Chapter 24
Clustering by K-Means Method and K-Medoids Method:
An Application With Statistical Regions of Turkey

Onur Önay
School of Business, Istanbul University, Turkey

ABSTRACT

Data science and data analytics are becoming increasingly important. It is widely used in scientific and real-life applications. These methods enable us to analyze, understand, and interpret the data in every field. In this study, k-means and k-medoids clustering methods are applied to cluster the Statistical Regions of Turkey in Level 2. Clustering analyses are done for 2017 and 2018 years. The datasets consist of "Distribution of expenditure groups according to Household Budget Survey" 2017 and 2018 values, "Gini coefficient by equivalised household disposable income" 2017 and 2018 values, and some features of "Regional Purchasing Power Parities for the main groups of consumption expenditures" 2017 values. Elbow method and average silhouette method are applied for the determining the number of the clusters at the beginning. Results are given and interpreted at the conclusion.

INTRODUCTION

Data science and data analytics are becoming more and more important with their wide application area. Spending time on the internet, visiting shopping sites, reading the news sites, using search engines, looking social media posts, adding comments to contents etc. each contributes to the formation of datasets. So data is constantly growing from social media, weather stations, government agencies, purchases and so on (Dichev, & Dicheva 2017). Data is very important source of knowledge. In business to design successful strategies and policies data science is widely used (Gibert et al., 2018). Hundreds of scientific studies and real-life applications can be found from the internet in data science and data analytics. A lot of applications can be found according to data which are collected in different areas. Data can be in a

DOI: 10.4018/978-1-7998-3053-5.ch024

variety of formats, such as numeric, text or image and etc. Data science and data analytics can help us understand the data by analyzing various methods. Data science includes mathematical and statistical analysis combined with information technology tools and builds systems and algorithms to discover the information, to detect the patterns, and create useful insights and predictions while doing this it uses techniques, such as classification, clustering, regression and association rule mining (Molina-Solana et al., 2017).

Clustering methods are used as data science methods and that can be used to understand the meaning of the data. The data are grouped into clusters and the resulting clusters are interpreted. There are many types of clustering methods, such as partitioning, hierarchical, grid-based and model-based methods (Kaur et al., 2014). In this study, k-means and k-medoids methods are used which are in partitioning clustering methods. Clusters are arranged with these methods by looking at the distances between the data.

In this study, Statistical Regions of Turkey in Level 2 are clustered with k-means and k-medoids clustering methods. Turkey has three different levels of Statistical Regions. They are “Level 1”, “Level 2” and “Level 3”. Details of the Statistical Regions of Turkey are given at the section two (background). Turkey has 7 geographical regions and their details are given before the analysis. The dataset 2017 (in Table 3) consists of “Distribution of expenditure groups according to Household Budget Survey (Horizontal %), 2015-2017, 2017”, “Gini coefficient by equivalised household disposable income 2017” and some features of “Regional Purchasing Power Parities for the main groups of consumption expenditures” for 2017. The dataset 2018 (in Table 2) consists of “Distribution of expenditure groups according to Household Budget Survey (Horizontal %), 2016-2018, 2018”, “Gini coefficient by equivalised household disposable income” for 2018. Statistical Regions of Turkey which are in the same or in a different cluster can be identified and interpreted by the clustering analysis according to datasets. Anyone who knows a region can make inferences about other regions in the same cluster. So data science methods help us understand the clustering of Statistical Regions of the Turkey in Level 2 according to topics of datasets.

There are seven sections in this study. Section one is the introduction. Section two is the background. In this section, there is information about the classification of Statistical Regions system of Turkey and some studies are given as examples from the literature which are related with the study. Section three is the main focus of the chapter part. The information about clustering, k-means, k-medoids, determination of the numbers of the clusters and information of the data are given in section three. Analyses and results are given in section four which is the solutions and recommendations section. Some ideas are given for the future research direction in section five. In the section six conclusions are given. And in final section references are given. The overview of the study can be shown by steps as following;

Step 1: Determine the research problem
Step 2: Do background research
Step 3: Give information of the data and analysis methods
Step 4: Analyze the data
Step 5: Communicate the results
Step 6: Conclusion