Marmara Üniversitesi Öneri Dergisi • Cilt 14, Sayı 52, Temmuz 2019, ISSN 1300-0845, ss. 356-371
DO1: 10.14783/maruoneri.594975

Araştırma Makalesi

CUSTOMER PORTFOLIO OF A CONSUMER GOODS BASED VIRTUAL STORE: IDENTIFYING CUSTOMER SEGMENTS WITH CLUSTER ANALYSIS*

TÜKETİM ÜRÜNLERİ BAZLI BİR SANAL MAĞAZA'NIN MÜŞTERİ PORTFOLYOSU: KÜMELEME ANALİZİ İLE MÜŞTERİ SEGMENTLERİNİN TESPİT EDİLMESİ

Begüm HATTATOĞLU**
Çağla ŞENELER***
Gökhan ŞAHİN *****
Fazlı YILDIRIM******

Abstract

In the last decade, analyzing and identifying customers became an irreplaceable need for companies. This research concentrates on discovering a company's customer segments using different machine learning algorithms, benchmarking different algorithms and its parameters to conclude the best results. Improvements in the technology provided several approaches to dive in and gain insights from a mass amount of data. Machine learning algorithms which is one of the most popular approaches was chosen to convey this empirical study. A dataset with mix categorical and numeric variables is analyzed with one of the conventional machine learning algorithms, namely Hierarchical Agglomerative Clustering Algorithm with Gower's distance. Kernel Principal Component Analysis is used

* Received: 09.12.2018; Accepted: 18.02.2019

** Yeditepe University, Department of Management Information Systems, ORCID ID: 0000-0002-8514-7778

*** Yeditepe University, Department of Management Information Systems, ORCID ID: 0000-0003-1817-9806

**** Yeditepe University, Department of Information Systems and Technologies, ORCID ID: 0000-0003-3980-8034

***** Yeditepe University, Department of Management Information Systems, ORCID ID: 0000-0002-8142-0466

for preprocessing due to the existence of categorical variables. K-prototypes Algorithm is chosen as benchmark algorithm that fits the qualities of the dataset with mixed categorical and numeric features. Benchmarking provides verification in respect to the accuracy of the results by evaluating the final clusters. Also, examining different parameters and comparing their effects on analysis results indicates the importance and vitality of them for machine learning algorithms, which need to be enlightened to do more accurate analyses. The results showed that both K-prototypes and HAC yield similar results proving that clusters mostly divided appropriately. However, there are a few significant points that are different at both algorithms' results, which should be examined in further study.

Keywords: Machine Learning, Kernel Principal Component Analysis, Clustering, Hierarchical Agglomerative Clustering, Gower's Distance, K-prototypes.

JEL Codes: C38

Öz

Son on yılda, müşterilerini analiz edip tanımlayabilmek, şirketler için vazgeçilmez bir ihtiyaç haline geldi. Bu araştırma, en iyi sonuçları elde etmek için farklı makine öğrenme algoritmaları kullanarak bir şirketin müşteri segmentlerini keşfetmeye, farklı algoritmaları ve parametrelerini karşılaştırmaya odaklanmıştır. Teknolojideki gelişmeler, çok büyük miktarda veriden yola çıkarak derinlemesine analizler yapmak ve iç görü kazanmak için çeşitli yaklaşımlar sağladı. Bu deneysel çalışmayı gerçekleştirmek için en popüler yaklaşımlardan biri olan makine öğrenme algoritmaları seçilmiştir. Karışık olarak kategorik ve sayısal değişkenlere sahip bir veri kümesi, geleneksel makine öğrenimi algoritmalarından biri olan Gower Uzaklığıyla birlikte Yığmacı Hiyerarşik Kümeleme Algoritması (HAC) ile analiz edildi. Kategorik değişkenlerin varlığı nedeniyle veri ön işlem süreci için Kernel Temel Bileşenler Analizi kullanıldı. Kprototypes Algoritması, veri kümesinin kategorik ve sayısal karakteristiği sebebiyle onun özelliklerine uyduğu için kıyaslama algoritması olarak seçilmiştir. Kıyaslama, sonuç kümelerini değerlendirerek elde edilen sonuçların doğruluğu konusunda bir doğrulama sağlar. Ayrıca, farklı parametrelerin incelenmesi ve bunların analiz sonuçları üzerindeki etkilerinin karşılaştırılması, daha doğru analizler yapmak için makine öğrenimi algoritmalarında parametrelerin önemli ve kritik olduğunu göstermektedir. Sonuçlar, hem Kprototipleri hem de HAC'nin, kümelerin çoğunlukla uygun şekilde bölünmüş olduğunu kanıtlayan benzer sonuçlar verdiğini göstermiştir. Ancak, her iki algoritmanın sonuçlarında bazı farklılıklar bulunmaktadır ve bunlar daha ileri çalışmalarda ele alınmalıdır.

Anahtar Kelimeler: Makine Öğrenimi, Kernel Temel Bileşenler Analizi, Kümeleme, Yığmacı Hiyerarşik Kümeleme, Gower Uzaklığı, K-prototypes Algoritması.

JEL Kodları: C38

1.INTRODUCTION

Today's business life experiences rapid changes at the technologies they engage every day. Since the world has entered the digital age, many things started to be done using technology such as handling business flows or tasks and its related data through computers and digital systems. With the development of technology, day to day transactions radically increased

and became more detailed compared to the past decades. Especially the usage of relational database with improved processing power of computers and its integration with other complex systems to accumulate mass amounts of data enabled companies to store the data and create an in-depth memory of the business of companies, which became a milestone for the emergence of data mining.

Increased competition among companies laid emphasis on the fact that if companies want to hold the competitive advantage in their hands, they must analyze the data they gathered to gain insights about what is currently happening in their business, and predict what might happen in the future based on the past, using data mining techniques. Muley and Anniruddha (2015) explain data mining as "an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data". The widest usage of data mining among companies is unraveling the hidden patterns in data with descriptive methods, or using data to forecast future direction of the business with predictive methods, and develop more accurate strategies while making decisions.

Since the markets became more competitive, it is harder to attract customers and sell the right product or service to them in order to increase profits, thus the importance of accurate market segmentation has been increasing steadily over the past years (Crabbe, Jones, & Vandebroek, 2011). That is why marketing became one of the most common fields that benefits from the descriptive analytics using data mining techniques to segment the customers. Descriptive analytics analyzes the historical data to discover and depict the underlying patterns in data, and helps to present the findings in an understandable way (Ghosh, 2017).

Many companies have done various marketing research using descriptive analytics to identify their customers and segment them to improve their marketing strategies alongside their Customer Relationship Management (CRM). Moreover, the research for CRM activities has started to be done intensively with data mining techniques in recent years (Kandeil, Saad, & Youssef, 2014). One of the most common data mining techniques used to segment customers, hence, to provide more personalized service and to improve CRM is clustering. Clustering is a technique to explore the groups (or clusters) underlying in the data by detecting and identifying the similarities of the members of the same clusters and the differences of dissimilar clusters (Oracle, 2003).

In the last decade, as the scope and functionality of online channels has improved and its usage has facilitated the purchases through online channels, it is quite commonly used by customers all around the world today. As a result, most of the businesses have not only physical stores but also have online/web stores. What is more, there are companies which have solely web store to sell their service or products and communicate their customers through online channels. Eventually, e-commerce market caused a massive population of

online customers which should also be segmented accurately to appeal to them in a customized way to maintain CRM and survive today's harsh competition. As a matter of course, the importance of online customer segmentation for these companies is clearly significant (Liu, Li, Peng, Lv, & Zhang, 2015). For these reasons, using clustering techniques to identify customer segments are vital for online businesses as well as the conventional businesses. This paper's purpose is based on identifying a company's customer segments using different machine learning algorithms, benchmarking the different algorithms to evaluate the accuracy of and its parameters to obtain the most accurate results.

The paper is organized as follows. Section 2 presents the methodology. Section 3 discusses estimation results. Section 4 concludes.

2.METHODOLOGY

The research is conducted based on a web store of a Turkish company which is exclusively open to its employees and the associated or partner companies' employees. This company is one of the biggest consumer products producer and distributer companies in Turkey, distributing different products ranging from napkins and tissues to canned food and sauces. Data is retrieved from both primary and secondary resources. The chart below is provided to envisage the process of the methods that will be used throughout the research.

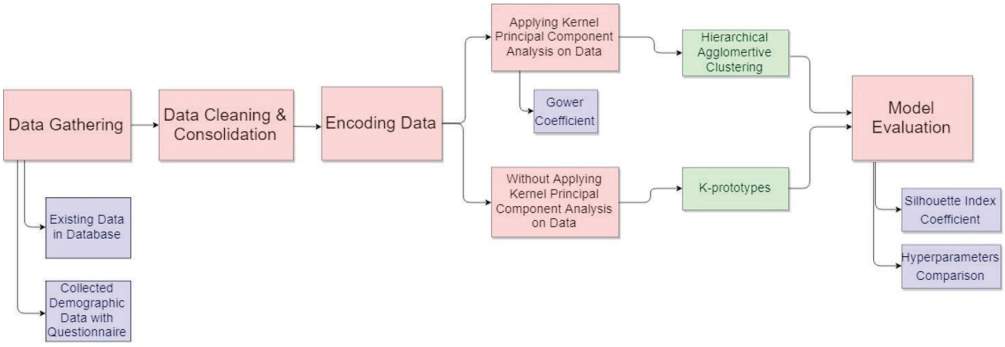

Figure 1: Process of the methods used throughout the research

2.1.Data Collection

There are two resources of data that is used for the analysis. First, current data in the database which contains sales records (each transaction, sold products, total amount of the purchase, the date of the transaction, and address information), general demographic information such as the date of birth, gender, e-mail address and given addresses by the customers.

Ad-hoc queries are used to extract and combine the data, then it is stored in excel sheets. The data obtained from database is also processed to achieve new fields from the raw data to gain more interpretable fields as input. The ages of customers are extracted from the date of births, average amount of purchase of each customer is extracted from the past order records, and the total number of children is found from the chosen age groups of each child by customers. These data are later used together with other raw data in the same input file.

The second source is the online questionnaire that is designed to collect more in-depth demographic information of customers. The information obtained from the questionnaire are the gender and date of birth to provide verification of previous data in the data warehouse, marital status, children data (if they are parents), education status, social media preference, and the favorite brand preference (that exists on the web store). The questionnaire was conducted in Turkish since the company is in Turkey and customers of the web store are Turkish. Questionnaire was sent to all customers that have membership of the web store and responded by 633 customers in total.

The sufficiency of the amount of data collected is decided according to the suggestion of Formann (1984) and Dolnicar (2002), which was stated as "the minimal sample size to include no less than 2^k cases (k=number of variables), preferably $5*2^k$". According to this calculation, minimum sample size requirement for this analysis is highly acceptable since the minimum sample size is $2^8 = 256$ while it is 633 samples for the study.

2.2.Instrumentation

The instrumentations that are used for the analysis are all based on Python language, which is used to code the algorithms. Spyder is used as an open source interactive development environment for Python language. There are several Python libraries used which are NumPy for scientific computation and n-dimensional array creation, SciPy for scientific computation which contains optimization, integration and linear algebra modules, and MatPlotLib for plotting and 2-D and 3-D visualization in various formats.

2.3.Data Analysis Methods

After collecting the data from the sources, the data is cleaned and consolidated for further process. Null value is assigned to the empty fields. The date of birth and gender data collected from the questionnaire is compared with the ones in database to consolidate whether the users are matching. Some of the users filled the questionnaire multiple times online. Thus, to eliminate multiplexing, the very last questionnaire filled and submitted by respondents is taken into the consideration and the previous submissions are deleted. Same identification hiding procedure is done to brand information. The identities of customers are hidden by assigning numeric MemberIDs. Date of birth field is replaced with age field. The ages

are separated into 5 distinct groups. Categorical variables are transformed into numerical variables, namely encoded, to make input more interpretable by the clustering algorithms. Clustering algorithms are unsupervised learning methods which do not use prior labels to learn the category tags of each instance in the dataset. These algorithms are used to obtain the segments, groups or clusters according to the measurable similarities and differences of features in the dataset (Jain, 2010). Encoded version of the dataset is used as the input of hierarchical agglomerative clustering algorithm after preprocessing. Since the dataset contains mostly categorical variables, the distance metric (for hierarchical agglomerative clustering) is chosen as Gower's similarity coefficient. Gower is a similarity measure for categorical, Boolean and numerical mixed data. This measure is specifically designed for clustering algorithms by J.C. Gower (Gower, 1971). The metric is calculated as below:

$$S_{ij} = \frac{\sum_{k=1}^{N} w_{ijk} s_{ijk}}{\sum_{k=1}^{N} w_{ijk}} \quad (1)$$

where w_{ijk} means the weight for variable k between observations i and j, s_{ijk} denotes the distance between i and j on variable k. Simply, this gives the weighted average of the distances on the different variables. The specialty of the calculation comes from the fact that equal/non equal calculation is used for categorical variables, while absolute difference is used for numeric variables. The coefficients distance is scaled in [0, 1] in order to prevent the different impacts of high scaled variables. For categorical variables, if i and j are equals, 1 is assigned, if not then 0 is assigned to scale those variables. For numeric variables, the absolute difference of the variables is divided to the range of variable to scale them. In mathematical notation, this means:

$$S_{ijk} = \frac{|x_{ik} - x_{jk}|}{r_k} \quad (2)$$

where r_k representa $\max(x_k) - \min(x_k)$ on the variable k (The explanation of Gower similarity is derived from the study of Jeroen van den Hoven (2015) and the journal of J.C. Gower (1971)).

After the dataset is encoded, next step is divided into 2 different approaches. In the first way, the whole clean dataset is used as an input without applying any pre-processing procedure. In the second way, since the dataset is not linearly structured and is categorical, Kernel's Principal Component Analysis (KPCA) is applied on it (PCA with the Kernel trick). PCA realizes orthogonal linear transformation to a coordinate system. The greatest variance of the dataset becomes the first principal component, and the second greatest variance becomes second principal component, respectively. These principal components becomes the new variables, containing more than 80% of the total information that previous dataset contains, without any redundancy. KPCA, which is used for non-linear systems, "…maps the input space into a feature space via non-linear mapping and then computes the principal components in that feature space" (Lee, Yoo, Choi, Vanrolleghem, & Lee, 2004). This analysis is

done because the dataset has 22 features (derived from 8 main characteristics) which results in high dimensionality that causes models to yield output more difficult. Thus, with Kernel PCA, the number of dimensions is reduced so that the clustering algorithms process the data easier.

In the first approach, Hierarchical Agglomerative Clustering (HAC) algorithm with ward linkage as a linkage metric and Gower's distance as a distance metric (because of the categorical variables) is applied on pre-processed data. It is a bottom-up approach, meaning that the algorithm starts by treating each instance as a cluster and then merges the clusters closest to each other step by step until obtaining one single cluster that contains all the instances. Also, the number of the clusters has to be specified before the HAC algorithm is run. This brings the need to try different number of clusters and assess which one is the best number of clusters for the results. While calculating the clusters, linkage metric is used as Ward's which minimizes the variance of clusters being merged. "Ward's is the only one among the agglomerative clustering methods that is based on a classical sum-of-squares criterion, producing groups that minimize within-group dispersion at each binary fusion" (Murgath & Legendre, 2014). The logic of the algorithm is explained by by Murgath and Legendre (2014) as the following:

"It was initially Wishart (1969) who wrote the Ward algorithm in terms of the Lance-Williams update formula. In Wishart (1969) the Lance-Williams formula is written in terms of squared dissimilarities, in a way that is formally identical to the following. Cluster update formula is:

$$\delta(i \cup i', i'') = \frac{w_i + w_i''}{w_i + w_{i'} + w_{i''}}\delta(i, i'') + \frac{w_{i'} + w_i''}{w_i + w_{i'} + w_{i''}}\delta(i', i'') - \frac{w_i''}{w_i + w_{i'} + w_{i''}}\delta(i, i') \quad (3)$$

with $w_{i \cup i'} = w_i + w_{i'}$."

In the equation, w_i resembles the cluster cardinality, while i, i', and i'' represents disjoint clusters (with cluster sizes $w_i, w_{i'}$, and $w_{i''}$). Since the Ward's method is based on minimum variance criterion, the initial cluster differences are defined with the squared Euclidean distance between points as below:

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2 d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2 \quad (4)$$

Also, other methods for linkage are used in order to benchmark the quality of outcomes and conclude the best linkage for the dataset.

In the second approach, k-prototypes algorithm is applied on dataset which is not encoded, since k-prototypes algorithm can handle both categorical and numerical variables together in a dataset. Because one of the foundational clustering algorithms, namely k-means, aims to minimize the average squared distance between instances (points) in the same cluster, it is not capable to handle categorical data. Hence, more improved version of k-means, which is k-prototypes, can

successfully handle mixed-typed variables and cluster more accurately. K-prototypes selects k-initial prototypes from the dataset, allocates every object O in a cluster whose prototype is the nearest to it, keeps retesting the similarity of objects against their assigned prototypes, reallocates the ones that are found that nearest to another cluster prototype, updates the changes of prototypes, and keeps doing it until there is no changed objects. The purpose of using two different clustering algorithms is to compare their performance and evaluate their results.

The results of the algorithms are evaluated with silhouette score analysis, which is used to determine how pure the cluster results of an algorithm are. This algorithm provides a graphical projection and the quality indexes, which indicates the strength of the clustering results (Rousseeuw, 1986). The Silhouette Coefficient is calculated by using the intra-cluster distance and the mean nearest-cluster distance for each sample. Silhouette scores can be between minus one and one (-1 to 1). The higher the silhouette score, the better the clustering is.

3.FINDINGS

After the dataset was processed through both pre-processing and learning with algorithm phases, it is found that there are several similarities and differences were detected in terms of both choice of the hyper-parameters and the type of kernel functions when the hierarchical agglomerative clustering was calculated. First of all, there is a significant difference between the types of kernel function and the derived results when their silhouette scores are compared, even if the hyper-parameters are chosen the same. Finding the type of kernel function that is compatible with the dataset is essential for high quality clustering. Another point that can be interpreted is when the algorithms are used with different hyper-parameters, both outputs yield the same optimal number of clusters according to their silhouette scores. It is observed that the results of rbf kernel based HAC algorithms with different hyper-parameters tend to behave in the same way. Also, since the cosine kernel function and its results with different hyper-parameters have the highest Silhouette Scores, the optimal number of clusters for this dataset can be concluded as 4 (with 2 Principal Components). The usage of manhattan and/or cosine as affinity is both considered reasonable and acceptable according to the results.

Another finding is that the number of principal components also affects the quality of the cluster. When the same clustering algorithm is executed with the same hyper-parameters and the same kernel function, the increase in the number of principal components decreases the clustering quality in this dataset. As can be seen, while the Silhouette Score of HAC with 3 Principal Components is 0.638, the score of HAC with 2 Principal Components is 0.684. Choosing two dimensions for this dataset as an input provides the best results. The reason why the results are spoiled is because the dimensionality is increased so the distribution of the data is changed and spread in space. As a result, it causes a difference in the number of clusters. The distribution of the dataset with 2 and 3 principal component in the space is given below:

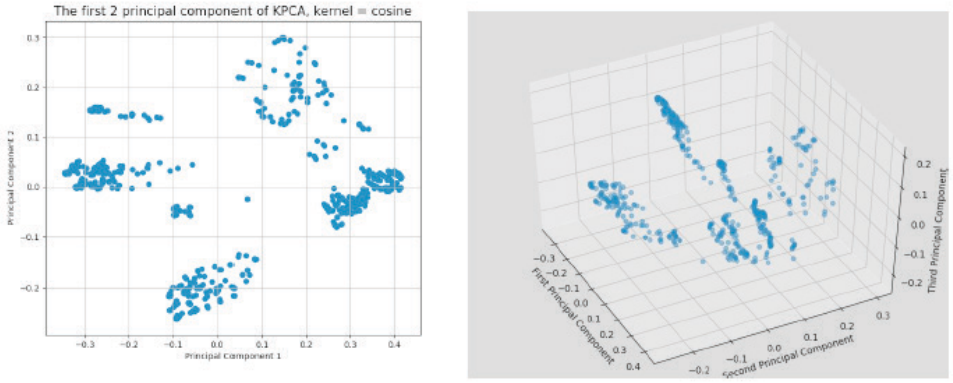

Figure 2: The distribution of input on 2D and 3D

The contribution of the input fields in the dataset to the clustering results is another important fact that needs to be examined. In the dataset, there is a field extracted from the questionnaire, which is the respondent customers' product category preference in the web store. It is recognized that this field randomizes the dataset so the results. The silhouette scores are compared when the input is provided with and without product category preference. It is found that there is a significant difference between the clusters yielded from these two different inputs. The results derived from the model that has the input without product category preference has better silhouette score. This fact is valid for other types of kernel functions and hyper-parameters as well.

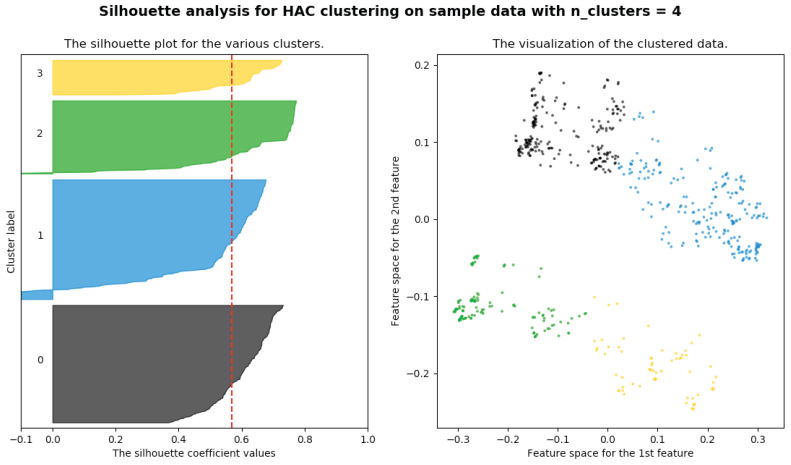

Figure 3: The visualization of HAC with cosine kernel function, cosine affinity and average linkage of the dataset with Product Category Preference

It is observed that the existence of product category preference field in the input dataset randomizes the distribution of the dataset on 2 dimensions and decreases the quality of the clustering results. It can be interpreted by the comparison of silhouette scores of the dataset with product category preference and without product category preference, when identical algorithms and hyper-parameters are chosen for both. The improvement on the results can be examined from the Figure 4.

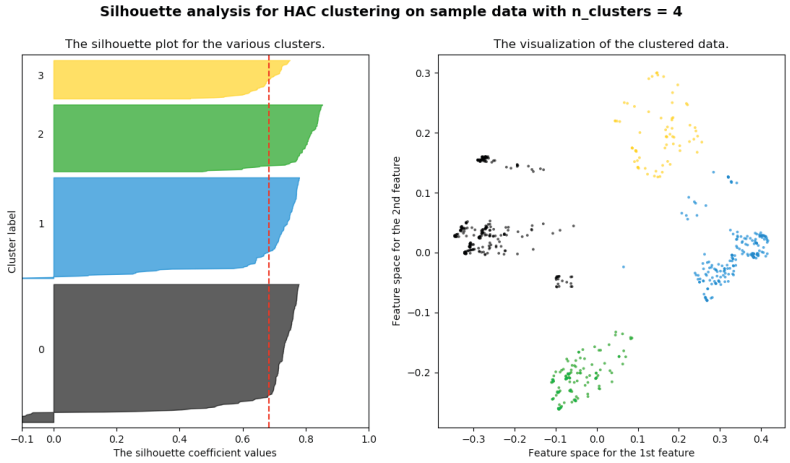

Figure 4: The visualization of HAC with cosine kernel function, cosine affinity and average linkage on the dataset without Product Category Preference

The results of silhouette scores also indicate that optimal number of clusters should be assessed carefully, and one analysis might not be enough for appropriate decision. In this study, while dendrogram graphics were indicating that the optimal number of clusters was 3, it was found that the optimal number was actually 4, when the silhouette scores of the 3-cluster algorithm and 4-cluster algorithm results are compared with the same metrics (see Table 1). It can be seen that the silhouette scores of 4-cluster algorithms in both with and without product category preference has the highest value among all other different amount of clustered algorithms' results. Consequently, it can be understood that multiple analyses should be realized to validate the optimal number of clusters in case the dataset is not fully conforming to the type of analysis.

Table 1: Silhouette Scores of two HAC algorithm results with and without product category preference using the same kernel and hyper-parameters

With Product Category Preference, KPCA(2 PC, kernel=cosine), affinity=cosine, linkage = average	**Without** Product Category Preference, KPCA(2 PC, kernel=cosine), affinity=cosine, linkage = average
For n_clusters = 3 The average silhouette_score is : 0.544394	For n_clusters = 3 The average silhouette_score is : 0.661451
For n_clusters = 4 The average silhouette_score is : 0.56955	For n_clusters = 4 The average silhouette_score is : **0.684267**
For n_clusters = 5 The average silhouette_score is : 0.533261	For n_clusters = 5 The average silhouette_score is : 0.624899
For n_clusters = 6 The average silhouette_score is : 0.540014	For n_clusters = 6 The average silhouette_score is : 0.608895
For n_clusters = 7 The average silhouette_score is : 0.514403	For n_clusters = 7 The average silhouette_score is : 0.627021

Note: 2 PC: Principal Components with 2 dimensions, n_clusters: number of clusters

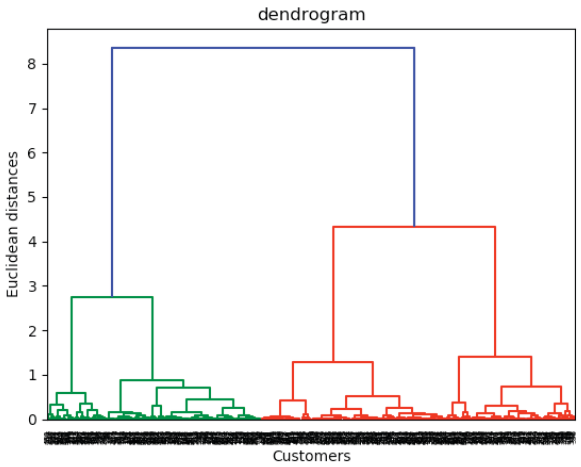

Figure 5: Dendrogram of HAC with cosine function in KPCA (with 2 dimensions), calculation method is chosen average, input is given without product category preference

Since we conclude that the optimal results are yielded from the HAC algorithms without the product category preference field, using cosine kernel function and cosine or manhattan distance affinity with average or complete linkage, the structure of the clusters are examined. According to the results, the clusters of both models (HAC and K-prototypes) are mainly divided according to the gender, marital status, age group and especially the social media habit. The first cluster has 253 customers who are young (age groups are 0 and 1, which correspond to the ages between 20 and 39) married females with

higher education without any children or with a child between 1-6 years old. This cluster has 100% Instagram usage habit as social media preference. Second cluster consists of 185 young (age group is dominantly 1, which corresponds to the ages between 30-39) and mostly married males with higher education who have at least a child in the most of the cases. This cluster of customers is mostly using Facebook, but a remarkable amount of the customers also use other social media channels as well. Third cluster has 123 people, 100% married males with higher education, who are in young-middle age group (age group 1-2 which corresponds between ages 30-49) that mostly has a child mostly older than 1 year old. This cluster preferred Twitter 100%. The fourth and the last cluster consists of only 71 females which are mostly married and has higher education. The members of this category are in young-middle age category as well. In this cluster, the customers either have no children, or has a child at least 6 years old. This cluster's social media preference is dominantly Facebook, while a respectable group of customers also prefer other social media channels which are not given. It is noticed that the average amount of spending does not indicate any significant difference among clusters. The price level of average spending fee in each cluster stayed approximately in the same level with insignificant differences (between 10-30 Turkish liras change detected in every model result).

Lastly, k-prototypes algorithm's cluster labels are examined. The centroid indicator is chosen according to Huang's proposal (Huang, 1997). However, since k-prototypes calculates the cluster centroids with both categorical and numeric variables, it has an output centroid array with both categorical and numeric values. Silhouette score is calculated with intra-cluster distance means, which are numerical, it cannot display the results of k-prototypes algorithm. Still, there are some indicators that help to interpret the results' quality in a reasonable sense. Clustering cost, which is the sum distances of all points to their respective cluster centroids. As it becomes smaller, the distances also become shorter; this indicates a tight cluster. K-prototypes algorithm is also executed with and without the product category preference to see how this input field affects the clustering cost. It is observed that the dataset without product category preference has lower clustering cost, so more accurately separated clusters. When the results of k-prototypes and HAC algorithms' results are compared, there are several similarities at cluster features: the optimal number of clusters are is the same, which is 4, The clusters are mainly based on the gender, 2 female 2 male clusters just like in the HAC, and with mostly young population from age groups 1 and 2, All the cluster members have higher education dominantly just like in the HAC cluster labels. According to the result of the k-prototypes with product category preference, the most dominant product category almost in all clusters is "dagitim", which means general distribution products of the company. However, we cannot examine its validity because it cannot be calculated a quality score for k-prototypes in this study. Even though there are several similarities, there is one contradicting result interpreted from

cluster centroids of k-prototypes results, which is all clusters have Instagram as centroid component which does not comply with HAC results. The problem cited above will be investigated in detail in the future study.

4.CONCLUSION

A dataset with mix categorical and numeric variables is analyzed with Hierarchical Agglomerative Clustering Algorithm with Gower's distance and Kernel Principal Component Analysis and K-prototypes Algorithm. When the cosine parameter is assigned to determine the type of kernel function yielded the best result for the research dataset. HAC Algorithm is optimized when its affinity is assigned to cosine with average linkage and manhattan affinity with complete linkage.

The number of clusters in this dataset is found to be 4. This is supported with both Silhouette Score and k-prototypes algorithm's results. The clusters are heavily based on gender and social media preferences, with the support of age category and children information. Since k-prototypes algorithm results could not be thoroughly analyzed due to the conditions, it is left to future study to figure out the patterns with Multidimensional Scaling or approach to dataset with more improved techniques such as latent class analysis. It is clearly seen from the results that choosing the most appropriate algorithm and hyper-parameter selection plays a crucial role on the quality of the analysis results. In order to obtain successful results, suitable input types for the algorithms and their logical structure should be known and the input dataset should be analyzed in detail to choose the most suitable algorithms with corresponding hyper-parameters that work the best with the dataset. Since the results of the analysis will be used for decision-making of a customer segmentation strategy in a company, choosing the correct algorithm and its corresponding hyper-parameters appropriate to the dataset play a critical role on the success of the analysis.

If the company wants to make advertisements or marketing campaigns based on its distinct customer segments, it can focus better by using specific social media platforms that customer segments prefer to reach them more successfully. For example, when the company wants to make advertisement about its baby care products, it can use Instagram to target young married females with children (0-6 ages) by opening an account or posting advertisements that can be seen on users' homepages. With this strategy, company can achieve low-cost delivery of more personalized advertisements on online platforms thanks to accurate targeting based on customer segments (Canhoto, Clark, & Fennemore, 2013).

As the usage of social media platforms continuously increases and becomes more popular, these platforms became one of the most important channels that companies reach their customers and create two-way communication (Kim & Ko, 2012). That is why identifying customer segments with social media preferences is useful for companies. There are several

studies that examines distinct customer segments among customers based on surveys and doing cluster analysis that are used in decision making strategies regarding marketing, advertisement or similar purposes. Amaro, Duarte and Henriques (2016) had a similar approach by using cluster analysis to identify distinct customer segments according to customers' social media characteristics in order to adapt online marketing strategies for travelers. Moreover, Vinerean, Cetina, Dumitrescu and Tichindelan (2013) used cluster analysis to make segmentation based on the responses regarding social media habits among different types of participants in their empirical study.

It is observed that the characteristics, variety, and the amount of population are important while identifying the customer segments. Having similar features at whole population complicates the ability to categorize customers in highly discrete segments for the conventional algorithms. The community consists of people mostly with higher education, who are in dynamic workforce age, and mostly married. It is harder for algorithms to distinct variables that show similar or common qualities frequently. This is why mostly gender and social media preference became the core indicator of different clusters in algorithms. The last fact that should be considered while interpreting the segmentation results is the product category and the variety that the company provides to its customers.

References

Amaro, S., Duarte, P. & Henriques, C. (2016). Travelers' use of social media: A clustering approach. *Annals of Tourism Research,* 59, 1-15.

Canhoto, A. I., Clark, M. & Fennemore, P. (2013). Emerging segmentation practices in the age of the social customer. *Journal of Strategic Marketing,* 21(5), 413-428.

Crabbe, M., Jones, B. & Vandebroek, M. (2011). *A comparison of two-stage segmentation methods for choice-based conjoint data: A simulation study.* Leuven: K.U. Leuven Faculty of Business and Economics.

Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation. *CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002 (ANZMAC 2002)*, (4-6), Melbourne, Retrieved December 2-4, 2002

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in Theorie und Anwendung.* Weinheim: Beltz.

Ghosh, P. (2017). *Fundamentals of Descriptive Analytics. Retrieved from Dataversity.* Retrieved from Data Education for Business and IT Professionals: http://www.dataversity.net/fundamentals-descriptive-analytics/

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics,* 27(4), 857-871. doi:10.2307/2528823

Hoven, J. V. (2015). *Clustering with Optimised Weights for Gower's Metric.* Amsterdam: University of Amsterdam.

Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, 21-34.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters,* 31(8), 651-666.

Kandeil, D. A., Saad, A. A. & Youssef, S. M. (2014). A two-phase clustering analysis for B2B customer segmentation. *International Conference on Intelligent Networking and Collaborative Systems* (221-228), Salerno: IEEE. doi:10.1109/INCoS.2014.49

Kim, A. J. & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research,* 65(10), 1480-1486.

Lee, J.-M., Yoo, C., Choi, W. S., Vanrolleghem, P. A. & Lee, I.-B. (2004). Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science,* 59(1), 223-234.

Liu, Y., Li, H., Peng, G., Lv, B. & Zhang, C. (2015). Online purchaser segmentation and promotion strategy selection: Evidence from Chinese E-commerce market. *Annals of Operations Research,* 233(1), 263-279. doi:https://doi.org/10.1007/s10479.013.1443-z

Muley, P. & Joshi, A. (2015). Application of data mining techniques for customer segmentation in real time business intelligence. *International Journal of Innovative Research in Advanced Engineering (IJIRAE),* 2(4), 106-109.

Murgath, F. & Legendre, P. (2014, October). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification,* 31(3), 274-295. doi:https://doi.org/10.1007/s00357.014.9161-z

Oracle. (2003). Descriptive Data Mining Models. In *In Oracle Data Mining Concepts 10g Release 1* (10.1) (pp. 4-1). Oracle.

Rousseeuw, P. J. (1986). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Vinerean, S., Cetina, I., Dumitrescu, L. & Tichindelean, M. (2013). The effects of social media marketing on online consumer behavior. *International Journal of Business and Management,* 8(14), 66-79.

Begüm HATTATOĞLU – hattatbegum@gmail.com

Received her B.A. degree in Management Information Systems as a top rank student and completed a minor degree in International Trade and Business at Yeditepe University. She is currently working as a Risk Assurance Consultant at PwC Turkey. Her research interests are data analytics, machine learning and process mining.

Çağla ŞENELER – cagla.seneler@yeditepe.edu.tr

Asst. Prof. Cagla Seneler is one of the first academics who graduated from Management Information Systems (MIS) undergraduate programmes in Turkey. She completed her BS and MA degrees in MIS Department of Boğaziçi University. She got her PhD from Computer Science Department of University of York, UK. She has been giving lectures in several universities in Turkey since 2004. She has more than 30 publications that are cited by many academics and one of her publications awarded as *Outstanding Paper Award Winner at the Literati Network Awards for Excellence 2011***.** Her research interests are human-computer interaction, user experience, learning styles, cultural differences, personalization, user interface characteristics, technology adoption and IoT.

Gökhan ŞAHİN – sahin@yeditepe.edu.tr

Received his B.Sc., M.Sc. and PhD degrees in physics at Boğaziçi University. He is currently an Assistant Professor at Yeditepe University. He teaches Applied Statistics, programming languages (C, C++, Java, Python) and web programming (php, .NET, Java servlets,JSF). His research interests include chaos, non-linear dynamics, statistical physics, machine learning, statistical properties of natural languages, metastable systems.

Fazlı YILDIRIM – fazli.yildirim@yeditepe.edu.tr

Assoc. Prof. Fazlı Yıldırım graduated from Istanbul University Computer Engineering Department in 1999. He received his MBA degree (2002) from Yeditepe University, Institute of Social Sciences and PhD in Business Administration from Işık University. In 2017, he received the title of Associate Professor in the field of Management Information Systems. He has many published articles in the field of electronic commerce, digital marketing, customer relations management, e-government and mobile applications and he has published a book about customer relationship management. He works at Yeditepe University in Electronic Commerce Department.