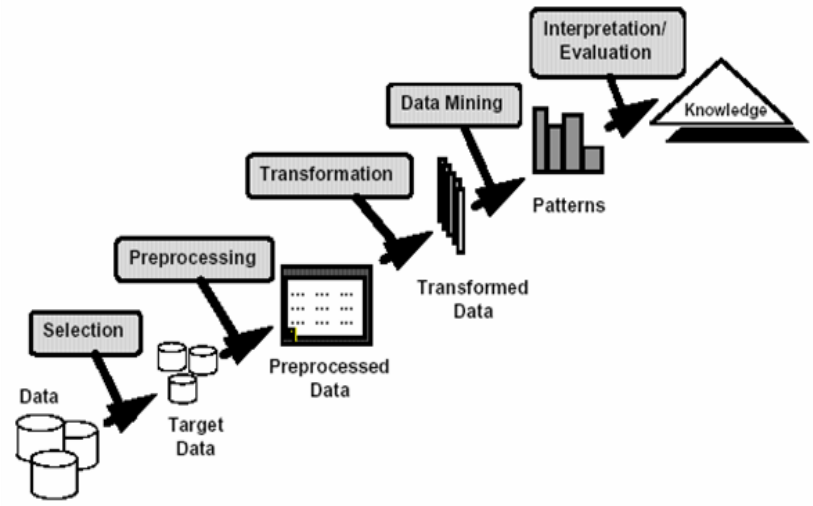
- **Market Analysis and Management**
- **Corporate Analysis & Risk Management**
- **Fraud Detection**

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

Market Analysis and Management:
Listed below are the various fields of market where data mining is used:

- **Customer Profiling** - Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** - Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** - Data mining performs Association/correlations between product sales.
- **Target Marketing** - Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc
- **Determining Customer purchasing pattern** - Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** - Data mining provides us various multidimensional summary reports.

KNOWLEDGE DISCOVERY:



"Data mining (DM), also called Knowledge-Discovery in Databases (KDD), is the process of automatically searching large volumes of data for patterns using specific DM technique."

Data Mining Tasks :
1. Classification: learning a function that maps an item into one of a set of predefined classes
2. Regression: learning a function that maps an item to a real value
3. Clustering: Identify a set of groups of similar items
4. Dependencies and associations: Identify significant dependencies between data attributes
5. Summarization: find a compact description of the dataset or a subset of the dataset.

2. Association rules:
Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness

- Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository.
- An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." .
  - Proposed by Agrawal et al in 1993.
  - It is an important data mining model studied extensively by the database and data mining community.
  - Assume all data are categorical.

- No good algorithm for numeric data.
- Initially used for Market Basket Analysis to find how items purchased by customers are related.
- Bread → Milk [sup = 5%, conf = 100%]

What Is Association Rule Mining?

Association rule mining

- Finding frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases.
- Understand customer buying habits by finding associations and correlations between the different items that customers place in their "shopping basket".

Applications:

- Basket data analysis, cross
- marketing, catalog design, loss
- leader analysis, web log analysis, fraud detection

What Is Association Rule Mining?

Rule form:

Antecedent → consequent [support, confidence]

(**Support and confidence are user defined measures of interestingness)**

How can Association Rules be used?
Let the rule discovered be

{Potato Chips}→{Bagels,...}

- **Potato chips as consequent** => Can be used to determine what should be done to boost its sales
- **Bagels in the antecedent** => Can be used to see which products would be affected if the store discontinues selling bagels
- **Bagels in antecedent and Potato chips in the consequent** => Can be used to see what products should be sold with Bagels to promote sale of Potato Chips

3. Market Basket Analysis:

An Overview Market basket analysis (MBA) is a data mining technique to discover associations between datasets. These associations can be represented in form of association rules. The formal statement of problem[7] can be stated as : Let I is a set of items {i1,i2,….,im}.Let D is a set of transactions such that T I. Each transaction is uniquely identified with with an identifier called TID. The method can be stated as if there are two subsets of product items X and Y then an association rule is in the form of X→Y where X I and Y I. It implies that if a customer purchases X, then he or she also purchases Y. Two measures which reflect certainty of discovered association rules are support and confidence. Support measures how many times the transactional record in database contain both X and Y. Confidence measures the accuracy of rule. As an example, the information that customers who purchase computers also tend to buy printer at the same time is represented in Association Rule below. Computer = Printer Support = 20%, Confidence = 80% Association rules are considered useful if they satisfy both a type equation here minimum support threshold and a minimum confidence threshold that can be set by users or domain consultants. Figure1 shows a typical Market basket analysis. This is a perfect example for illustrating association rule mining. This market basket analysis system will help the managers to understand about the sets of items are customers likely to purchase. This analysis may be carried out on all the retail stores data of customer transactions. These results will guide them to plan marketing or advertising approach. For example, market basket analysis will also help managers to propose new way of arrangement in store layouts. Based on this analysis, items that are regularly purchased together can be placed in close proximity with the purpose of further promote the sale of such items together. If consumers who purchase computers also A Survey on Association Rule Mining in Market Basket Analysis 41

Example 1:

Customer	Items
1	Orange juice ,soda
2	Milk, Orange juice, window cleaner
3	Orange juice, Detergent
4	Orange juice, Detergent ,soda
5	Window cleaner, soda

	Orange juice	Window cleaner	Milk	Soda	Detergent
Orange juice	4	1	1	2	1
Window cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

- **Orange juice** and **soda** are more likely to be purchased together than any other two items
- **Detergent** is never purchased **with window cleaner or milk**.
- **Milk** is never purchased with **soda or detergent**.

Basic Concepts:
Given:
(1) Database of transactions,
(2) Each transaction is a list of items purchased by aν customer in a visit

Find:
All rules that correlate the presence of one set of items (itemset) with that of another set of items E.g., 35% of people who buys salmon also buys cheese

Rule Basic Measures:
A⇒B [s, c]

A)Support: denotes **the frequency of the rule within transactions**. A high value means that the rule involves a great part of database.

support(A⇒ B) [s, c]) = P(A∪B)

B)Confidence : denotes the **percentage of transactions containing A which also contain B**. It is an estimation of conditioned probability.

Confidence (A⇒B([s, c]) = P(A|B) = Supp(A,B)/Sup(A).

Example1 2:

Tr1	Shoes, Socks, Tie, Belt
Tr2	Shoes, Socks, Tie, Belt, Shirt, Hat
Tr3	Shoes, Tie
Tr4	Shoes, Socks ,Belt

Transaction	Shoe	Socks	tie	belt	Shirt	Scarf	Hat
1	1	1	1	1	0	0	0
2	1	1	1	1	1	0	1
3	1	0	1	0	0	0	0
4	1	1	0	1	0	0	0

Socks ⇒ Tie **Support is** **50% (2/4)**
Confidence is **66.67%(2/3)**

LIMITATIONS:

The discovery of purchasing patterns in these multiple stores changes with time as well as location. In this multiple store chain basic association rules are not effective. The problem of finding association rules using market basket analysis can be solved using the basic A**priori algorithm**

(4)Apriori algorithm:

It is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association_rules which highlight general trends in the database: this has applications in domains such as market_basket_analysis.

Example 3

Assume that a large supermarket tracks sales data by stock-keeping unit (SKU) for each item: each item, such as "butter" or "bread", is identified by a numerical SKU. The supermarket has a database of transactions where each transaction is a set of SKUs that were bought together.
Let the database of transactions consist of following itemsets:

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

We will use Apriori to determine the frequent item sets of this database. To do so, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the *support threshold*.

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. We obtain the following result.

Item	**Support**
{1}	3
{2}	6
{3}	4
{4}	5

All the itemsets of size 1 have a support of at least 3, so they are all frequent.
The next step is to generate a list of all pairs of the frequent items.

For example, regarding the pair {1,2}: the first table of Example 2 shows items 1 and 2 appearing together in three of the itemsets; therefore, we say item {1,2} has support of three.

Item	**Support**
{1,2}	3
{1,3}	1
{1,4}	2

{2,3}	3
{2,4}	4
{3,4}	3

The pairs {1,2}, {2,3}, {2,4}, and {3,4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {1,3} and {1,4} are not. Now, because {1,3} and {1,4} are not frequent, any larger set which contains {1,3} or {1,4} cannot be frequent. In this way, we can *prune* sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Item	Support
{2,3,4}	2

In the example, there are no frequent triplets -- {2,3,4} is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.
We have thus determined the frequent sets of items in the database, and illustrated how some items were not counted because one of their subsets was already known to be below the threshold.

Algorithm Apriori(*T*)

$C_1 \leftarrow$ init-pass(T);
$F_1 \leftarrow \{f \mid f \in C_1, f.\text{count}/n \geq minsup\}$; // n: no. of transactions in T
for ($k = 2$; $F_{k-1} \neq \emptyset$; k++) **do**
  $C_k \leftarrow$ candidate-gen(F_{k-1});
  for each transaction $t \in T$ **do**
    for each candidate $c \in C_k$ **do**
      if c is contained in t **then**
        $c.count$++;
    end
  end
 $F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$
end
return $F \leftarrow \bigcup_k F_k$;

Conclusion:

We have shown how Market basket analysis using association rules works in determining the customer buying patterns. Data mining refers to extracting knowledge from large amount of data. Market basket analysis is a data mining technique to discover associations between datasets. Association rule mining identifies relationship between a large set of data items. When large quantity of data is constantly obtained and stored in databases, several industries are becoming concerned in mining association rules from their databases. This method examines customer buying patterns by identifying associations among various items that customers place in their shopping baskets. The identification of such associations can assist retailers expand marketing strategies by gaining insight into which items are frequently purchased by customers. It is helpful to examine the customer purchasing behavior and assists in increasing the sales.

References:

[1]. Jiawei Han and Micheline Kamber, "Data Mining: Concepts andν Techniques", 2 edition (4 Jun 2006)
[2]. Gordon S. Linoff and Michael J. Berry, "Data Mining Techniques:ν For Marketing, Sales, and Customer Relationship Management", 3rd Edition edition (1 April 2011)
[3]. Vipin Kumar and Mahesh Joshi, "Tutorial on High Performanceν Data Mining ", 1999
[4]. Rakesh Agrawal, Rama krishnan Srikan, "Fast Algorithms forν Mining Association Rules", Proc VLDB, 1994
[5]. Rakesh Agrawal and Ramakrishna Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.