Chapter 8
Association Rule

ABSTRACT

Businesses are increasingly seeking to understand consumer behavior and purchasing habits in order to analyze the relationship between different products purchased by customers. By leveraging the association rule mining technique, businesses can identify complementary products and bundle them together to increase sales. This chapter provides an overview of association rule mining, a form of unsupervised learning that enables data scientists to analyze the relationship between data items in a dataset. The chapter explains the Apriori algorithm, a feature of association rule mining, and highlights how data scientists can collaborate with domain experts to achieve business objectives such as product matching and testing the efficiency of association rule results. Readers can follow a step-by-step guide to experience the association rule mining technique using RapidMiner, enabling them to develop their understanding of this valuable analytical tool.

INTRODUCTION

In business sectors, there is a need to sell multiple products at the same time. For example, superstores often offer several products to their customers, such as offering soap along with shampoo. This kind of product matching is done by analyzing the association rules on customer purchasing behavior (Ren, 2021: Yingzhuo & Xuewen, 2021), which is the customers who buy soap often buy shampoo as well. Therefore, association rules are used to process large data sets to analyze the correlation rules hidden within the data (Paruechanon & Sriurai, 2019: Hong & Nan 2021: Salman & Sadkhan, 2021). The technique is also used to analyze the correlation rules that can occur together; for example, when a car is broken, it can lead to an accident while driving. In processing big data with the association rules, data scientists may be given a number of rules. Data scientists, therefore, are necessary to analyze the benefits and efficiency of association rules in practical applications in industries. Data mining techniques are not only used to analyze business data, but it has also been applied in many industries such as medical diagnosis, Web Mining and Bioinformatics (Toumi, Gribaa, & Ben Abdessalem Karaa, 2021).

DOI: 10.4018/978-1-6684-4730-7.ch008

ASSOCIATION RULE

In analyzing data with Association Rule, data scientists can use the following format (Zhang & Zhang, 2002).

Soap -> Shampoo

From such an association rule, it means that when customers buy soap, they will also buy shampoo. Another metric used for association rule analysis is Confidence, which is the percentage of the availability of rules-based data within the dataset. For example, if the Confidence value is 60 percent, and within the dataset there are 30 instances of soap purchases out of the total 100 instances, it means there is 60 percent of shampoo purchase data, or there are 18 times that soap along with shampoo are purchased.

Another measure is Support, which refers to the percentage of compliant data in the entire dataset. For example, a Support value of 2% means that the entire dataset has compliant instances, which is there are 2 instances of buying soap along with shampoo out of all 100 instances.

To analyze association rules with Confidence and Support values, data scientists can calculate the values of both as in the following example.

The store has a total of 100,000 customer purchases. Of all the data, the total number of milk purchases is 40,000 times, and the total number of fruit purchases is 15,000. Of all the 40,000 milk purchases, the number of purchases of milk along with fruit is 10,000 times. Confidence is calculated, and it results as 25%, which means there is the frequency of buying fruit together with milk at 25 percent. And the Support value is 10 percent which means that the frequency of buying fruit along with milk is at 10 percent of the total data. After analyzing the association rules, data scientists will obtain association rules consisting of a large number of Confidence and Support values. Therefore, the data scientists need to determine the Support and Confidence Threshold in order to measure whether each association rule can be used in analysis. For example, when data scientists define Support and Confidence Thresholds as 50%, after analyzing the data the Confidence and Confidence Thresholds must not be less than 50% to be considered a Strong Association.

The data ready for analysis with association rules are usually in Boolean format. For example, to buy equals 1, and not to buy equals 0. The analysis of association rules based on Boolean can be performed as in the following example.

Table 1. Example of correlation rule analysis from Boolean data

Frequency of Each Purchase	Soap	Shampoo	Detergent	Dishwasher
1st Purchase	1	1	0	0
2nd Purchase	1	0	0	1
3rd Purchase	1	1	0	0
4th Purchase	1	1	1	1