Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **22** (2022) 031302 (552-559) AKU J. Sci. Eng. **22** (2022) 031302 (552-559)
DOI: 10.35414/akufemubid.1018302

Araştırma Makalesi / Research Article

Comparison of Hierarchic Clustering Methods with Cophenetic Correlation Coefficient in Big Data*

Sinan SARAÇLI[1], Murat AKŞİT[2]
[1] *Afyon Kocatepe Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Afyonkarahisar.*
[2] *Big Cat Research, İstanbul, Turkey.*

Sorumlu yazar e-posta: ssaracli@aku.edu.tr ORCID ID: http://orcid.org/0000-0003-4662-8031
e-posta: murat@bigcatresearch.com ORCID ID: http://orcid.org/0000-0002-1982-4849

Geliş Tarihi: 19.11.2021 Kabul Tarihi: 07.06.2022

Keywords
Cophenetic Correlation Coefficient; Big Data; Cluster Analysis; Data Mining

Abstract

The aim of this study is to compare hierarchical clustering methods by Cophenetic Correlation Coefficient (CCC) when there is a big data. For this purpose, after giving information about big data, clustering methods and CCC, analyzes are carried out for the related data set. The 2015 air travel consumer report, which was used in the application part of the study and published by the US Ministry of Transport, was used as big data. Libraries of the Python programming language installed on the Amazon cloud server, which includes open-source big data technologies, were used for data analysis. Since there is big data in the study, in order to save time and economy, the variables used in the study were first reduced by feature selection method, standardized and analyzed over the final 4 different data sets. As a result of the clustering analysis, it was observed that the highest CCC was obtained with the Average clustering method for all of these four different data sets.

Büyük Veride Hiyerarşik Kümeleme Yöntemlerinin Kofenetik Korelasyon Katsayısı ile Karşılaştırılması

Anahtar kelimeler
Kofenetik Korelasyon Katsayısı; Büyük Veri; Kümeleme Analizi; Veri Madenciliği

Öz

Bu çalışmanın amacı büyük veri söz konusu olduğunda hiyerarşik kümeleme yöntemlerini Kofenetik korelasyon katsayı ise karşılaştırmaktır. Bu amaçla büyük veri, kümeleme yöntemleri ve Kofenetik korelasyon katsayısı hakkında bilgiler verildikten sonra ele alınan veri seti için analizler gerçekleştirilmiştir. Çalışmanın uygulama kısmında kullanılan ve büyük veri olarak ABD ulaştırma bakanlığı tarafından yayınlanan 2015 yılı hava seyahat tüketici raporu kullanılmıştır. Veri analizi için açık kaynaklı büyük veri teknolojilerini içeren Amazon bulut sunucusuna kurulan Python programlama diline ait kütüphanelerden yararlanılmıştır. Çalışmada büyük veri söz konusu olduğundan, zamandan ve maliyetten tasarruf amacıyla çalışmada kullanılan değişkenler ilk olarak özellik seçimi yöntemi ile indirgenmiş, standardize edilmiş ve nihai 4 farklı veri seti üzerinden çözümlemeye gidilmiştir. Kümeleme analiz sonucunda bu dört farklı veri setinin tamamı için en yüksek Kofenetik korelasyon katsayısının ortalama bağlantı kümeleme yöntemi ile elde edildiği gözlemlenmiştir.

1. Introduction

Due to the increasing data volume and diversity, data has reached dimensions and differences that cannot be processed with traditional methods. Data, called as big data, is compiled in unconventional ways beyond what we are used to (such as wireless sensors, blogs, e-mail, social media, etc.) and in larger sizes beyond what is expected and from many different sources. According to the literature, it has been observed that the data types are homogeneous and do not come with a specific format. This is one of the main challenges faced by researchers dealing with data

* This study is a part of Murat Akşit's MS thesis, supervised by Sinan Saracli at Afyon Kocatepe University Institute of Science.

science. Other problems faced by researchers can be listed as follows; It is the need for large storage space and the need for a server with high hardware features. In order to meet these needs, there has been an increase in the hardware capacities of the computers and software diversity. In this way, big data technologies have emerged. Thanks to these technologies, large amounts of data can be processed in real time and practically. One of the most used methods in analysing big data is cluster analysis. Clustering analysis provides a better understanding of the data in the clusters created. Although the clustering method is the most frequently used method in the literature, no study has been encountered in which clustering methods in large data are compared with the Cophenetic Correlation Coefficient (CCC).

2. Literature review

2.1. Big Data

Researches examining and working on big data emphasized that there cannot be a single common definition on this subject, but different definitions can be made according to the area of use. According to Vinod (2013), big data is a concept that typically defines the size of data hundreds of times of Terabit or Petabit. According to Rubistein (2013), it argues that big data is in operational and application terms as "the use of confidential information and surprise correlations with statistics and data mining techniques by integrating different digital data sets of enterprises, government or organizations" (Demirtaş and Argan 2015). Big data consists of five components.

- Size of Data
- Speed of Data
- Diversity of Data
- Value of Data
- Verification of Data (Takcı and Aydemir 2018)

2.2 Cluster Analysis

Cluster analysis is one of the most significant data mining process to group objects according to their similarities and to obtain summary information about objects belonging to the same group through these groups (Yılmaz and Patır 2011).

2.2.1. Distance Criteria

Distance criteria is the distance between two unit is less than or equal to the sum of the distances of these two units to a third unit (Yılmaz and Patır 2011).

2.2.1.1. Canberra Distance

Canberra distance measure is a sensitive measure of distance for small points that take non-negative values and have a value close to zero (Kazaz 2019, Ziviani *et al*. 2004). The Canberra distance measure can be defined as a measure of absolute functional differences between the properties of a pair of data points. The Canberra distance is calculated with the formula given in equation 1:

$$d_{Canberra}(x_i, x_j) = \sum_{l=1}^{d} \frac{|x_{il} - x_{jl}|}{|x_{il}| + |x_{jl}|} \quad (1)$$

2.2.1.2. Euclidean Distance

Euclidean Distance is one of the significant and common classical measures of similarity used in various clustering algorithms such as K-means and hierarchical clustering. Euclidean distance can be defined as the distance between two points or vectors in Euclidean norm (Kumar and Toshniwal 2016). Euclidean distance is calculated with the formula given in equation 2:

$$d_{Euclidean}(T_1, T_2) = \sum_{j=1}^{n} \sqrt{(T_{1j} - T_{2j})^2} \quad (2)$$

2.2.1.3. Minkowski Distance

Minkowski distance is defined as a metric in a vector space that can be considered as a generalization of both Euclidean distance and Manhattan distance (Kumar and Toshniwal 2016). The Minkowski distance between two points T1 and T2 on the p order can be determined as T1 = (T11, T12, ..., T1n) and T2 = (T21, T22, ..., T2n). Minkowski distance is calculated with the formula given in equation 3:

$$d_{Minkowski}(T_1, T_2) = (\sum_{i=1}^{n} |T_{1i} - T_{2i}|^2)^{\frac{1}{p}} \quad (3)$$

2.2.2. Hierarchical Clustering Methods

Backer (1995) defines hierarchical clustering as a partition sequence in which each partition is nested to the next partition in the series. There are two

main types of hierarchical clustering; agglomerative and divisive. Agglomerative methods start with each object being a cluster and continue by gradually combining the two closest clusters until all objects form a single cluster; n is a series of successive fusions of objects in groups (Sakarya 2007).

2.2.2.1. The Single-linkage (SL)

The single-linkage method is the oldest model developed by Polish researchers in 1950s (Murtagh and Contreras 2012). It was first defined by Florek *et al.* (1951) and later by Sneath (1957) and Johnson (1967). The distance between two clusters (C1) and (C2∪C3) is defined as the minimum distance between any sample in a set and any other sample Everitt *et al.* (2011) and can be obtained by equation 4 (Carvalho *et al.* 2019).

$$d(C_1, C_2 \cup C_3) = min[d(C_1, C_2), (C_1, C_3)] \quad (4)$$

Additionally, this method tends to produce unbalanced and scattered clusters ("chained"), especially in large data sets. Thus, it does not consider the cluster structure (Everitt *et al.* 2011).

2.2.2.2. Complete-linkage (CL)

Complete- Linkage clustering method is similar to the single- link method except for the distance between the two clusters (C1) and (C2∪C3). It is described as the largest distance between pairs of samples in each set, rather than the smallest Everitt *et al.* (2011) Mardia *et al.* (1989) Carvalho *et al.* (2019) and can be obtained by equation 5:

$$d(C_1, C_2 \cup C_3) = max[d(C_1, C_2), (C_1, C_3)] \quad (5)$$

This method tends to find compact clusters of equal diameters (maximum distance between objects). It does not take into account the cluster structure (Everitt *et al.* 2011).

2.2.2.3. Average- Linkage (AL)

Average-linkage Clustering Method is also known as the unweighted pair group method using the average approximation (UPGMA). The distance between two sets is the average of the distance between all pairs of samples consisting of one sample from each group (Everitt *et al.* 2011). The distance between clusters is determined by the Lance-William correlation:

$$d(C_1, C_2 \cup C_3) = \frac{n_2.\mathrm{d}(C_1, C_2) + n_3.\mathrm{d}(C_1, C_3)}{n_2 + n_3} \quad (6)$$

n_2 and n_3 are the number of samples in cluster C_2 and C_3, respectively (Carvalho *et al.* 2019).

2.2.2.4. Ward Clustering Method

Ward-Clustering Method obtains new clusters by minimizing intra-cluster variance. Among these clusters, the cluster with the lower error square value is chosen (Çelik 2017). Ward clustering method is calculated as given in equation 7.

d= distance between two clusters
x= observation
n= number of data

$$= \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \quad (7)$$

2.3. Cophenetic Correlation Coefficient (CCC)

The CCC is a coefficient calculated to evaluate the fit between raw data distances and the distance measures used (Ponde *et al.* 2016). It is widely preferred to evaluate both an appropriate distance measure of data set classification and the efficiency of various clustering techniques (Saraçlı *et al.* 2013). The high CCC indicates that it is the most accurate clustering and distance criterion for the data set (Ponde *et al.* 2016). CCC is calculated as given in equation 8.

x (i, j) = |Xi – Xj |= Euclidean distance
t (i, j) = |Ti – Tj |= Dendrogram distance

$$c = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t}{\sqrt{\sum_{i<j}[x(i,j) - x)^2] \sum_{i<j}[t(i,j) - t)^2]}} \quad (8)$$

2.4. Feature Selection

Feature selection is an important set of algorithms used to achieve more consistent results by improving the correct classification rates or performances of the methods used in machine learning systems (Gazeloğlu 2020). Feature selection can be defined as the selection of the best

subset that can present the data set by selecting k features among n features from the data set (Budak 2018). Feature selection is used to determine the variables that will not affect the result in the data set before starting the analysis. This method is one of the first and significant steps in big data and data mining processes (Guyon and Elisseeff 2003).

The methods used in feature selection are composed of three groups as; Filtering methods Coiled Methods and Embedded Methods (Rong *et al.* 2019).

In this study related with the aim of the study, which emphaises the CCC, Correlation based future selection is considered.

2.4.1. Correlation Based Feature Selection

Correlation-based feature selection chooses on the basis of finding subsets of the data set that have the highest correlation coefficient and that contain different features (Emrah and Akın 2019). Correlation-based feature selection method is calculated as given in equation 9.

k= The numbers of features in subset
$\overline{r}_{ci}$= Average correlation between y and property
$\overline{r}_{ii}$= Average internal correlation of properties between each other

$$M_s = \frac{k\overline{r}_{ci}}{\sqrt{k + k(k-1)\overline{r}_{ii}}} \quad (9)$$

3. Material and Method

This study conducted cloud server service offered by Amazon to perform cluster analysis. Amazon cloud server service is a collection of web services that allow many developments for enterprise applications, big data projects and mobile applications to be developed in cloud infrastructure. Thus, Amazon elastic computing cloud was activated on this cloud service. Moreover, Amazon elastic computing cloud, is a cloud computing environment with a specific operating system, specific computing, storage and networking features that provides mechanisms for starting and managing virtual machines (Kokkinos *et al.* 2015). Another significant cloud computing environment used Amazon Elastic MapReduce (EMR) is built on this server. Amazon Elastic MapReduce (EMR) service is a data processing platform that includes open-source big data technologies such as Hadoop and Spark developed by Amazon and is used to process and manage data quickly. Python programming language is preferred to perform cluster analysis. Dask and Sicikit-Learn libraries were used in the study.

In this study, 2015 Air Travel Consumer Report data set is used, which was published by the US Department of Transportation. Thus, this data set published as free and open access. Before starting the clustering analysis, feature selection process was carried out in order to determine the variables that would not affect the result in the data set. This feature selection provided the opportunity to work with the data set containing more observations, since the variables that would not affect the result were removed from the data set. Correlation-based feature selection among filtering methods was preferred here.

As a result of the feature selection, a new data set was created by removing other variables from the data set. The multivariate normality assumptions of this data set were provided. Next, the variables were standardized because the units of the variables were different. For this purpose Z score standardization is applied to all data set.

Following that, 4 different data sets were created by random selection method to represent this data set and fit into the memory. The number of observations was randomly selected in all data sets. The second data set was created by removing the first data set from the total data set. This method has also been used in the creation of other data sets. In this way, different data sets were created as given in Table 1.

Table 1. Number of observations and variables belonging to the 4 selected clusters.

Cluster	*Number of observations*	*Variables*	*Airline Companies*
1st Data Set	10,859	*Taxi Entry Time *Taxi Departure Time *Wheel Closing Time	*United Airlines *American Airlines *US Airlines *Frontier Airlines *JetBlue Airlines
2nd Data Set	51,428	*Time between closing and opening the wheels *Distance	*Skywest Airlines *Alaska Airlines *Spirit Airlines *Southwest Airlines

3rd Data Set	72,553	*Total Delay Time *Flight Time	*Delta Airlines *AtlanticSoutheast Airlines *Hawaiian Airlines
4th Data Set	108,568		

After all these processes were completed, the clustering analysis process was started. Initially, the CCCs of the first data set were calculated. The clustering method and the distance criterion where the CCC get the highest value were determined. These processes were applied for all other data sets.

4. Results

For the 1st data set, CCC vas the greatest for Average-Linkage (AL) as the clustering method and Euclid distance as the distance criterion. Detailed results are given in Table 2.

Table 2. Cophenetic correlation coefficients for the 1st data set

Distance Criterion	*Clustering Method*	*Cophenetic correlation*
Euclidean	SL	0,577
Euclidean	CL	0,698
Euclidean	**AL**	**0,783**
Euclidean	Centroid	0,757
Euclidean	Ward	0,480
Canberra	SL	0,608
Canberra	CL	0,575
Canberra	AL	0,773
Minkowski	SL	0,577
Minkowski	CL	0,698
Euclidean	AL	0,577

When the clustering method is AL and the distance criterion is Euclidean, the dendrogram graph indicates that the Airline companies are divided into 3 clusters with 11 units of distance value. According to the results, UA (United Airlines) in a single cluster, AA (AmericanAirlines) and US (US Airways) in same cluster, F9 (FrontierAirlines), B6 (JetBlueAirlines), OO (SkywestAirlines), AS (Alaska Airlines), NK (SpiritAirlines), WN (SouthwestAirlines), DL (Delta Airlines), EV (AtlanticSoutheastAirlines) and HA (HawaiianAirlines) are colected in the other cluster as given in Figure 1.

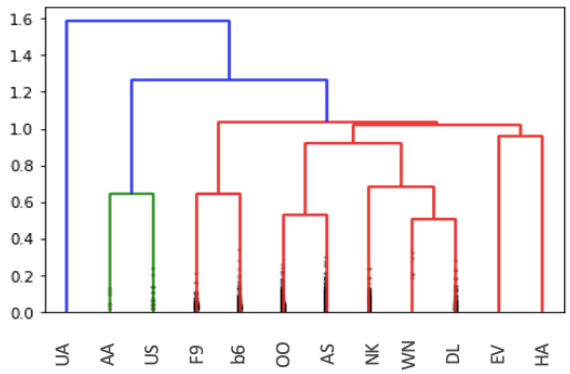

Figure 1. Dendrogram chart for Average-Linkage as Clustering method and Euclidean as distance criterion.

For the 2nd data set, CCC vas the greatest for AL as the clustering method and Canberra distance as the distance criterion. Detailed results are given in Table 3.

Table 3. Cophenetic correlation coefficients for the 2nd data set

Distance Criterion	*Clustering Method*	*Cophenetic correlation*
Euclidean	SL	0,524
Euclidean	CL	0,644
Euclidean	AL	0,753
Euclidean	Centroid	0,750
Euclidean	Ward	0,574
Canberra	SL	0,597
Canberra	CL	0,588
Canberra	**AL**	**0,764**
Minkowski	SL	0,524
Minkowski	CL	0,644
Minkowski	AL	0,751

When the clustering method is AL and the distance criterion is Canberra, the dendrogram graph indicates that the Airline companies are divided into 2 clusters with 6 units of distance value. When these clusters are examined, in the first cluster, UA (United Airlines), AA (AmericanAirlines), US (US Airways), F9 (FrontierAirlines) and B6 (JetBlueAirlines) take part together and in the second cluster, OO (SkywestAirlines), AS (Alaska Airlines), NK (SpiritAirlines), WN (SouthwestAirlines), DL (Delta Airlines), EV (AtlanticSoutheastAirlines) and HA (HawaiianAirlines) take part tohether as it can be seen from Figure 2.

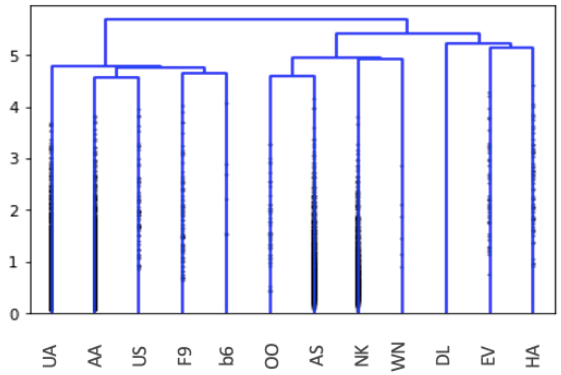

Figure 2. Dendrogram chart for Average-Linkage as Clustering method and Canberra as distance criterion.

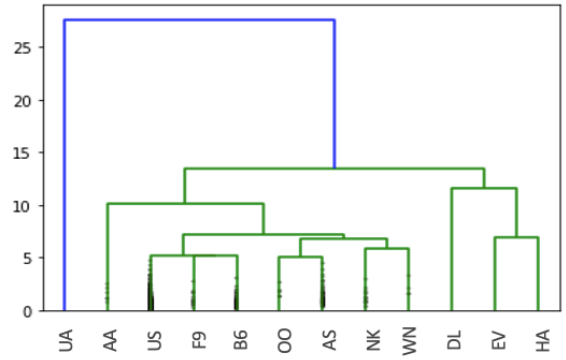

Figure 3. Dendrogram chart for Average-Linkage as Clustering method and Euclidean as distance criterion.

For the 3rd data set, CCC vas the greatest for AL as the clustering method and Euclidean distance as the distance criterion. Detailed results are given in Table 4.

Table 4. Cophenetic correlation coefficients in the 3rd data set

Distance Criterion	*Clustering Method*	*Cophenetic correlation coefficient*
Euclidean	SL	0,510
Euclidean	CL	0,671
Euclidean	**AL**	**0,774**
Euclidean	Centroid	0,765
Euclidean	Ward	0,542
Canberra	SL	0,612
Canberra	CL	0,554
Canberra	AL	0,768
Minkowski	SL	0,510
Minkowski	CL	0,671
Minkowski	AL	0,771

When the clustering method is AL and the distance criterion is Euclidean, the dendrogram graph indicates that the Airline companies are divided into 2 clusters with a distance value of 15 units. Results indicate that while UA is in a single cluster, other airlines AA, US, F9, B6, OO, AS, NK, WN, DL, EV and HA are in the other as it can be seen from Figure 3.

For the 4th data set, CCC vas the greatest for Centroid as the clustering method and Euclidean distance as the distance criterion. Detailed results are given in Table 5.

Table 5. Cophenetic correlation coefficients in the 4th data set

Distance Criterion	*Clustering Method*	*Cophenetic correlation coefficient*
Euclidean	SL	0,492
Euclidean	CL	0,717
Euclidean	AL	0,760
Euclidean	**Centroid**	**0,779**
Euclidean	Ward	0,465
Canberra	SL	0,579
Canberra	CL	0,555
Canberra	AL	0,768
Minkowski	SL	0,492
Minkowski	CL	0,717
Minkowski	AL	0,750

Dendrogram chart for Centroid as Clustering method and Euclidean as distance criterion given in Figure 4. indicate that, while UA, AA and US are in the same cluster, F9, B6, OO and AS are in the other cluster, NK and WN together in an other cluster and finally DL, EV and HA are in other cluster. As it can be seen from the figüre, there are 4 clusters fort his data set.

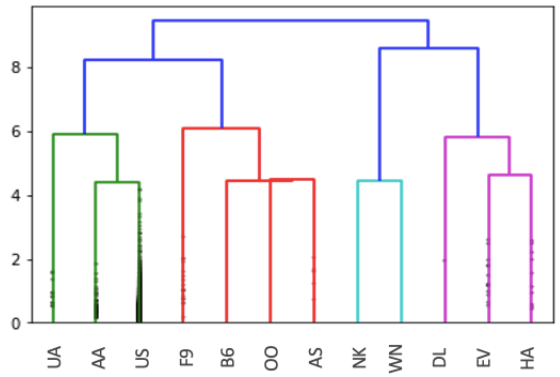

Figure 4. Dendrogram chart for Centroid as Clustering method and Euclidean as distance criterion.

5. Discussion and Conclusion

In this study, hierarchical clustering methods were compared with the CCC by using big data Technologies like Elastic computing cloud server, one of the cloud server services offered by Amazon. In order to facilitate the processing of big data on this server, Amazon Elastic Map Reduce (EMR) service, which includes open-source big data technologies such as Hadoop, Spark, were activated.

Moreover, Python libraries are used in clustering analysis. The first of these libraries, Dask is used to create a virtual server that can run in parallel on datasets that cannot fit in the main memory. The next one is the Sicikit-Learn library. Sicikit-Learn library is used to perform cluster analysis and calculate CCCs.

Before starting the cluster analysis, feature selection process was carried out in order to determine the variables that would not affect the result in the data set. Correlation-based feature selection, which is a sub-method of method filtering feature selection, was used. At this point, variables with a correlation coefficient less than 0.8 were removed from the data set.

Following that, 4 different data sets were created by random selection method by representing the main population from the data set. Clustering methods were applied by calculating distance criteria in each data set. Finally, the clustering method and the distance criterion were determined, where the CCC gave the highest value.

As a result of the study, it was observed that in the first data set, when the clustering method was average linkage and the distance criterion was Euclidian, the CCC gave the highest result. It was observed that in the 2nd data set, when the clustering method is average linkage again and the distance criterion is Canberra, the CCC gave the best result. Similar to first data set, for the 3rd data set, CCC vas the greatest for AL as the clustering method and Euclidean distance as the distance criterion. For the 4th data set, Centroid as Clustering method and Euclidean as distance criterion which gave the highest CCC, there were found four final clusters. The results of this stduy indicate that, Average-Linkage as Clustering method is giving the highest CCC for most of the clustering analysis.

Previous studies also indicate that (Silva and Dias 2013, Carvalho *et al.* 2019, Kumar and Toshniwal 2016, Ponde *et al.* 2016, Saraçlı *et al.* 2013), the CCC gave the highest result in the average linkage method.

We hope that this study, which was designed as considering the previous studies, will contribute to the literature as it was aimed at determining the best clustering method in big data by using big data Technologies.

İt has been also observed that one of the biggest problems for the researchers who apply clustering analysis in big data is, hardware deficiency. The solution for this problem can be recommended as using Amazon EMR, Python and Dask.

According to the findings obtained from the study, it is also recommended to use the average linkage method in different big data sets such as marketing, e-commerce, etc.

5. References

Backer, E., 1995. Computer-Assisted Reasoning in Cluster Analysis. Prentice Hall, Hertfordshire, 214.

Budak, H., 2018. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım: *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, **22**, 21-31.

Carvalho, P.R., Munita, C.S. and Lapolli, A.L., 2019. Validity Studies Among Hierarchical Methods of Cluster Analysis Using Cophenetic Correlation Coefficient. *Brazilian Journal of Radiation Sciences*, **7**, 1-14.

Çelik, S., 2017. Büyük Veri Teknolojilerinin İşletmeler İçin Önemi. *Social Sciences Studies Journal*, **3**(6), 873-883.

Demirtaş, B. and Argan, M., 2018. Büyük Veri ve Pazarlamadaki Dönüşüm: Kuramsal Bir Yaklaşım. *Pazarlama ve Pazarlama Araştırmaları Dergisi*, **8**(15), 1-21.

Emhan, Ö. and Akın, M., 2019. Filtreleme Tabanlı Öz Nitelik Seçme Yöntemlerinin Anomali Tabanlı Ağ Saldırısı Tespit Sistemlerine Etkisi. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, **10**(2), 549-559.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D., 2011. Cluster analysis. London: Edward, 73-169.

Florek, K., Lukaszewıez, L. and Perkal, L., 1951. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, **2**, 282-285.

Gazeloğlu, C., 2020. Prediction of heart disease by classifying with feature selection and machine learning methods. *Progress in Nutrition*, **22**(2), 660-670.

Guyon, I. and Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. The Journal of Machine Learning Research, **3**, 1157-1182.

Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Kazaz N.M.E., 2019. Veri Madenciliğinde Kümeleme Analizi Yöntemlerinin İncelenmesi Ve Sağlik Bilimleri Alanindaki Uygulamalari. Yüksek Lisans Tezi, İstanbul Ünversitesi, Sağlık Bilimleri Enstitüsü, İstanbul, 45.

Kokkinos, P., Varvarigou, T.A., Kretsis, A., Soumplis, P. and Varvarigos, E.A., 2015. SuMo: Analysis and Optimization of Amazon EC2 Instances. *J Grid Computing*, **13**, 255-274.

Kumar, C. and Toshniwal, D., 2016. Analysis Of Hourly Road Accident Counts Using Hierarchical Clustering and Cophenetic Correlation Coefficient (CPCC). *Journal Big Data*, **3**(13), 2–11.

Mardia, K.V., Kent, J.T. and Bibby, J. M., 1989. Multivariate Analysis. London: Academic Press, 385.

Murtagh, F. and Contreras, P., 2012. Methods of Hierarchical Clustering, Data Mining and Knowledge Discovery. *Wiley-Interscience*, **2**(1), 86-97.

Ponde, P., Shirwaikar, S. and Gore, S., 2016. Hierarchical Cluster Analysis on Security Design Patterns. *Association for Computing Machinery*, **92**, 1-6.

Rong, M., Gong, D. and Gao, X., 2019. Feature Selection and its Use in Big Data: Challenges, Methods, and Trends, *IEEE Access*, **7**, 19709–19725.

Rubistein, I.S., 2013. Big Data: The end of privacy or a new beginning? *International Data Privacy*, **3**(2), 74-86.

Sakarya, B., 2007. From Delphi to Scenario by Using Cluster Analysis: Turkish Foresight Case, Middle east technical university. Doctoral dissertation, 119.

Saraçlı, S., Dogan, N. and Dogan, I., 2013. Comparison of Hierarchical Cluster Analysis Methodsby Cophenetic Correlation. *Journal of inequalities and Applications*, **203**, 1-8.

Silva, A.R. and Dias, C.T.S., 2013. A cophenetic correlation coefficient for Tocher's method. *Pesquisa Agropecuária Brasileira*, **48**(6), 589-596.

Sneath, P.H.A., 1957. The application of computers to taxonomy. *J. General Microbiology*, **17**, 201-226.

Takcı, H. and Aydemir, N., 2018. Büyük Veri Yaklaşımıyla Birden Çok Bilgi Erişim Merkezinin Kolektif Kullanımı. *Bilişim Teknolojileri Dergisi*, **11**(2), 123–129.

Vinod, B., 2013. Leveraging big data for competitive advantage in travel. *Journal of revenue and pricing management*, **12**(1), 96-100.

Yılmaz, Ş. and Patır, S., 2011. Kümeleme Analizi ve Pazarlamada Kullanımı. *Akademik Yaklaşımlar Dergisi*, **2**(1), 91–113.

Ziviani, A., Fdida, S., Ezende, J.F. and Duarte, M.B., 2004. Toward a Measurement Based Geographic Location Service. *Lecture Notes in Computer Science*, **3015**, 43-52.