Chapter 4
An Optimized Clustering Quality Analysis in K-Means Cluster Using Silhouette Scores

F. Mohamed Ilyas
Bharath Institute of Higher Education and Research, India

S. Silvia Priscila
Bharath Institute of Higher Education and Research, India

ABSTRACT

Data-driven problem-solving requires the capacity to use cutting-edge computational methods to explain fundamental phenomena to a large audience. These facilities are needed for political and social studies. Quantitative methods often involve knowledge of concepts, trends, and facts that affect the study programme. Researchers often don't know the data's structure or assumptions when analysing it. Data exploration may also obscure social science research methodology instruction. It was essential applied research before predictive modelling and hypothesis testing. Clustering is part of data mining and picking the right cluster count is key to improving predictive model accuracy for large datasets. Unsupervised machine learning (ML) algorithm K-means is popular. The method usually finds discrete, non-overlapping clusters with groups for each location. It can be difficult to choose the best k-means approach. In the human freedom index (HFI) dataset, the mini batch k-mean (MBK-mean) using the Hamely method reduces iteration and increases cluster efficiency. The silhouette score algorithm from Scikit-learn was used to obtain the average silhouette co-efficient of all samples for various cluster counts. A cluster with fewer negative values is considered best. Additionally, the silhouette with the greatest score has the optimum clusters.

DOI: 10.4018/979-8-3693-1355-8.ch004

INTRODUCTION

Automation is practically everywhere in organisations; each department creates a certain amount of transactions. These operations are carried out in continuous streaming sequences of data objects. The main problem for researchers is handling this volume and amount of streaming data (Kumar & Shankar Hati, 2021). The problem is coping with high-dimensional, large-volume big data sources that change frequently. The so-called data streams are enormous, unrestricted streams of data that come in and go out continuously, and the data is unavailable for access and future treatment (Lohith et al., 2015). The database in the data stream may include supervised and unsupervised data, the fundamental methodologies used in ML algorithms (Lohith, Singh & Chakravarthi, 2023). With unlabeled data from the dataset, the unsupervised algorithm finds hidden data structures. Clustering algorithms are frequently used to identify related data groupings based on the dataset's hidden structures, which may also be regarded as a key component of data science (Sarker, 2021). Clustering is an effective data science tool. The maximum level of similarity within a cluster and the maximum level of dissimilarity across clusters are used to discern the cluster structure of a data set using this method (Marar et al., 2023; Sholiyi et al., 2017).

Hierarchical clustering was the original method social and biologist scientists used, and cluster analysis has developed into a speciality of statistical multivariate analysis. There is also unsupervised ML involved. Clustering algorithms are statistically classified as nonparametric procedures and probabilistic model-based approaches (Nomula et al., 2023). Clustering uses the mixture likelihood technique employed in probability model-based approaches since it implies that the data points derive from a mixture probability method. The Expectation and Maximization (EM) algorithm is the most popular model-based algorithm (Yu et al., 2018). Clustering techniques for nonparametric approaches can be separated into hierarchical and partitional techniques, which are more popular (Yang et al., 2018). These techniques are based primarily on a subjective measure of similarity or dissimilarity values.

Customer segmentation was previously accomplished via ML. Unsupervised ML is employed. K-means or Hierarchical Clustering combined with the PCA (Principal Component Analysis) approaches are used (Pradana, 2021). Customer ratings were calculated using K-means and RFM (Frequency, Monetary, Recency) Analysis (Erickson et al., 2017). Some of these studies segment them only based on prediction numbers or numerical values generated by ML, such as spending, RFM Score, and annual income (Andrews & Hemberg, 2018), rather than categorising them using qualitative and descriptive data. If a question arises, in which city does the client group with the maximum national income reside? Therefore, we must look into the data more to find the answer. Another issue is how to integrate the data properly. Data that has been organised well will make analysis and report preparation easier. Additionally, the data's quality needs to be taken into account. Better data governance requires overcoming issues with data, such as duplication, disparate formats, missing data, and filthy data (Amezquita et al., 2020).

Additionally, unsupervised clustering techniques frequently divide a set of unlabeled data into various groups with related characteristics (Kiselev et al., 2019). Because single-cell transcriptomics datasets can contain millions of unlabeled observations (or cells), the most prominent clustering methods are used in healthcare (Kiselev et al., 2017). Cells will be arranged into groups with distinct labels that roughly correspond to genuine biological categories (Risso et al., 2018; Sudheesh et al., 2023b). In this perspective, various clusters can be viewed as diverse cell kinds or cell states, which can be investigated further in subsequent analyses (Li et al., 2020). The most prevalent partitional algorithm for clustering is k-means. The algorithm divides N cells into k clusters, with every cluster's centroid (or mean profile) representing the cells in that cluster. This approach is frequently applied independently and as a part of