Chapter 7
Clustering

ABSTRACT

Clustering is employed to divide a data set into an appropriate number of groups. Clustering is a form of unsupervised learning, which means a data scientist can bring labelled features of interest into the mining model. Furthermore, after dividing the data set, the data scientist can label each cluster. In business, clustering is used to analyze a customer or product segment that matches a target market. This chapter introduces clustering techniques including k-means, hierarchical clustering, and DBSCAN as well as techniques to indicate the efficiency of the clustering analysis. Data scientists can assess the efficiency of clustering analysis in two ways. Firstly, subjective measurement is where a data scientist consults a domain expert to confirm the efficiency of the cluster analysis, and secondly, data scientists can use objective measurements that test the efficiency of the cluster analysis result based on calculations. This chapter demonstrates cluster analysis adoption with RapidMiner so that readers can follow the process step-by-step.

INTRODUCTION

Clustering is Unsupervised-Learning data mining technique (Govindaraj et al., 2020: Abbas et al, 2021). Data scientists are able to group the quantitative data without defining the target variables (Labeled) nor dividing the dataset into 2 parts to teach the machine learning. Clustering technique gather similar data into one group and bring dissimilar data into another group. Therefore, clustering can be applied in a wide range of industries, such as customer segmentation according to customer purchasing behaviors (Punhani et al., 2021). The customers in the same group tend to have similar purchasing behaviors and expected prices. Clustering is applied in order that the business sectors can design marketing plans suitable for each group of customers. It can also diagnose the stage of cancer in each patient from the sizes of the tumors growing in the patient's organs (Kumar, Ganapathy & Kang, 2021).

Although the process of data science provides computational methods to obtain the optimal number of clusters, to specify the number of clusters depends on the objectives of the analysis. Data scientists need to consult with domain experts to gain the exact number of clusters to be analyzed. For example, clustering the sizes of shirts that will be produced by analyzing the clustering data from the purchase

DOI: 10.4018/978-1-6684-4730-7.ch007

history datasets. In this way, data experts can advise on a number of clusters of shirt sizes that are appropriate, such as small, medium, and large.

Clustering can be used to analyze the centroid of the data in each cluster. It can be applied to set the car sales price from the purchase history. Each car model can be produced in many specifications. After the analysis, the car dealers can set various prices of each specification under the same car model by using the data in each group. For example, Model A car with the entry-level specification is 1,300,000 baht. Model A car with the intermediate specification is 1,600,000 baht. Model A car with the highest specification is 1,700,000 baht. As mentioned in the example, the price determination is from the analysis using clustering technique. Data scientists can perform a centroid analysis to each group of data according to the customer purchasing behavior.

K-MEANS CLUSTERING

Once data scientists are asked about the nature of the data, such as the purchasing behavior of each group of customers, they need to find a consistent data set to answer those questions (Ginting, 2021: Puspasari et al., 2021). The dataset initially received is characterized as Unlabeled Data, the data which has not yet been clustered nor defined for its name or meaning. Therefore, these data sets can be clustered and defined according to the objectives of the data analysis, such as definition of the customers; the middle-class or the wealthy.

K-Means Clustering Algorithm

Clustering data can be done by bringing 2 Data Objects; the shirt width and the shirt length; to create a Data Point; the sizes. Data scientists can develop a display in the Scatter Plot format to see the intersection at the data point created from the 2 data objects, and then colorize each data point according to clusters. For example, the yellow data point represents the medium size, and the brown data point represents the large size.

Figure 1. Unlabeled data and data after clustering