A PREDICTION OF CUSTOMER BEHAVIOR USING APRIORI ALGORITHM WITH ASSOCIATION RULE MINING WITH REAL TIME DATA

P.Nithya M.Sc.,M.Phil., Assistant Professor,

Department of Computer Science & Information Technology,

Nadar Saraswathi College of Arts & Science, Theni

ABSTRACT

Data mining is the process of extracting useful information from the large amount of data stored in the database. Data mining tools and techniques helps to predict business trends those can occur in near future. Data mining is the procedure of mining knowledge from the data. The information or knowledge extracted so can be used for the following applications are Market Analysis, Fraud Detection, Customer Retention, Production Control, Science and Exploration. Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket. Association rule mining is an important technique to discover hidden relationships among items in the transaction. The goal of this paper is to experimentally evaluate an Apriori algorithm for predicting customer behavior. It is a classic algorithm used in data mining for learning association rule. It is very important for effective market basket analysis and its helps the customer in purchasing their items with more ease which increase the sales of the markets. A key concept in Apriori algorithm is the anti-montoicity of the support measure. It assumes that all subsets of a frequent item set must be frequent, similarly for any infrequent item set all its supersets must be infrequent too.

Frequent Pattern Mining is a very important undertaking in data mining. Apriori approach applied to generate frequent item set generally espouse candidate generation and pruning techniques for the satisfaction of the desired objective.

I. INTRODUCTION

Data mining is the process that uses a variety of data analysis tools to discover pattern and relationships in data that may used to make valid prediction. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems[4]. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

1.2 Association Rule Mining

Association rules mining is an important branch of knowledge mining research. They are used for finding frequent patterns and associations among sets of items in transactional databases, relational databases, and other information repositories [1]. An association rule is the relationship between two disjoint itemsets, X and Y. An association rule is of the form: - X -> Y X => Y: - When X occurs, Y also occurs. Given a set of items $I = \{I1, I2,\ldots,Im\}$ and a database of transactions $D = \{t1, t2,\ldots\ldots,tn\}$ where $ti = \{Ii1, Ii2\ldots.\ Iik\}$ and $Iij \in I$, an association rule is an implication of the form X=>Y where $X, Y \subseteq I$ are sets of items called item sets and $X \cap Y = \emptyset$. Association rule mining has been used in a retailing where discovering of purchase patterns between products is very useful for decision making.

1.2.1 Frequent Itemset

Frequent pattern analysis allows a researcher to systematically identify patterns that emerge from database. Frequent pattern mining comprises frequent item set mining and association rule induction[3].

1.2.2 Market Basket Analysis

Market Basket Analysis is a knowledge mining technique that is widely used to identify consumer patterns such that if customer buys certain group of items then customers are likely to buy another group of items[6]. Market basket analysis is an important component in retail organizations. It is a very useful technique for finding out co-occurrence of items in consumer shopping baskets.

1.2.3 Support

It is the measure of how often the collections of items in an association occur together as percentage of all transactions. Support(s) for an association rule *X* =>*Y* is the percentage of transactions in the database that contains *X* U *Y*. Every association rule has support. The rule that has very low support may occur simply by chance.

1.2.4 Confidence

Confidence for an association rule X=>Y is the ratio of the number of transaction that contain both antecedent and consequent to the number of transaction that contain only antecedent. A rule with low confidence is not meaningful. Confidence (α) for an association rule x=>Y is the ratio of number of transactions that contains X U Y to the number of transactions that contains X.

1.2.5 Minimum Threshold Values

The strength of an association rule can be measured in terms of its support and confidence. The rules derived from itemsets with high support and high confidence. The number of association rules that can be derived from a dataset are large. Interesting association rules are those whose support and confidence are greater than minimum support and minimum confidence[2]. The number of association rules discovered is affected by a user's decision concerning the minimum support threshold and minimum confidence threshold. Threshold values can be set by user or domain export. It may be decided on the basis of number of transactions in database. Association rule need to satisfy a user specified minimum support and user specified minimum confidence at the same time. support and confidence values occur between 0% and 100%.

II. METHODOLOGY

2.1 APRIORI ALGORITHM

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.

Basic Concentrations:

1. **Apriori Property**: **a.** Any subset of frequent itemset must be frequent. **b**. An itemset is called a candidate itemset if all of its subsets are known to be frequent.
2. **Join Operation**: To find Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself.
3. **Prune step:** Remove those candidates in *Ck* that cannot be frequent.

2.2 ALGORITHM STEPS:

1. **Find all frequent itemsets:** This step finds all frequent itemsets using minimum support count.
2. **Generating Association Rules from Frequent Itemsets:** The frequent itemsets found in step (1) are used to generate association rules as: For each frequent itemset "I", generate all nonempty subsets of I. For every nonempty subset s of I, output the rule "s → (I-s)" if support_count (I) / support_count(s) >= minimum confidence threshold.

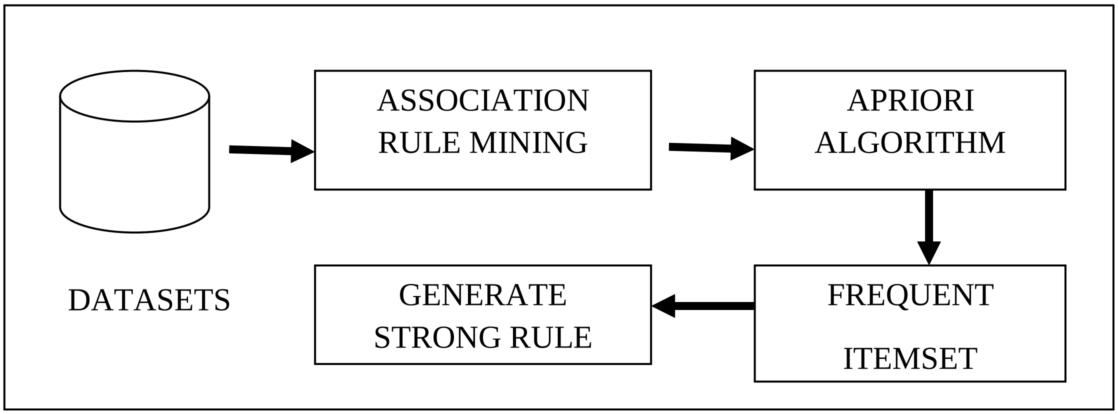

Figure 2.1

Definition 2.2.1: The association rules can be formally defined as:

- If the support of item-sets *X* is greater than or equal to minimum support threshold, *X* is called frequent item-sets.
- If the support of item-sets *X* is smaller than the minimum support threshold, then *X* is called infrequent item-sets[5].

Definition 2.2.2: The support of an item-set is the fraction of the rows of the database that contain all of the items in the item-set. Support indicates the frequencies of the occurring patterns.Sometimes it is called frequency. Support is simply a probability that a randomly chosen transaction t contains both item-sets A and B. Mathematically,

Support, $s(X \rightarrow Y) = \sigma (XUY)/N$

Definition 2.2.3: Confidence denotes the strength of implication in the rule. Sometimes it is called Accuracy. Confidence is simply a probability that an item-set B is purchased in a randomly Chosen transaction t given that the item-set A is purchased. Mathematically,

Confidence, $c(X \rightarrow Y) = \sigma (X\ UY)/ \sigma (X)$

Definition 2.2.4: Minimum-Support$=\dfrac{\text{No.of Transaction that contain both X\&Y}}{\text{Total no.of transaction}}$

Definition 2.2.5: Minimum-Confidence $= \dfrac{\text{No.of Transaction that contain X\&Y}}{\text{Transaction that contain only X}}$

The Apriori Algorithm

```
Initialize: k:= 1, C1 = all the 1 - itemsets;
read the database to count the support of C1 to determine Li.
L1:= {frequent 1-itemsets};
k := 2; // k represents the pass number //
while (k-1 ≠Null set) do
begin
Ck := gen_candidate_itemsets with the given $L_{k\text{-}1}$
prune(Ck)
for all transactions t   T do
Calculate the support values;
Lk:= All candidates in Ck with a minimum support;
```

```
k : = k + l ;
end
Candidate Generation:
gen_candiadate_itemsets with the given Lk-1 as follows:
Ck=ϕ
for all itemsets   do
for all itemsets   do
if l1[1]= l2 [1] ^ l1[2]= l2 [2] ^….. ^l1[k-1]<l2 [k-1] ^
then l1[1], l1[2],……, l1 [k-l], l2[k-l]
Ck = Ck {c}
Pruning
Prune(Ck)
for all c Ck
for all (k-1)- subsets d of c do
if d   Lk-1
then Ck = Ck - {c}
```

Figure 2.2: Apriori Algorithm

III. DATA SET DESCRIPTION

The apriori algorithms were developed for frequent item set generation and Association rule mining. After the development of this algorithm, it is tested by using the Customer dataset.

3.1 Customer Data Set Description

The Customer data set contains four attributes. First attribute is transaction ID (TID), which is unique and the remaining attributes are ITEMS. The Customer data set provides the details of the transaction made by customer. Table 3.1 displays the Customer Data set

TID	ITEM1	ITEM2	ITEM3
TXl	Bread	Butter	Milk
TX2	Ice-cream	Bread	Butter
TX3	Bread	Butter	Noodles
TX4	Bread	Noodles	Ice-cream
TX5	Butter	Milk	Bread
TX6	Bread	Noodles	Ice-cream
TX7	Milk	Butter	Bread
TX8	Ice-cream	Milk	Bread
TX9	Butter	Milk	Noodles
TX10	Noodles	Butter	Ice-cream

Table 3.1: Customer Data Set

IV. IMPLEMENTATION OF APRIORI ALGORITHM

A large supermarket tracks sales data by Stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together.

Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. By using the consumer database given in the table 3.1, The support of an association pattern is the percentage of task - relevant data transactions for which the pattern is true.

Support (A -> B) = P (A U B)

Support (A -> B) = no. of Tuple containing both A & B /Total no. of Tuples

Step 1: The first step of Apriori is to count the frequencies, called the supports, of each member item separately is shown in Table 4.1

Items	Support
Bread	0.8
Butter	0.7
Ice-Cream	0.5
Milk	0.5
Noodles	0.5

Table 4.1: Candidate 1

Step 2: The pruning step eliminates the item sets which are not found to be frequent (i.e. the item set is less than or equal to the minimum support) shown in Table 4.2

Level 1= {{Bread}, {Butter}, {Ice-cream}, {Milk}, {Noodles}}

Items	Support
Bread	0.8
Butter	0.7
Ice-Cream	0.5
Milk	0.5
Noodles	0.5

Table 4.2: Level 1

Step 3: To generate the itemset candidate 2 item set has minimum support, so do all subsets.

Candidate2= {{Bread, Butter}, {Bread, Ice- cream, {Bread, Milk}, {Bread, Noodles},
{Butter, Ice-cream}, {Butter, Milk}, {Butter, Noodles}, {Ice-cream, Milk},
{Ice-cream, Noodles}, {Milk, Noodles}}

Step 4: Read the combination of 2 items to count the support. The frequent 2-item set and their support counts is calculated that is shown in Table 4.3

Item-sets	Support
{Bread, Butter}	0.5
{Bread,Ice-cream}	0.4
{Bread, Milk}	0.4
{Bread, Noodles}	0.2
{Butter, ice-cream}	0.3
{Butter, Milk}	0.4
{Butter, Noodles}	0.3
{Ice-cream, Milk}	0.1
{Ice-cream,Noodls}	0.3
{Milk, Noodles }	0.3

Table 4.3: Candidate 2

Step 5: The pruning step eliminates the item sets which are not found to be frequent (i.e. the item set is less than or equal to the minimum support) that shown in Table 4.4

Level 2={{Bread, Butter} ,{ {Bread, Milk} ,{ {Bread, Noodles} ,{ {Bread, ice-cream} , {Butter, Milk} , {Butter, Noodles}, {Noodles, ice-cream } }

Item-sets	**Support**
{Bread, Butter}	0.5
{Bread, Milk}	0.4
{Bread, Noodles}	0.2
{Bread, ice-cream}	0.3
{Butter, Milk}	0.4
{Butter, Noodles}	0.3
{Noodles, ice-cream }	0.3

Table 4.4: Level 2

Step 6: To generate itemset candidate 3 item set has minimum support, so do all the subset that shown in Table 4.5

Candidate 3= {{Bread, Butter, Milk},{Bread, Butter, Noodles},{Bread, Butter ,Ice-cream} ,
{Bread, Milk,Noodles},{Bread, Milk,Ice-cream},{Bread,
Noodles,Ice-cream}, {Butter, Milk,Noodles}, }

item-sets	support
{Bread, Butter, Milk}	0.3
{Bread, Butter, Noodles}	0.1
{Bread, Milk, ice-cream}	0.1
{Bread, Butter, ice-cream}	0.0
{Butter, Milk, Noodles}	0.1
{Bread, Milk, Noodles}	0.0
{Noodles, ice-cream, Bread}	0.2

Table 4.5: Candidate 3

Step 7: The pruning step eliminates the item sets which are not found to be frequent (i.e. the item set is less than or equal to the minimum support) that shown in Table 4.6

Level 4= {{Bread, Butter, Milk}}

Item-sets	**Support**
{Bread, Butter, Milk}	0.3

Table 4.6: Level 3

The confidence of a rule A -> B, is the ratio of the number of occurrences of B given A, among all other occurrences given A. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern A -> B

Confidence (A -> B) = P (B |A) means the probability of B that all know A

Confidence (A-> B) = no. of Tuple containing both A & B /no. of Tuples containing A

Step 8: Here only one item-set which satisfy the minimum support value. So after three iteration, only one item-set filtered.

Frequent Item-set= {Bread, Butter, Milk}

Association rules for frequent item-sets

Rules	Confidence(percentage)
{Bread}->{Butter, Milk}	37
{Bread, Butter}->{Milk}	60
{Bread, Milk}->{Butter}	75
{Butter}->{Bread, Milk}	42
{Butter, Milk}->{Bread}	75
{Milk}->{Bread, Butter}	75

Table 4.7: Confidence of items

Step 9: If the minimum confidence threshold is 70 percentages then discovered rules are

{Bread, Milk}-> {Butter}
{Butter, Milk}-> {Bread}
{Milk}-> {Bread, Butter}

Because the confidence value of these rules are greater than minimum confidence threshold value which is 70 percent. So in the simple language if a customer buy Bread and Milk he is likely to buy Butter. A customer buy Butter and Milk is likely to Bread. A customer buy Milk is likely to buy Bread and Butter.

V. CONCLUSION

In this paper, Apriori algorithm is used to find frequent items in a given transaction of database. Apriori algorithm finds the tendency of a customer on the basis of frequently purchased itemset. The algorithm was tested on sample dataset.

Association rule have been used to extract the useful information from the large database. In this work customer behavior is analyzed using Apriori algorithm with association rule mining.

VI. REFERENCES

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 487-499, 1994.

[2] Anand H.S. and Vinodchandra S.S.,” Applying Correlation Threshold on Apriori Algorithm,” 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology

[3] Arpan Shah Pratik A. Patel, “A Collaborative Approach of Frequent Item Set Mining: A Survey,” International Journal of Computer Applications (0975 – 8887)

[4] Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan “A SURVEY ON DATA MINING TECHNIQUES,”International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

[5] Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, “An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction”, IEEE 2011.

[6] Loraine Charlet Annie M.C and Ashok Kumar D, ,” Market Basket Analysis for a Supermarket based on Frequent Itemset Mining , ”IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694- 0814.