AN APPLICATION OF K-MEANS CLUSTERING FOR CUSTOMER SEGMENTATION IN ONE LUXURY GOODS COMPANY

Hana Stefanovic[1], PhD; Stefana Janicijevic[1], PhD; Goran Bjelobaba[2], MSc; Ana Savic[3], PhD
[1] Comtrade Information Technology School of Applied Studies, Belgrade, SERBIA,
hana.stefanovic@its.edu.rs, stefana.janicijevic@its.edu.rs
[2] National Bank of Serbia, Belgrade, SERBIA, Goran.Bjelobaba@nbs.rs
[3] School of Electrical and Computer Engineering of Applied Studies, Belgrade, SERBIA, ana.savic@viser.edu.rs

Abstract: *In this paper K-means clustering algorithm is applied in order to classify customers into several groups showing the similarity within a group is better than among groups. After determining the relevant client's attributes in a SQL Server database, K-means is applied in MATLAB programming environment, using fixed number of clusters. Each centroid defines one of the clusters, while each data point is assigned to the nearest centroid, based on the squared Euclidean distance. In this research, centroids are randomly generated, while the separation distance between the resulting clusters is analyzed and illustrated using the Silhouette index. The analysis and results presented in this paper could determine a similarity in purchasing or using the services by a population cluster in one luxury goods company, to develop market segments, to identify repetitive behavior or trends in order to evaluate client actions and to create some new customer loyalty campaigns.*

Keywords: *cluster analysis, dendrogram, K-means, Silhouette index*

1. INTRODUCTION

K-means clustering is one of the simplest unsupervised machine learning algorithms [1]. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes [2]. The objective of *K*-means is to group similar data points together and discover underlying patterns [3-5]. To achieve this objective, *K*-means looks for a fixed number of clusters in a dataset. *K*-means algorithm identifies *K* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible [6], [7]. To process the learning data, the *K*-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids [8-10].
In this paper *K*-means is applied in order to segment clients for next marketing campaign in one luxury goods company. From this point of view, *K*-means is used to find out the most significant clients of company through clustering. The main goal is to identify relevant clients, who are also loyal, and to use their profile to create new digital marketing campaigns [11-13].
The database of company has been observed with data between the December 2010 and April 2018 year. There are more than 9.000 customer purchases records in data base, more than 1200 customer interaction records, and more than 250 distinct customers. An algorithm is written in MATLAB, including the results and interpretation.
Clustering methodology proposed in this paper includes: loading and cleaning data (removing nulls, NaNs and removing outliers from database), preprocessing of data (reformatting of dates, indexing labels...), data analysis in order to find candidates for good quality features (visualize data), splitting data set on test and train sets, variable clustering to remove features with similar impact, testing and statistical comparison of candidates, feature vector normalization (mapping the distribution into 0.0–1.0 range), multiple clustering model testing with the selected features, and best model selection and final clustering, which is applied using hierarchical and non-hierarchical cluster methods in MATLAB [14]. Nearest neighborhood method is used as linkage clustering method, while the Silhouette index [15] is used to profile clusters and to indicate the clusters' separation, using the Euclidean or some another distance metric.
The main target and result is to attract new clients based on analyzed profiles and behavior patterns. Thus, the desired profile of the company's possible clients will be created from the data on existing loyal clients. As a result, the company management team will be able to create a digital marketing campaign that will target exactly this market segment, or to create alternative campaigns for the other significant client segments as well.

2. ALGORITHM OBJECTIVES AND SOME OPTIMIZATION STRATEGIES

In this paper, the clients are grouped into five clusters:
Cluster1: Medium interaction shopaholic customers (smaller spenders who prefer high-index types of items and who are contacted a moderate amount of time),
Cluster2: Low interaction modest customers (smaller spenders who prefer high-index types of items and who are contacted a fairly low amount of times. They have a significant amount of purchases),
Clustre3: High interaction rich customers (big spenders who are contacted many times with high-index types of contact, but who prefer low-index types of items),
Clustre4: High interaction modest customers (smaller spenders who prefer high-index types of items and who are contacted many times),
Cluster5: Low interaction rich customers (big spenders who are contacted moderately with low-index types of contact and who prefer low-index types of items).
Let $X_\tau = \{X_1,...X_N\}$ be the set of data points, where $C=(C_1,...,C_K)$ presents clustering into K groups. Let $d(X_k, X_l)$ be the Euclidean distance between X_k and X_l. Let $C_j = \left\{ X_1^j,..., X_{m_j}^j \right\}$ present jth cluster, $j=1,\ldots,K$, where $m_j=|C_j|$. An average distance a_i^j between i-th vector in clusterC_j and other vectors in the same cluster is [4]:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{k=1 \\ k \neq i}}^{m_j} d\left(X_i^j, X_k^j\right), \quad i = 1,...,m_j \tag{1}$$

Minumum average distance between i-th vector in cluster C_j and all the vectors in cluster C_k, $k=1,...,K$, $k \neq j$, is:

$$b_i^j = \min_{\substack{n=1,...,K \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{k=1 \\ k \neq i}}^{m_n} d\left(X_i^j, X_k^n\right) \right\}, \quad i = 1,...,m_j \tag{2}$$

Silhouette index of i-th vector in cluster C_j is:

$$s_i^j = \frac{b_i^j - a_i^j}{\max\left(a_i^j, b_i^j\right)} \tag{3}$$

So, $-1 \le s_i^j \le 1$. Silhouette index of cluster C_j is:

$$S_j = \frac{1}{m_j} \sum_{j=1}^{m_j} s_i^j \tag{4}$$

while the global silhouette value is:

$$S = \frac{1}{K} \sum_{j=1}^{K} S_j \tag{5}$$

taking values from -1 to 1.
The Silhouette plot [15] illustrated in Fig.1 shows that the data is split into five clusters, having large silhouette index values (0.7 or greater), indicating that the clusters are well separated. Silhouette index is calculated according to (5).
Clusters are formed such that objects in the same cluster are similar, and objects in different clusters are distinct. *K*-means clustering is a partitioning method that treats observations in data set as objects having locations and distances from each other. It partitions the objects into *K* mutually exclusive clusters, such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster is characterized by its centroid, or its center point, while an iterative algorithm that assigns objects to clusters is used, so that the sum of distances from each object to its cluster centroid, over all clusters, is a minimum. At each iteration, the algorithm reassigns points among clusters to decrease the sum of point-to-centroid distances, and then recomputes cluster centroids for the new cluster assignments. In this paper, the Euclidean distence, the squared Euclidean distance and cosine distance are used.
Grouping into clusters, with visualization, following clients' distribution according to their attributes (number of purchases, number of times contacted, total sales value) is given in Fig.2, Fig.3, Fig.4, Fig.5, and Fig.6, respectively.
The relationship between attributes (number of purchases, number of times contacted, total sales value) with data points categorized into five clusters, presented in different colors, are shown in Fig.7 and Fig.8.

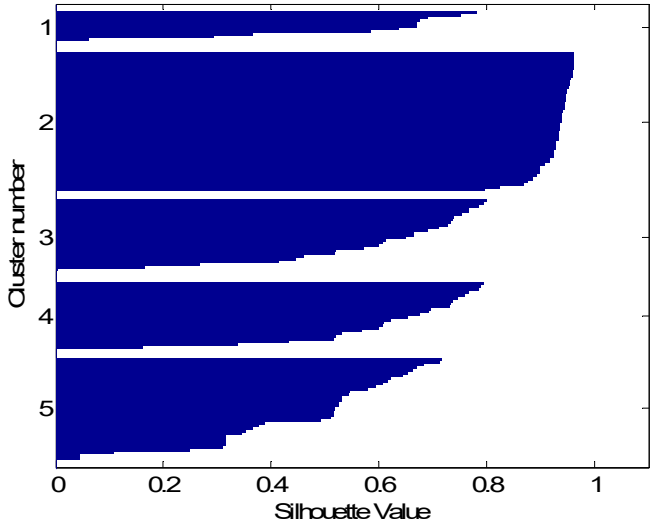

Figure 1: Silhouette index value

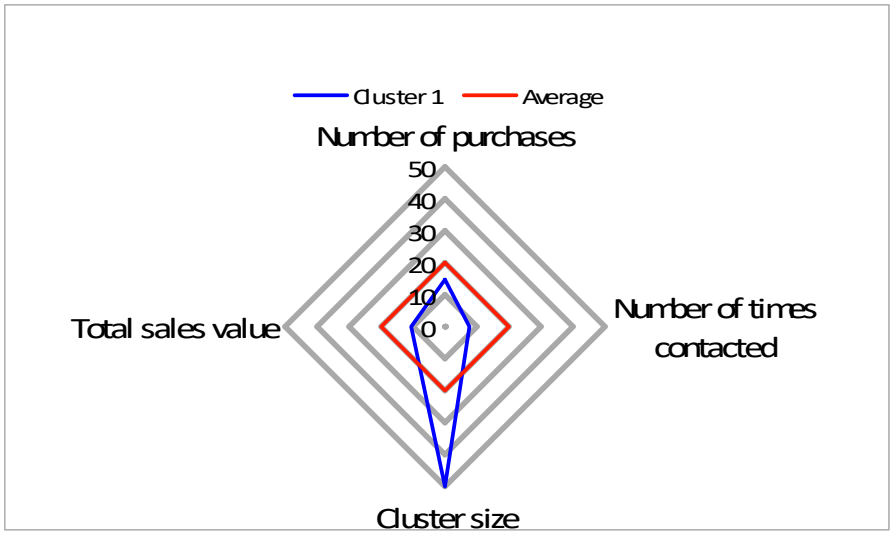

Figure 2: Cluster1

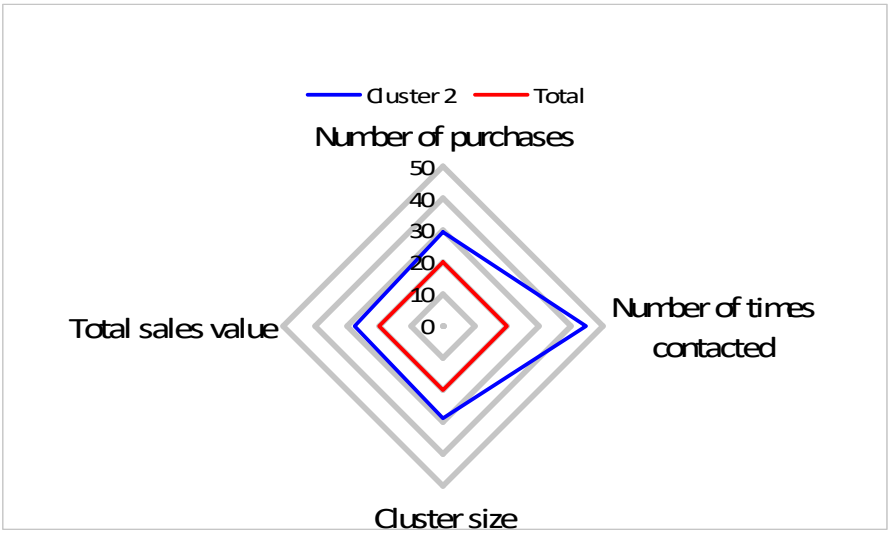

Figure 3: Cluster2

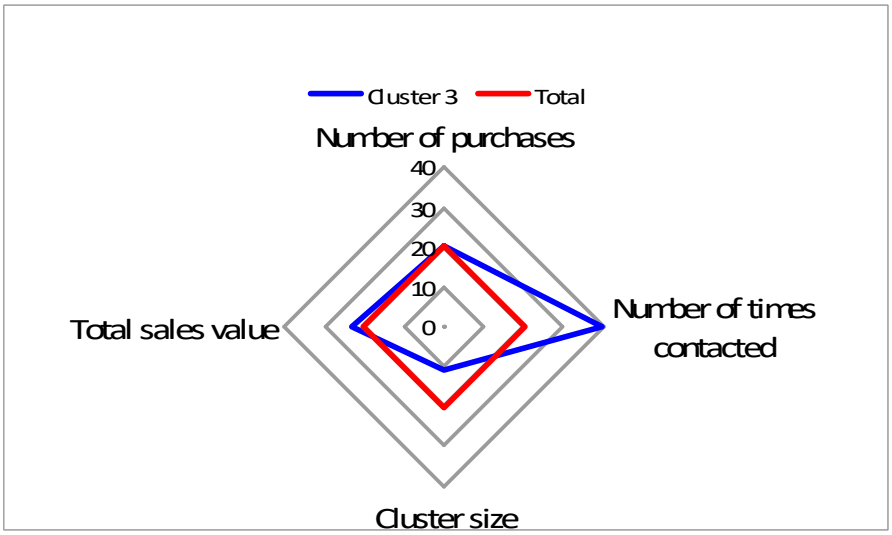

Figure 4: Cluster3

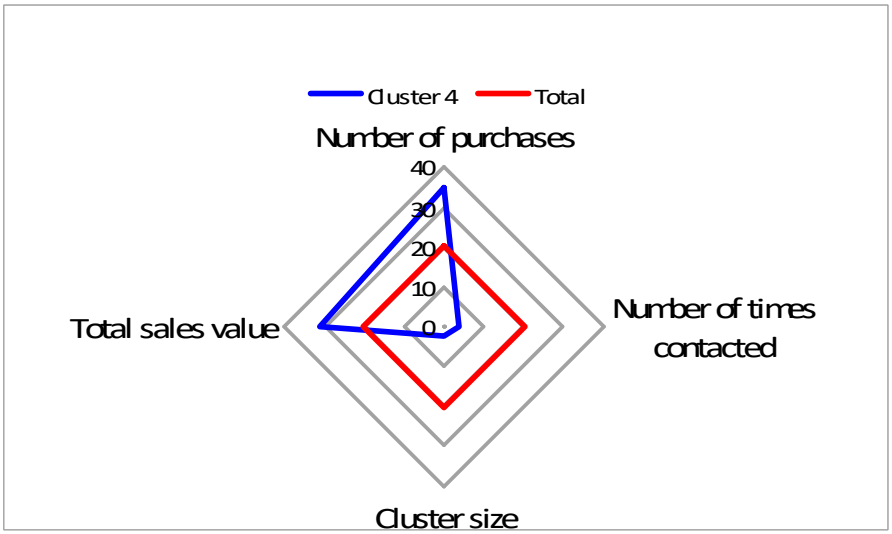

Figure 5: Cluster4

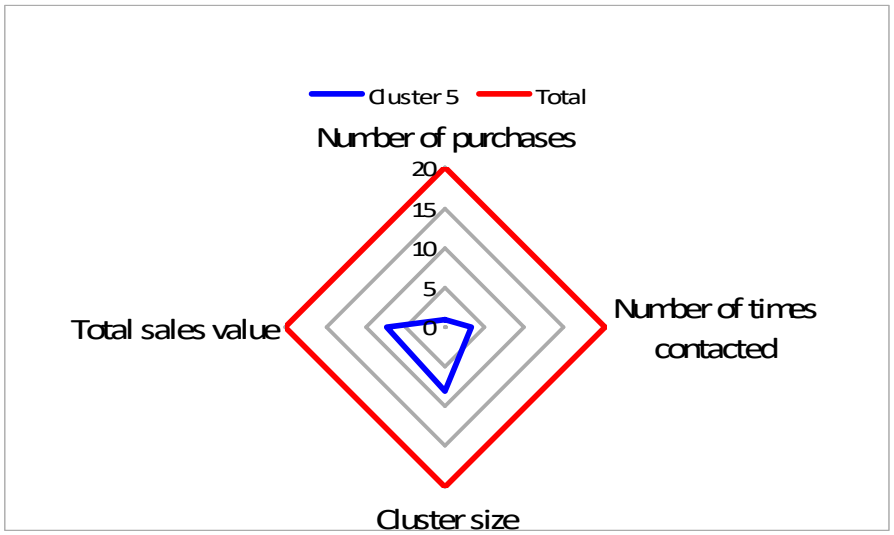

Figure 6: Cluster5

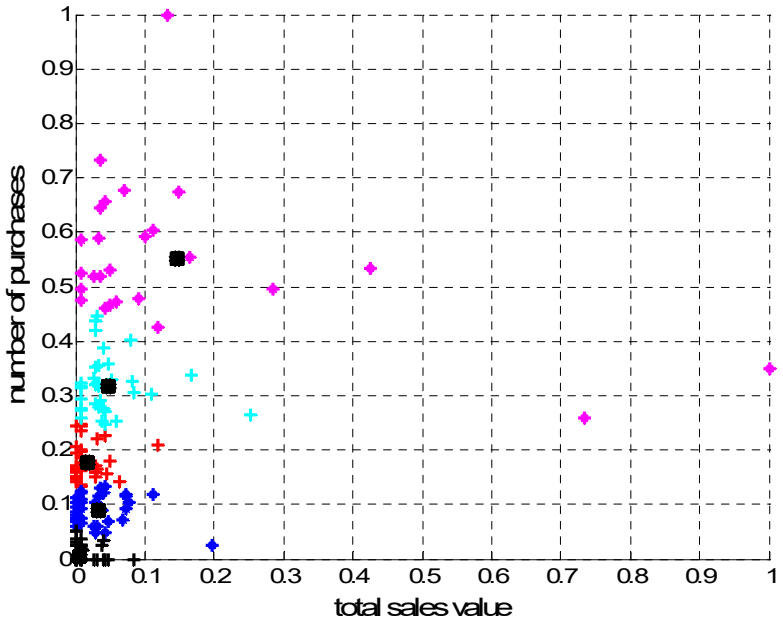

Figure 7: Data points in number of purchase vs total sales graph

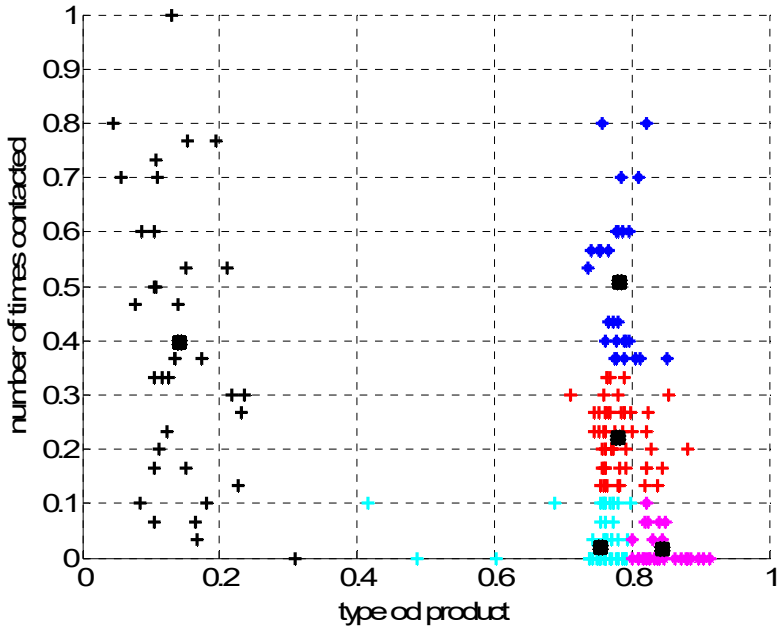

Figure 8: Data points in number of purchase vs total sales graph

3. PRODUCING NESTED SETS OF CLUSTERS USING HIERARCHICAL TECHNIQUES

Hierarchical clustering groups data into a multilevel cluster tree or dendrogram [14]. This technique can help in choosing the level of clustering that is most appropriate for specific application.
In order to visualize the hierarchy of clusters, a dendrogram is plotted, using Euclidean and cosine distance measure, which is presented in Fig 9. and Fig.10, respectively.
Creating a hierarchical cluster tree allows to visualize, all at once, what would require considerable experimentation with different values for *K* in *K*-means.
In this paper the single linkage, also called nearest neighbor is applied [16]. The nearest neighbor uses the smallest distance between objects in the two clusters. The centroid linkage, using the Euclidean distance between the centroids of the two clusters, could be also applied.
The root nodes in dendrogram trees confirm what *K*-means clustering produced: there are five distinct groups of observations, but specifying a linkage height that will cut the tree below the three highest nodes, we also could create four clusters.

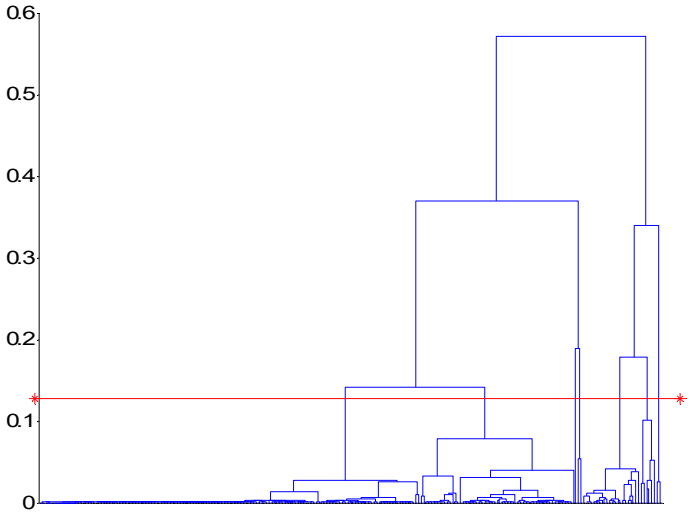

Figure 9: Dendrogram plotted using Euclidean distance measure

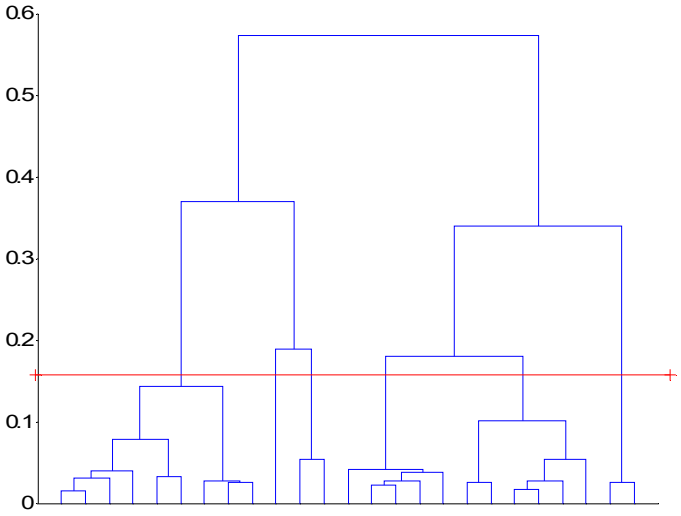

Figure 10: Dendrogram plotted using cosine distance measure

4. CONCLUSION

This paper contains description and demonstration of simple MATLAB-based *K*-means algorithm, used to iteratively re-assign clients to the nearest cluster center, with randomly selected *K* points as initial cluster center. In order to optimize the number of clusters choice, the hierarchical techniques are also used. Analysis given in this paper could help in attracting and keeping clients, and it also allows company to better understand its clients, the market in which they are active, their competitors, and other factors that can impact their own business and profitability.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[2] A. Agresti, Categorical Data Analysis, 2nd ed., Wiley, New York, 2002.

[3] T. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., Wiley, New York, 2003.

[4] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[5] A. K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recogn. Lett. 31, 2010, pp. 651-666.

[6] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Syst. Appl., 2013, pp. 200-210.

[7] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic", Journal of the Royal Statistical Society, 2001, Series B. 32(2) pp. 411–423.

[8] H. Stefanović, R. Veselinović, G. Bjelobaba, A. Savić, "An adaptive car number plate image segmentation using K-means clustering", Proceedings of Int. Scientific Conference on Information Technology and Data Related Research-SINTEZA 2018, 2018, pp. 74-78.

[9] Z. Wang, Z. Xu, S. Liu, and Z. Yao, "Direct clustering analysis based on intuitionistic fuzzy implication", Applied Soft Computing 23, 2014, pp. 1-8.

[10] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient kmeans clustering algorithm: Analysis and implementation", IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 2002, pp. 881–892.

[11] J. Nayak, D.P. Kanungo, B. Naik, and H.S. Behera, "Evolutionary improved swarm-based hybrid K-means algorithm for cluster analysis", Proceedings of Int. Conf. on Computer and Communication Technologies, Springer, New Delhi, 2016, pp. 343-352.

[12] C. McDaniel and R. Gates, Marketing Research, 10th ed, John Wiley & Sons, 2014.

[13] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications 36, 2009, pp. 4176–4184.

[14] https://www.mathworks.com/help/stats/dendrogram.html

[15] https://www.mathworks.com/help/stats/silhouette.html

[16] https://www.mathworks.com/help/stats/nearest-neighbors-1.html