Data Visualization and Data Mining for Retailers

Kaustubh P. Nagwekar  
UG Student  
Department of Computer Engineering  
PVPP College of Engineering, Mumbai, Maharashtra, India – 400022

Kishanu S. Chowdhary  
UG Student  
Department of Computer Engineering  
PVPP College of Engineering, Mumbai, Maharashtra, India – 400022

Gaurav M. Shejwal  
UG Student  
Department of Computer Engineering  
PVPP College of Engineering, Mumbai, Maharashtra, India – 400022

Vinod N. Alone  
Assistant Professor  
Department of Computer Engineering  
PVPP College of Engineering, Mumbai, Maharashtra, India – 400022

Abstract

In today’s world, the retail industries are facing many challenges. Retailers usually have large amounts of sales data which can be used to overcome these challenges by extracting some knowledge from it and make some future prediction. To extract knowledge from huge amount of data sets, Data Mining techniques are used. Through our system, our aim is to provide simple, lightweight and user-friendly tool, that can extract information directly from the operational databases in order to provide quick results to support decision making and strategic planning.

Keywords: Data Visualization and Data Mining, Retail industry, C4.5 algorithm, Apriori algorithm, Association rules

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data set involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems in combination of one or more.

Data Mining also helps in reducing information overload along with the improved decision-making by searching for relationships and patterns from the huge dataset collected by retailers [1].

Data mining prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [2]. From the last decade data mining have got a rich focus due to its significance in decision making and it has become an essential component in various industries [2].

The idea of our proposed system evolved because of the increased demand of strategic information which can supports business decision in retail industry and has to be extracted from operational database quickly. We are designing the system in such a way that it is simple, lightweight & more user friendly.

Retailers have been collecting enormous amounts of data throughout the years, just like the banking industry, and now have the tool needed to sort through this data and find useful pieces of information. For retailers, data mining provides information on product sales trends, customer buying habits and preferences, supplier lead times and delivery performance, seasonal variations, customer peak traffic periods, and similar predictive data for making proactive decisions.

The existing system are providing different types of data mining algorithms and accepts inputs in different formats but are little-bit time consuming and sometimes it is difficult to understand the results. These systems are mostly designed to work effectively in different types of domains.

II. EXISTING SYSTEM

A. Weka tool [4]:
Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

B. RapidMiner[5]:
RapidMiner is a data science software platform which provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. RapidMiner is written in the Java programming language.
III. Comparison

Based on the analysis of the related work with our proposed system, following are the observations:

Table – 1
Comparison (Proposed System and Existing System)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proposed System</th>
<th>Existing System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intended users</td>
<td>Specifically designed for retailers.</td>
<td>Designed to use in different types of domains.</td>
</tr>
<tr>
<td>Input Sources</td>
<td>Connect to the database using database connector.</td>
<td>Accepts different types of formats e.g.: csv file, arff file etc.</td>
</tr>
<tr>
<td>Technology used</td>
<td>Java &amp; Python Programming language is used because they are platform independent &amp; supports high level data structures.</td>
<td>Implementation is done using mostly Java Programming language.</td>
</tr>
</tbody>
</table>

IV. Proposed System

The proposed design is divided into six modules:-

A. Data Source:
The data source will be the database of retail stores which contains the daily sales records. This database contains various information such as Customer ID, purchase product, product category, product brand, price, purchase date etc.

B. Data Extraction:
The data will be extracted from the data sources using Java and Python, which are programming languages and environment for statistical computing and graphics. Eclipse is free and open-source IDE for Java and Python.

C. Java and MySQL connection:
The retrieval and storage of data from a MySQL database with Java is possible by using Java MySQL connector and java.sql package. This package simply needs to be installed and imported in the source code.

D. Python and MySQL connection:
The retrieval and storage of data from a MySQL database with Python is possible by using Python MySQL connector and mysql.connector package. This package simply needs to be installed and imported in the source code.

E. Classification and Association Algorithms:
Classification and Association algorithms are applied on the data to classify the data based on some criteria and association between the different item sets in database. C4.5 algorithm is use for classification and Apriori algorithm is use for association mining.

F. User Interface:
A user interface is use to accept the inputs and display the results in an appropriate manner in the form of statistical and Diagrammatical representation such as graphs and pie charts. Analyzing the conclusion, we will display the association rules between the different item sets and also display the decision tree.

Fig. 1: Architecture of the proposed system
V. ALGORITHMS

A. C4.5 Algorithm:

C4.5 is a statistical classifier algorithm used to generate a decision tree.
This algorithm has a few base cases:
1) All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
2) None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
3) Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

In our system C4.5 algorithm is implemented using Java Programming Language (Java SE 8).
Inputs and outputs of C4.5 algorithms is as follows:

![Image 1: C4.5 Algorithm Inputs](image1.jpg)

**Fig. 2: Inputs for C4.5 Algorithm**

![Image 2: C4.5 Algorithm Decision Tree](image2.jpg)

**Fig. 3: Output of C4.5 Algorithm**
B. Apriori algorithm:-

The Apriori Algorithm is an influential algorithm for mining frequent item-sets for boolean association rules. This algorithm learns association rules and is applied to a database containing a large number of transactions. Association rule learning is a data mining technique for learning correlations and relations among variables in a database. Two thresholds are set as minimum support and minimum confidence.

In our system Apriori algorithm is implemented using Python Programming Language (Python v3.4.4).

Inputs and outputs of Apriori algorithms is as follows:

![Apriori Algorithm Input](image1)

![Apriori Algorithm Output](image2)

VI. ADVANTAGES

Our system has simple design and specifically designed for retailers. Our system has simple, light-weight & more user friendly interface. It directly extract information from the operational database so to provide results quickly.

VII. APPLICATIONS

Following are some of the application area in retail sector [3]:-

A. Acquiring and Retaining Customer:

It is more costly to reach new customers than to get existing one. So by knowing existing customers purchase behaviour, direct marketer can predict customers need and interest in buying particular product. Using this type of prediction retailer can retain existing customers by providing discounts or offer, attract customers and acquire customers.

B. Market Basket Analysis:

Market basket analysis is a technique in understanding what items are likely to be purchased together according to association rule. It provides valuable indications about customers, shopping patterns by showing associations among various items. This type
of item association is useful for shelf design, deciding the location and promotion of items by means of combination. So that customers can easily locate item and this analysis helps in product cross-selling.

C. Customer Segmentation and Target Marketing:

Segmentation is to divide the market into several parts by certain characters. Data mining can be used in grouping or clustering customers based on the behavior. This type of information is useful to define similar customers in a cluster, holding on good customers and identify likely responders for target marketing.

VIII. Conclusion

In this paper we have presented a way to analyze sales information from retail stores. Our system directly extract information from the operational databases in order to provide quick results to support decision making and strategic planning to compete in the retail sector.

REFERENCES