2017 3rd International Conference on Electrical Information and Communication Technology (EICT), 7-9 December 2017, Khulna, Bangladesh

Customer Segmentation using Centroid Based and Density Based Clustering Algorithms

A. S. M. Shahadat Hossain
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi - 6204, Bangladesh
shahadat.ruet.cse@gmail.com

***Abstract*—In recent years, customer segmentation has become one of the most significant and useful tools for e-commerce. It plays a vital role in online product recommendation system and also helps to understand local and global wholesale or retail market. Customer segmentation refers to grouping customers into different categories based on shared characteristics such as age, location, spending habit and so on. Similarly, clustering means putting things together in such a way that similar type of things remain in the same group. Due to having similarities between these two terms, it is possible to apply clustering algorithms for ensuring satisfactory and automatic customer segmentation. Among different types of clustering algorithms, centroid based and density based are the most popular. This paper illustrates the idea of applying density based algorithms for customer segmentation beside using centroid based algorithms like k-means. Applying DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm as one of the density based algorithms results in a meaningful customer segmentation.**

***Keywords*—Customer segmentation, Market segmentation, Clustering algorithms, Centroid based clustering, Density based clustering.**

I. Introduction

The concept of ‘Customer Segmentation’ which is also alternatively known as ‘Market Segmentation’ was introduced by Smith in 1956. It was stated as “Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing preferences, attributable to the desires of consumers for more precise satisfaction of their varying wants” [1]. To ensure customer satisfaction and optimal profit, customer segmentation helps to study and understand behaviors of customers. Actually, it is done by analyzing different types of information about customers.

Customer segmentation can be done focusing on different aspects such as demographic, geographic, behavioral and so on. Among them, this paper concentrates mostly on behavioral perspective as it is the most effective and practical one. Spending habit can be considered as one of the behavioral instances of customers which varies from one to another.

Clustering means partitioning a set of data in a set of groups of similar data. Clustering algorithms refer to those machine learning algorithms which are associated with unlabeled data. Among all other types of clustering algorithms, centroid based and density based algorithms are the most popular two types. In ‘Centroid Based’ clustering, clustering is done based on some randomly initialized points and minimum distance from a point to others. On the other hand, in ‘Density Based’ clustering, points are clustered based on their densities in a particular region.

In spite of the existence of a few works related to cluster based customer segmentation, no one of them has considered applying any density based clustering algorithm. Therefore, this paper implements DBSCAN algorithm as one of the density based algorithms while applying k-means with different distance metrics as a centroid based clustering algorithm.

The rest of the paper is organized as follows: section II discusses related work, section III gives a brief overview of the theoretical terms used in this paper and section IV presents the implementation. Besides, section V discusses on the experimental results. Finally, section VI concludes focusing on future work.

II. Related Work

Namvar et al. [2] introduced a new customer segmentation method consisting of two phase clustering. They showed that combining demographic data with two phase clustering results in a relatively better clustering. Hruschka and Natter [3] compared performances of feedforward Neural Network and k-means algorithms for cluster based market segmentation and found that cluster analysis done by Neural Network is better than of k-means clustering. Besides, Wu and Lin [4] studied cluster based customer segmentation model, Lee and Park [5] surveyed customer satisfaction for satisfactory customer segmentation and Teichert et al. [6] studied a specific case of customer segmentation in airline industry while customer segmentation in case of banking selection was presented by Anderson et al. [7]. How data driven customer segmentation is done in case of tourism was shown by Dolnicar [8].

III. Background Study

The theoretical terms behind this work such as customer segmentation, clustering, centroid based algorithms, density based algorithms etc. are briefly discussed here:

978-1-5386-2307-7/17/$31.00 ©2017 IEEE

Fig. 1: Customer Segmentation (Source: Adapted from [9])

A. Customer Segmentation

Customer Segmentation is a process to divide customers of a consumer or business market in groups based on some shared characteristics. There are various aspects on which customer segmentation can be done such as demographic, behavioral, geographic and so on; however, among them, customer segmentation based on user behavior is the most effective one as it considers attitudes, opinions, interests and some related criteria of the customers. Focusing on spending habits of the customers and few other behavioral aspects is expected to provide a relatively better customer segmentation. Figure 1 illustrates the concept of customer segmentation.

B. Clustering

Cluster analysis or clustering is a general task to group a set of objects in such a way that same type of objects remain in same group. It is one of the main aspects of data mining and a common technique for statistical data analysis used in different fields including machine learning, pattern recognition, image analysis, bioinformatics, data compression and so on. There are mainly four types of clustering algorithms- centroid based, density based, hierarchy based and distribution based clustering algorithms. Among them, centroid based algorithms work using some randomly initialized points called 'centroid' while density based algorithms consider the density of points in a region, hierarchy based algorithms work assuming that the nearby objects are more similar and distribution based algorithms work based on statistical distribution models.

1) Centroid Based Clustering: In centroid based algorithms, a predefined number of points are selected randomly which are called 'Centroid' where that predefined number denotes the number of expected clusters. After that, each point is assigned into a cluster based on the closest centroid. Then, position of centroids are changed to the mean position of their respective clusters. Finally, this process goes on iteratively until it finds expected number of meaningful clusters. A number of centroid based algorithms have been proposed to date. K-means, k-medoids, fuzzy c-means are some of them; however, k-means is the most popular among them.

a) K-means: K-means is a popular way to do vector quantization. Target of this algorithm is to group n objects into k clusters. Let $X = \{x_i\}, i = 1, 2, 3, ..., n$ be the set of d-dimensional points and $C = \{C_k, k = 1, 2, 3, ..., k\}$ where C is set of clusters. Let μ_k be the centroid of cluster c_k. The squared error between μ_k and the points in cluster C_k is to be minimized according to Equation 1 as follows:

$$J(C_k) = \sum_{X_i \epsilon C_k} ||X_i - \mu_k||^2 \tag{1}$$

According to Equation 2, actual objective of k-means is to minimize the sum of the squared error over all k clusters.

$$J(C_k) = \sum_{k=1}^{K} \sum_{X_i \epsilon C_k} ||X_i - \mu_k||^2 \tag{2}$$

k-means works as follows:

1) Initializes k number of centroids randomly within the data domain.
2) Clusters the data into k groups by assigning each data point to its closest centroid based on the distance between them.
3) Calculates the mean of all objects in each cluster and moves the centroid to that position.
4) Repeats steps 2 and 3 until the same points are assigned to each cluster in consecutive iteration.

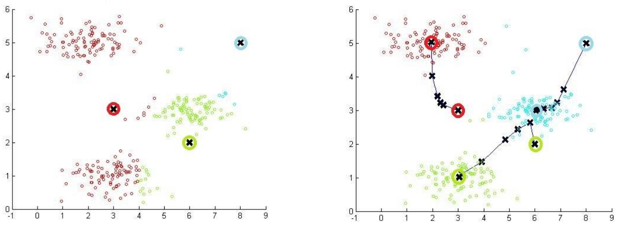
Any of the several distance metrics [10] can be used to calculate the minimum distance between any two points such as: Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance etc.

Distance between two n-dimensional vectors X and Y can be calculated using Euclidean distance according to Equation 3 and Manhattan distance according to Equation 4.

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{3}$$

$$D(X, Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{4}$$

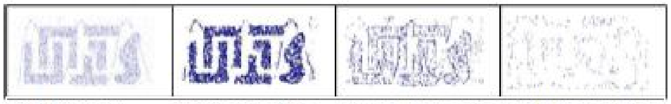
Figure 2 illustrates an example of initial and final positions of centroids for k-means.



(a) Initial position

(b) Final position

Fig. 2: Initial and final position of centroids during k-means



(a) All points (b) Core points (c) Border points (d) Noise points

Fig. 3: Points identified by density based clustering algorithms (Source: Adapted from [11])

2) Density Based Clustering: Density-based clustering means creating clusters having different densities. Density refers to the number of data points within a specific region around a point P. This region is generally a N-sphere with radius ϵ which represents the maximum distance between any data point and P. Usually, these algorithms consider data points with lower density as noise or border points as shown in Figure 3. DBSCAN and OPTICS are two most commonly used density based clustering algorithms. There is a key drawback of this type of algorithms that they expect some kind of density drop to detect cluster borders.

a) DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) discovers clusters with arbitrary shape based on the ϵ neighborhood principle which means that points within ϵ radius are neighbors. It was first proposed by Ester et al. in 1996 [12]. Size of a cluster can be controlled by a parameter called 'MinPts'(number of minimum points to be considered as a cluster) . DBSCAN outperforms many other centroid based clustering algorithms in case of noisy datasets.

DBSCAN algorithm works as follows:

1) Categorizes each point as either core point, border point or noise point.
2) Eliminates noise.
3) Makes new clusters with unclustered core points by assigning each point into a cluster based on neighborhood.

Time complexity of this algorithm is $O(n\log n)$ where, n denotes the number of total points in dataset to cluster.

b) OPTICS: Ordering Points To Identify the Clustering Structure (OPTICS) is an algorithm presented by Ankerst et al. in 1999 [13]. Working principle of this algorithm is similar to DBSCAN algorithm. But it outperforms DBSCAN in case of data with varying density. It finds intrinsic clustering structure by producing a special order of the database with respect to its density based clustering structure. Unlike DBSCAN, it puts more importance on choosing number of minimum points to form a cluster instead of fixing the radius. It makes a priority heap of data points to keep nearest points in neighbor list while number of minimum points roughly controls the size of cluster. Thus, this algorithm ensures better clustering. Besides, this algorithm provides flexibility of controlling the cluster size.

IV. Implementation

As it is described earlier that this paper implements customer segmentation using two types of clustering algorithms, implementation is done following the given steps:

A. Experimental Setup

Implementation is done on Intel Core i5 CPU which has a processor speed of 3.2 GHz with a 4 GB RAM. WEKA (Waikato Environment for Knowledge Analysis) [14] and scikit-learn [15] are used. WEKA is developed by University of Waikato, New Zealand as a suite of machine learning software written in Java while scikit-learn is a machine learning library for Python programming language.

B. Dataset

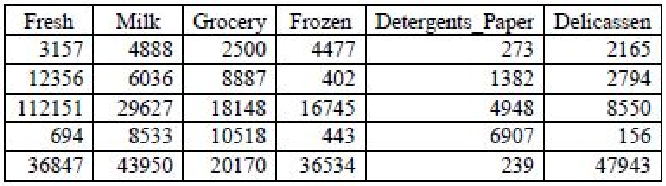
This paper uses Wholesale customers dataset [16] from University of California Irvine machine learning repository for customer segmentation. That dataset refers to clients of a wholesale distributor and contains data regarding consumption of different items by customers showing their annual spending. There are 440 instances and 8 attributes in total. Each instance in the dataset represents a customer while the attributes show each customer's annual spending on different commodities. To ease the implementation, two attributes from the dataset named as 'Channel' and 'Region' are excluded since they are not that much related to the spending habits of customers. Moreover, interest of this paper lies in considering the spending habits of customers rather than anything else for segmentation. Hence, this paper will be using the other 6 attributes on which the spending of customers are recorded. According to the University of California Irvine machine learning repository, this dataset is originated from another larger dataset referred in a paper [17]. Table 1 shows few samples of the dataset used.

C. Customer Segmentation using Centroid Based Algorithm

In this paper, k-means has been chosen among all centroid based algorithms for implementation as it is considered as the most popular one of its type. Before applying on dataset, K-means needs to decide the value of k and the distance metric to use. The appropriate number of total clusters, k can be estimated using a few methods [18]. This paper considers 2 to 5 clusters of customers. Euclidean distance and Manhattan distance are used here as distance metrics for k-means though finally they do not show that much differences.

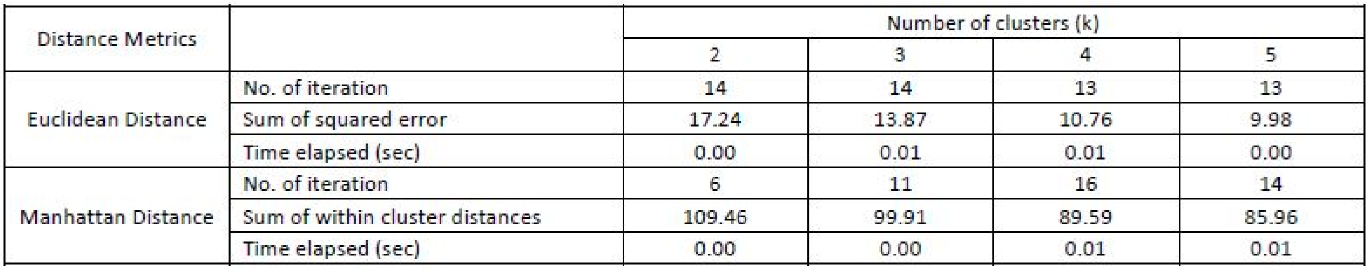
To scale all the values within a range between -1 and 1, 'Feature Normalization' process is applied.

Table I: Few samples of the dataset used



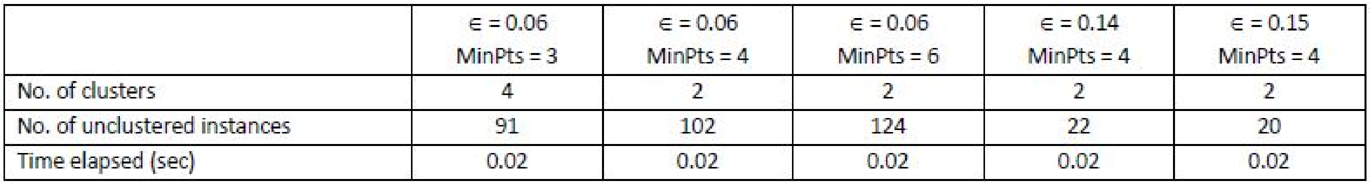
Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
3157	4888	2500	4477	273	2165
12356	6036	8887	402	1382	2794
112151	29627	18148	16745	4948	8550
694	8533	10518	443	6907	156
36847	43950	20170	36534	239	47943

Table II: Results on applying k-means with different distance metrics and different values of k



Distance Metrics		Number of clusters (k)			
		2	3	4	5
Euclidean Distance	No. of iteration	14	14	13	13
	Sum of squared error	17.24	13.87	10.76	9.98
	Time elapsed (sec)	0.00	0.01	0.01	0.00
Manhattan Distance	No. of iteration	6	11	16	14
	Sum of within cluster distances	109.46	99.91	89.59	85.96
	Time elapsed (sec)	0.00	0.00	0.01	0.01

Table III: Results on applying DBSCAN with different radius of neighborhoods (ϵ) and MinPts



	∈ = 0.06 MinPts = 3	∈ = 0.06 MinPts = 4	∈ = 0.06 MinPts = 6	∈ = 0.14 MinPts = 4	∈ = 0.15 MinPts = 4
No. of clusters	4	2	2	2	2
No. of unclustered instances	91	102	124	22	20
Time elapsed (sec)	0.02	0.02	0.02	0.02	0.02

Feature Normalization is done according to the following Equation 5:

$$x' = \frac{x - \bar{x}}{x_{max} - x_{min}} \quad (5)$$

where, x is the original value, $\bar{x}$ is the mean, x_{max} is the maximum value, x_{min} is the minimum value and x' is the normalized value.

Finally, k-means is applied on the dataset varying the value of k roughly and the experimental results found for two different distance metrics are shown in Table II.

D. Customer Segmentation using Density Based Algorithm

Due to the vast popularity of DBSCAN over other density based clustering algorithms, this paper implements that algorithm on the dataset. Applying DBSCAN algorithm does not need to define the total number of cluster to make from the user. It is itself able to determine the number of meaningful clusters using its two parameters: radius of neighborhoods ϵ within which points are considered neighbor to each other and MinPts which refers to the number of minimum points needed to be considered as a cluster. Table III shows the experimental results on applying DBSCAN algorithm on the dataset varying ϵ and MinPts. For each combination of these two parameters, number of clusters determined by the algorithm, number of unclustered instances and time elapsed are recorded and shown.

V. Discussion

Table II shows that the sum of squared error for Euclidean distance is decreasing with the increasing of total number of clusters. Because, the more clusters are created, the more compact and dense the points become in a cluster. Similarly, about Manhattan distance, the sum of the distances within cluster is also decreasing as the points become closer when the number of clusters increases. Besides, Table III shows that DBSCAN can find the total number of clusters itself without being defined explicitly. If the number of points needed to get recognition as a cluster is increased, then more number of unclustered instances are found as they can not form cluster.

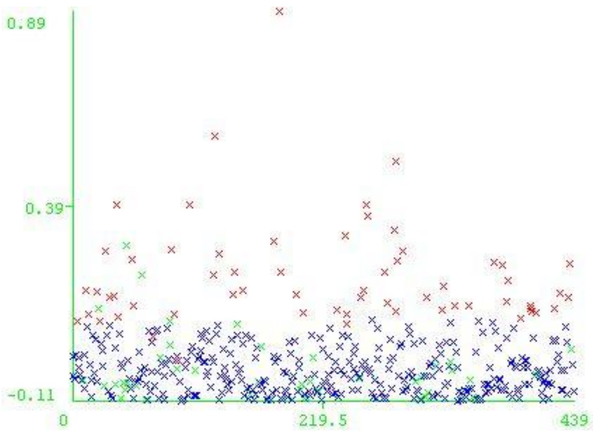
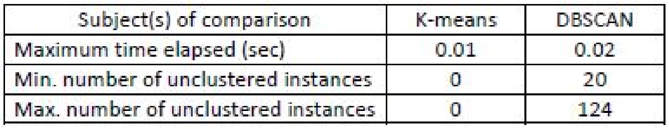


Fig. 4: Clusters found by k-means for 440 customers (red, green and blue colors represent instances of different clusters)

From the visualization of the experimental results found from implementation, customer segmentation observed for the dataset considered is shown in Figure 4. It is found by applying k-means algorithm where value of k is 3 and Euclidean distance is considered as distance metric. From this figure, it seems that there are two prominent clusters of customers marked by red and blue colors and one other cluster marked by green color which does not seem that meaningful. Customers of this green cluster can be considered as noise or anomalous from density based perspective.

To observe the effectiveness of these two algorithms, performances shown by both of the algorithms can be compared based on the time elapsed and also the number of unclustered instances.

Table IV: Comparison between performances of k-means and DBSCAN for customer segmentation



Subject(s) of comparison	K-means	DBSCAN
Maximum time elapsed (sec)	0.01	0.02
Min. number of unclustered instances	0	20
Max. number of unclustered instances	0	124

Table IV shows a comparison between the performances of k-means and DBSCAN for customer segmentation. According to the data of this table, Figure 5 shows that DBSCAN takes relatively longer time than k-means; however, in spite of consuming more time, it provides a meaningful customer segmentation as well as detects some anomalous customers whose spending habits are different.

If the data of the table other than only time elapsed are considered, then concentration should also be put on the total number of unclustered instances. It can be seen that for using any type of distance metric, k-means algorithm clusters all the points of the dataset. Hence, both the minimum and the maximum number of unclustered instances found from this algorithm are zero. On the other hand, minimum and maximum number of the unclustered instances found using DBSCAN algorithm are 20 and 124 respectively.

In case of customer segmentation, these unclustered instances can play an important role. Because, this unclustered instances which are also alternatively known as noise from density based algorithm perspective can be used to identify customers having different behaviors. For example, customers having unusual spending habits can be detected considering their respective data points as noise by applying DBSCAN algorithm.

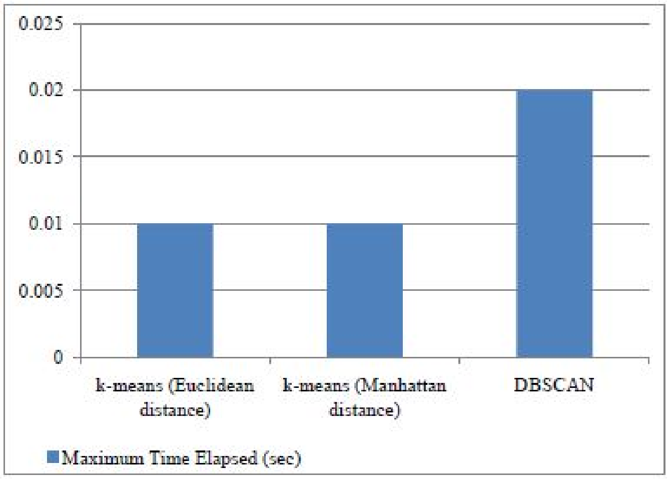


Fig. 5: Maximum time elapsed (sec) by k-means and DBSCAN for customer segmentation

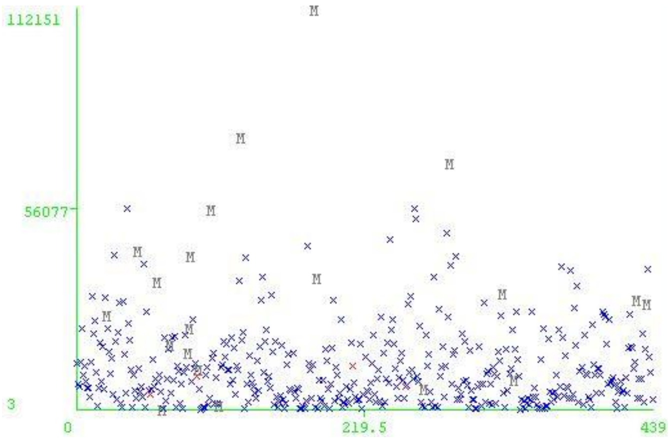


Fig. 6: Identifying customers having unusual spending habits with the help of DBSCAN algorithm

For real life applications and industrial use, performance often gets more priority than time taken. Hence, in order to get satisfactory customer segmentation, customers showing different behaviors also need to be identified beside just grouping them based on shared characteristics. For that reason, time consumed by the DBSCAN algorithm can be ignored. Because, unlike k-means clustering, DBSCAN algorithm does not only categorize customers into groups based on shared characteristics but also is able to spot customers having unusual behavior like relatively more spending habits. Figure 6 shows how DBSCAN algorithm is able to identify customers (denoted by 'M') having aberrant type of spending habits with the help of its unclustered instances. In this figure, vertical axis shows annual spending of customers where it clearly shows that DBSCAN identifies customers who spend relatively more than others. This identification is very useful to detect potential customers. That is why this paper claims that using density based clustering algorithm can result in relatively more meaningful clustering or customer segmentation.

VI. Conclusion & Future Work

In this paper, k-means and DBSCAN algorithms are chosen to represent centroid based and density based algorithms for data clustering respectively. Results obtained from the implementation of these two algorithms show that any of them can be used for customer segmentation; however, in contrast to k-means, DBSCAN provides an extra option to find unusual customers having different spending habits which is very effective to ensure customer satisfaction and optimal profit. Besides, the results found from the implementation of density based clustering algorithm seem meaningful for the dataset considered. Hence, it can be said that density based clustering algorithms are also worth considering to apply in order to get satisfactory customer segmentation. Hopefully, to get more satisfactory customer segmentation in future, while using Neural Network for cluster analysis, other types of clustering algorithms will also be applied on different datasets and their performances will be evaluated as well.

References

[1] W. R. Smith, “Product differentiation and market segmentation as alternative marketing strategies,” *Journal of marketing,* Vol. 21, No. 1, pp. 3-8, July 1956.

[2] M. Namvar, M. R. Gholamian, S. KhakAbi, “A Two Phase Clustering Method for Intelligent Customer Segmentation,” *International Conference on Intelligent Systems, Modelling and Simulation,* 2010.

[3] H. Hruschka, and M. Natter, “Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation,” *European Journal of Operational Research,* Vol. 114, No. 2, pp. 346-353, July 1999.

[4] J. Wu, and Z. Lin, “Research on customer segmentation model by clustering,” *Proceedings of the 7th international conference on Electronic commerce,* 2005.

[5] J. H. Lee, Jang Hee, and S. C. Park, “Intelligent profitable customers segmentation system based on business intelligence tools,” *Expert systems with applications,* Vol. 29, No. 1, pp 145-152, 2005.

[6] T. Teichert, E. Shehu, and I. V. Wartburg, “Customer segmentation revisited: The case of the airline industry,” *Transportation Research Part A: Policy and Practice,* Vol. 42, No. 1, pp. 227-242, 2008.

[7] W. T. Anderson, E. P. Cox, and D. G. Fulcher, “Bank selection decisions and market segmentation,” *Journal of Marketing,* Vol. 40, No. 1, pp 40-45, January 1976.

[8] S. Dolnicar, “A review of data-driven market segmentation in tourism,” *Journal of Travel & Tourism Marketing* Vol. 12, No. 1, pp 1-22, July 2002.

[9] ReachForce.com, “4 Ways to Segment Your Big Data and Super-Boost Your Lead Generation Efforts,” [Online] Available: https://www.reachforce.com/blog/4-ways-to-segment-your-big-data-and-super-boost-your-lead-generation-efforts [Accessed: August 18, 2017].

[10] A. Singh, A. Yadav, and A. Rana, “K-means with Three different Distance Metrics,” *International Journal of Computer Applications,* Vol. 67, No. 10, 2013.

[11] L. Ertöz, M. Steinbach, and V. Kumar, “Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data,” *Proceedings of the 2003 SIAM International Conference on Data Mining,* 2003.

[12] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” *Kdd ,* Vol. 96, No. 34, pp. 226-231, 1996.

[13] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, “OPTICS: Ordering Points To Identify the Clustering Structure,” *ACM Sigmod record,* Vol. 28, No. 2, pp. 49-60, 1999.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: an Update,” ACM SIGKDD explorations newsletter, Vol. 11, No. 1, pp. 10-18, 2009.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* Vol. 12, pp. 2825-2830, October 2011.

[16] “Wholesale Customers Data Set,” [Online] Available: https://archive.ics.uci.edu/ml/datasets/wholesale+customers [Accessed: September 11, 2017].

[17] N. Abreu, “Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional,” *Mestrado em Marketing,* ISCTE-University Institute of Lisbon, 2011.

[18] T. M. Kodinariya, and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” *International Journal of Advance Research in Computer Science and Management Studies,* Vol. 1, No. 6, pp. 90-95, 2013.