International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

A Hybrid Approach to Analyse Customer Purchase Behaviour

Vishva Shah[1], Bhavin Sheth[2], Vipul Solanki[3], Sumedh Telang[4], Niti Desai[5]
[1, 2, 3, 4]*Student, Computer Engineering Department, Rajiv Gandhi Institute of Technology, Versova, Mumbai*
[5]*Professor, Computer Engineering Department, Rajiv Gandhi Institute of Technology, Versova, Mumbai*

Abstract: ***Analyzing purchase behavior of customers is one of the analytics techniques used to interpret the complex behavior and market trends based on sales. It is a data driven method to generate reports and draw conclusions about the behavior patterns of customers and present market. The process consists of calculating the behavior of a customer for a retail enterprise based on the history of transactions. The knowledge derived from this model helps the retailer to make better business decisions and build marketing and sales strategies that are more profitable.***
Predicting behavioral patterns is a Customer Relationship Management practice that a company uses to manage and analyze customer interactions and data throughout the customer lifecycle for enhancing the business relationship with customers. Machine learning algorithms Support Vector Machines and Apriori are implemented to achieve the outcome of analysis.
The proposed system uses R programming to analyze the data of a retail enterprise. R programming is a powerful statistical language. The unstructured raw data is pre-processed to undergo analysis. The cleaned data is operated upon by regression models and the analysis report is plotted in graph. This report can be used by the retailer to understand the sales trends and customer behavior. The analysis will be represented in an online shiny dashboard. In support of this predictive model, a customer-feedback module for product review makes the proposed system more interactive.
Keywords: ***Predictive analytics, machine learning, support vector machines, shiny, apriori, product review***

I. INTRODUCTION

The major intention behind customer relationship management is to identify customer interests and purchase pattern. History of transactions is very useful to conduct this analysis. The attributes that define the customer behavior pattern are transaction attributes such as Transaction ID, Product Name, Unit Price, and Quantity. These are analyzed using predictive modeling techniques and the report is presented in the form of graphs and statistical measures in R.
Two main modelling machine learning algorithms implemented are Support Vector Machine (SVM) for linear classification and Apriori algorithm for association rule generation. SVM is used to transform the input data and identify the optimal boundary between the possible outputs. Apriori algorithm analyzes the association between sales of various items and determines the frequency of items being sold together as an item set. To gain insight on customer opinion about the products, a feedback model is implemented.

II. PROBLEM STATEMENT

A retailer wants to undertsand the behavior of customers and the sales statistics. The knowledge of customer's opinion will give the retailer an insight to the buyer's needs and ways to modify his marketing or store design. It will help the retailer to make profitable business decisions such as which item when put on promotion is likely to trigger the sales of another associated item.

III. TRANSACTION DATASET

Our dataset is structured data of retail outlet downloaded from UCI Machine Learning Repository. It has around 50 items with 982 records. The data set contains transaction records in comma separated version. The attributes we consider are:

1) Order ID -unique numerical value for each transaction
2) Customer Name- first name & last name of customer
3) Order Date -date of transaction
4) iv. Order Priority- can be low, medium or high
5) Product Name- name of the item purchased
6) Unit Price-cost of item per unit
7) Quantity-total number of items purchased

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

IV. STRATEGY

The transaction data set is first pre-processed. Various scatter plots are generated based on the value of attributes in the transaction data. The data is then fed to the Apriori algorithm. It generates item sets of associated products based on their frequency of occurrence. This information is used to plot histograms to compare sales of all products in company of one particularly frequent product. We build subsets of the transaction dataset and perform clustering for predicting sales trends and represent the behaviour in form of a graph.

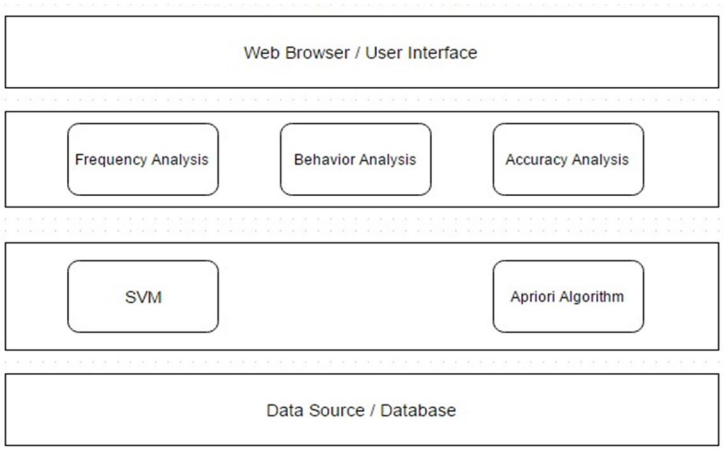

Fig. 1 Analysis Architecture

Two algorithms are implemented to achieve better accuracy in the analysis of the given data set. The two algorithms used are:

A. Apriori Algorithm

Apriori [1],[2] is an algorithm to perform association rule mining. It operates on databases containing transactions such as collection of items purchased by customers, or details of any website. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time.

The frequency of occurrence of the selected items is known as support. The item set that does not satisfy the minimum support is removed and the rest of the item set forms the level 1 frequent item set. For level 2, each item from level 1 is paired with every other item and support is calculated again for the item set. The item set that doesn't satisfy the minimum support condition is eliminated and the remaining sets forms the Level 2.

1) *Support:* Support [1],[3],[4] is defined as the frequency of occurrences of items divided by the size of the transaction. Support is calculated as follows:-

The support of itemset $\boldsymbol{X}$ with respect to $\boldsymbol{T}$ is defined as the proportion of transactions $\boldsymbol{t}$ in the dataset which contains the itemset $\boldsymbol{X}$.

$$\boldsymbol{supp(x)} = \frac{|\{\boldsymbol{t} \in \boldsymbol{T}; \boldsymbol{X} \subseteq \boldsymbol{t}\}|}{|\boldsymbol{T}|}$$

2) *Confidence:* Confidence [1],[3],[4] is the probability or the likelihood of the combination of item sets being purchased by the customers

The Confidence is calculated for the item sets as follows:-

The confidence value of a rule, $\Rightarrow Y$, with respect to a set of transactions T , is the proportion of the transactions that contains X which also contains Y

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

B. Support Vector Machines

SVM [5] are supervised learning models that analyse data and recognize patterns. Classification and regression analysis is performed using SVM. An SVM model is a representation of observations as points in space which are mapped such that the observations of distinct categories are divided by a clear gap that is as wide as possible.

SVM can also be used for non-linear classification using kernels; however it is a linear classification model using SVM and predicting the quantity of sales of the most frequently purchased item. In the dataset, the most frequently purchased item is found to be frankfurter. The sales of frankfurter in classified into two classes HIGH and LOW depending upon its quantity being purchased in various months of the year.

Now, this is a two-class separable dataset so there are many possible linear separators. SVM draws a decision boundary in the middle of the void between data items of the two classes. The SVM defines the criterion to be looking for a decision boundary such that it is maximally far away from any data point in a feature space. This distance from the decision surface to the closest data point determines the margin of the classifier. This method of construction means that the decision function for an SVM is completely specified by a subset of the data which defines the position of the separator. These points are referred to as the support vectors i.e. in a vector space, an observation can be thought of as a vector between the origin and the observation point.

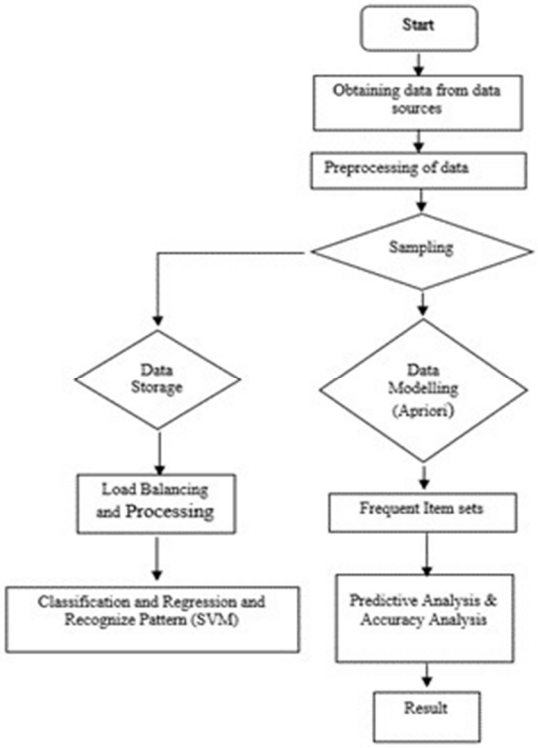

Fig. 2 System Flow

V. ALGORITHM

Input: Transaction Dataset ‘data’ in CSV format
With following attributes: -
Order ID: numeric value
Customer Name: nominal value
Order Date: MM/DD/YYYY
Order Priority: nominal attribute with values low, medium, high, critical
Product Name: nominal value
Unit Price: numeric value
Quantity: numeric value
Value of constraints:-
Support value: supp
Confidence value: conf
Item set length: minlen
TopN: topLevel

A. Output

1) Scatter plots for attributes Product name, Customer name, Quantity and Unit Price
2) Most frequent items(MFI)
3) Support value, Confidence value, Lift value
4) Frequent item sets, association ruleset
5) A relative-items plot
6) SVM classification graph showing prediction of sales behaviour
7) Accuracy of prediction

B. Procedure

1) Step 1: Data preparation Sample the data to form subsets with necessary attributes Decompose order date to get order month

2) Step 2: Begin with length of item set=1 Perform step 3 through step 5 till length of item set=minlen

3) Step 3: Calculate support of item set X as

$$\boldsymbol{supp(x)} = \frac{|\{\boldsymbol{t} \in \boldsymbol{T}; \boldsymbol{X} \subseteq \boldsymbol{t}\}|}{|\boldsymbol{T}|}$$

4) Step 4: Calculate confidence of item set X as

$$\boldsymbol{conf(X \Rightarrow Y)} = \frac{\boldsymbol{supp(X \cup Y)}}{\boldsymbol{supp(X)}}$$

5) Step 5: Eliminate all item sets which have support value less than threshold support. They are eliminated from candidate item-set list

6) Step 6: Fetch topN frequent items and plot the graph product versus frequency for topN items

7) Step 7: Plot the relative items graph by comparing frequencies of most frequent item and its associated items

8) Step 8: Make a dataset 'newdata' as a subset having records of transactions including the most frequent item with attributes orderID, value, order Month and Quantity

9) Step 9: Plot the predictive behavior of sales quantity(high/low) of most frequent item using support vector mechanism

10) Step 10: Generate confusion matrix by finding the values of true high, true low, false high and false low predictions

Calculate accuracy as

$$classification\ accuracy = (\frac{correct\ predictions}{total\ predictions}) * 100$$

VI. STATISTICAL MEASURES OF PERFORMANCE

After deeper analysis about the performance of classification system, the measures derived from this model are as follows:

A. Accuracy

It tells us the proportion of correct answers that the classifier can produce based on the model built.

B. Recall

It calculates the chances of getting all the answers correct i.e. the proportion of positives that are correctly identified as such. It is also referred as fraction of relevant instances that are retrieved. Statistical measures Sensitivity and True Positive Rate are same as recall.

C. Precision

It measures how many correct answers can be produced by the classifier on an average i.e. fraction of retrieved instances that are relevant.

VII. RESULTS

The analysis report produced by this model consists of various results which are as follows:

A. Customer Name vs Quantity

The scatter plot relating the customer vs quantity shows the quantity of products bought by various customers. This graph can give an idea about the customers who have purchased large quantity of products.

B. Association Rule Sets

Second result obtained is the rule set generated from apriori algorithm in relation with the given dataset. These rule set [1],[3] explains how the products are related to other purchased items. This characteristic is explained using support and confidence defined by apriori.

The retailer can use this analysis to decide right items for promotion and which one to exclude, ultimately maximizing profit of the store.

C. Frequently Purchased items

After deciding the rulesets, the next step of the analysis process is to guess the most frequent item. The most frequent item graph is plotted based on how frequent an item occurs in the ruleset.

D. Lift ratio

Lift ratio [3],[4] tells how significant the relationship between the antecedent and consequent item sets is. It is the ratio of confidence to the expected confidence. It is visualized as a horizontal bar plot.

Large lift ratio implies higher significance in the association rules. Items with low lift ratio have lower chance of forming an effective association rule set of items.

E. Relative Items Plot

It is the plot of frequently purchased items relative to the item being purchased by most of the customers.

	Attributes				Analysis	Future decisions that can be made based on analysis
	Product Name	Customer Name	Quantity	Unit price		
Scatter plot 1	✓		✓		Products which have been purchased in large quantity	Which products need more marketing and offers
Scatter plot 2	✓			✓	How expensive each item is	Which products need to be priced lower to maximize sales
Scatter plot 3		✓	✓		Customers who have purchased large quantity of products	Offer discount on next purchase to customer who makes bulk purchases
Scatter plot 4		✓		✓	Spending capacity of customer	Retain customers who make more profitable transactions

Table 1 Analysis of Scatter Plots

VIII. CONCLUSION

This report explains how the behaviour analysis is done with the help of machine learning algorithms, association rule mining and visualization libraries.

The results achieved are

1) The relation between various products which are related and a ruleset is generated with the help of apriori algorithm.
2) The retailer can use the item set rules to conduct cross-promotional programs & promote items.
3) The analysis of the most frequent item is established with the help of linear classification using SVM.
4) SVM classification plot helps the retailer to decide that in which month of the year an item should be promoted depending on the sales of its item sets derived from the association rules.
5) The accuracy achieved through this prediction ranges from 68% – 79%.

REFERENCES

[1] SunithaVanamala, L.Padma sree, S.Durga Bhavani "Efficient Rare Association Rule Mining Algorithm" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 3, May-Jun 2013, pp.753-757 753 | P a g e

[2] "Sales analysis using product rating in data mining techniques" Sushant Bhagwat1, Vishnu Jethliya2, Ankit Pandey3, Lutful Islam4, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-730

[3] "Building an Association Rules Framework to Improve product Assortment Decisions" Tom Brus etal, Department of Economic Science, Limburg University centre, Belgium , Data mining and Knowledge discovery , October 2003

[4] Michael Hahsler, Christian Buchta, Bettina Gruen, Kurt Hornik and Christian Borgelt, "Package arules", 2016.

[5] David Meyer, "Support Vector Machines – The Interface to libsvm in package e1071", FH Technikum Wien, Austria, 2015