TELKOMNIKA, Vol.17, No.1, February 2019, pp.110~117
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i1.9394

Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour

Siti Monalisa*[1], Fitra Kurnia[2]
[1]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Riau, Indonesia
[2]Department of Informatics Engineering, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Riau, Indonesia
Corresponding author, e-mail: siti.monalisa@uin-suska.ac.id[1], fitra.k@uin-suska.ac.id[2]

Abstract

The aim of study is to discover outlier of customer data to found customer behaviour. The customer behaviour determined with RFM (Recency, Frequency and Monetary) models with K-Mean and DBSCAN algorithm as clustering customer data. There are six step in this study. The first step is determining the best number of clusters with the dunn index (DN) validation method for each algorithm. Based on the dunn index, the best cluster values were 2 clusters with DN value for DBSCAN 1.19 which were minpts and epsilon value 0.2 and 3 and DN for K-Means was 1.31. The next step was to cluster the dataset with the DBSCAN and K-Means algorithm based on the best cluster that was 2. DBSCAN algorithm had 37 outliers data and K-means algorithm had 63 outliers (cluster 1 are 26 outliers and cluster 2 are 37 outliers). This research shown that outlier in DBSCAN and K-Means in cluster 1 have similarities is 100%. But overal outliers similarities is 67%. Based the outliers shown that the behaviour of customers is a small frequency of spending but high recency and monetary.

Keywords*: customer behaviour, DBSCAN, dunn index, K-means, outlier and RFM models*

1. Introduction

Clustering is the process of dividing the objects into groups so that the objects within a group have similarities with each other and those objects have no resemblance to the objects in the other group. clustering is also referred to as segmentation data [1]. Clustering has been widely used in various fields such as in the case of hotspot data clustering, customer segmentation, customer behaviour and more. The clustering method of subscriber grouping has been widely used as in the research [2–6].The algorithm used in the cluster are many kinds such as K-Means, Self Organizing Map, DBSCAN and others [6–13].However, the common and simple algorithm used is the K-Means algorithm [14].

Algorithm K-Means has been successfully applied to various fields [15]. However, this algorithm is very sensitive to the choice of starting point [4] and also sensitive to the outliers because these objects distort the average value of clusters [1]. This is because the determination of the number of clusters in the K-Means algorithm is determined by the user [1], [12], [16]. However, many researchers have now discovered the method of cluster validity in determining the best number of clusters on a dataset. One of method validity is Dunn Index Method was developed by Dunn [17].

In addition to K-means, another clustering method is DBSCAN. This algorithm is different to the K-Means because it does not require the user to specify the number of clusters produced. DBSCAN has better performance compared with K-means. This has been demonstrated in the study [8] suggesting that DBSCAN has higher sensitivity and better segmentation. DBSCAN is designed to find the dataset portion containing cluster and noise changing [9] by using the Epsilon (eps) and Minimal Point (minpts) parameters that are useful in determining the distance and minimum number of neighbors and core point. In addition to handling the noise, DBSCAN can also find outliers in arbitrary clusters [18]. Outliers are objects in datasets that are much different from the rest of the objects in the data set [15] which do not contain enough number of points (minpts) in forming the clusters [19]. Outliers are often discarded because they are considered noise [1] but actually the detection of outlier data or

Received May 21, 2018; Revised November 8, 2018; Accepted November 30, 2018

so-called anomaly data is necessary if there is a dataset that provides important information to the system [19]. Important information derived from customer data collection can be obtained from the results of data analysis. Because each customer does not have the same behaviour [14], [19] then the data needs to be analyzed to find profitable customers. One of model that is able to measure customer behaviour is the RFM Models with three criterias namely Recency, Frequency and Monetary.

Customers who have different behaviours can cause outliers. Outliers found in customer's data can generate favorable customer behaviour or vice versa. If profitable customer is detected as an outlier and the outlier is discarded, this will harm the company because there is important information about profitable customers such as customer profiles, etc.

Through the DBSCAN and K-Means algorithms, outliers in customer data sets can be found to see different behaviours among customers. The outlier in K-Means is determined by determining the distance between the object and a group of objects [15]. The K-Means algorithm classifies datasets into several clusters and checks whether objects in the cluster are detected by outliers [15]. Objects detected as outliers are objects that are far from the core point of the cluster [15]. Because these two algorithms are able to find outliers, this study will compare and test the outliers in each algorithm with the same customer data collection. This aims to determine the consistency of outliers in the customer data collection with RFM Models through the DBSCAN and K-Means algorithms. Furthermore, the data contained in the outliers in both algorithms will be analyzed to see the information whether the outliers contain profitable customers or vice versa so that decisions need to be made in providing services to customers.

2. Research Method

2.1. RFM Model

RFM model is the model developed by Hughes in 1994 in estimating subscriber life value [20] and customer loyalty behaviour [21] with 3 variables: recency (R), Frequency (F) and Monetary (M). Recency is the customer's time interval since the last purchase with certain period of time, Frequency is the number of purchases made by customers in a certain period, and Monetary is the amount of money that customers spend to the company in a certain period [21]. Customer RFM values is needed to be known to assist companies in marketing because high customer RFM is more responsive to promotions, more likely to repeat order and most profitable purchases [22] when it is compared to customers with low RFM scores.

To knowing the RFM value of the customer [2] using the RFM value, the symbol '↑' is a value higher than the average value, the symbol '↓' is a value lower than the average value. This means that the higher the value it will be better for the company and the lower of the average it will get worse for the company. But for R, the symbol ↓ means the lower of the average then the better for the company and the symbol ↑ means higher than average then the value is not good for the company. The cluster belonging to the symbol R ↓ F ↑ M ↑ is named with Loyal Customer, the symbol R ↑ F ↓ M ↓ is called Lost Customer, the symbol R ↓ F ↓ M ↓ is called New Customer and the symbol R ↓ F ↑ M ↓ called Prospect Customer. The symbols are explained in Table 1 [2].

Table 1. Characteristic of the Cluster

No.	Symbol RFM	Cluster Name	Information
1	R ↓ F ↑ M ↑	Loyal Customer	This customer group is customers who has recently made a purchase with a high number of transactions and the amount of money spent is high too.
2	R ↑ F ↓ M ↓	Lost Customer	This group of customers is customers that has long made no purchase with a low number of transactions and the money spent is low too
3	R ↓ F ↓ M ↓	New Customer	This customer group is the customer has just made a purchase with a low number of transactions and the money is still low
4	R ↓ F ↑ M ↓	Prospect Customer	This customer group is a customer who has just made a purchase with a high number of transactions but the money is still low

2.2. DBSCAN Algorithm

DBSCAN is an algorithm developed by [9] with 2 input parameters, namely epsilon (eps) and minimum points (Minpts). Eps is the maximum point distance in forming a cluster and

minpts are the minimum number of points on a formed cluster [9]. The number of clusters in this algorithm is determined by the value of eps and mints inputted by user. As its name, which is density-based spatial clustering of applications with noise (DBSCAN), this algorithm is able to detect noise when there are different data points with other data sets [9]. In addition to noise, this algorithm is also able to define anomalous data or data outliers in the data series using the same 2 parameters where the point will be considered outlier if it does not contain sufficient number of dots in forming clusters or minpts determined previously [23].

The steps of the DBSCAN are as follows:

1) Select any point or object randomly from the data set as candidate corepoint
2) If the selected object qualifies as a core point having minpts and epsilon which has benn specified by user, then the object will form a new cluster with its neighbor object. Calculate the distance between a corepoint object and a neighbor object using *euclidean distance* formula [8]:

$$dxy = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where n is number of points in the sequence, yi is the mean of sequence and xi is sequence of data points within each window.

3) Objects which are not included as corepoint or neighbor objects in step 2 will then be processed by making them as the next corepoint candidate. If it meets as corepoint then the object will form the next cluster with its neighbor object. And so on until all the objects in the data set are tested.

If the object that has been tested does not qualify as a corepoint or a neighbor object, they can be categorized as an outlier/noise that is the object which corepoint distance is larger than epsilon and the number of reachable densities is less than the user-specified minpts.

2.3. K-Means Algorithm

K-Means is an algorithm that is categorized into partition clustering methods [5]. This algorithm aims to collect all datasets into determined clusters [24]. Steps in the K-means method are as follows:

1) Determine the number of clusters
2) Select the initial centroid randomly according to the number of clusters that have been determined
3) Calculate the distance of data to the centroid with the euclidean distance formula in (1).
4) Renew the centroid by calculating the average value of each cluster
5) Return to step 3 if there is still data moving clusters or centroid value changes.

In steps 1 and 2 these steps, the number of clusters is determined by dunn index method to identify optimal number of cluster [6]. This is because K-means is very sensitive to starting point selection to part items into specified clusters [25].

2.4. Dunn Index

Dunn Index (DI) is a cluster validity to identify optimal number of cluster with the highest value having the best clusters [26]. Dunn index is calculated based on the following equation [26]:

$$D_{nc} = \min_{i=1,\ldots,nc}\left\{\min_{j=i+1,\ldots,nc}\left(\frac{d(c_i,c_j)}{\max_{k=1\ldots,nc}\,diam(c_k)}\right)\right\} \quad (2)$$

where $d(ci, cj)$ is different function between cluster ci and cj defined as:

$$d(ci, cj) = \min_{x\in c_i, y\in c_j} d(x,y) \quad (3)$$

and $diam(C)$ is cluster diameter probably considered as cluster dispersion size. Cluster diameter of C can be defined as flows:

$$diam(C) = \max_{x,y\in C} d(x,y) \quad (4)$$

3. Methodology

The steps in this study consisted of 6 steps. The first step is to determine customer data with 1866 R, F and M attributes from the Herbal Penawar Alwahida Indonesia retail company. The second step is to normalize the data with the aim that each attribute R, F and M does not have a long range because the value of M is the value of money with different units of rupiah with the value of recency and frequency using the following formula [1]. This method performs a linear transformation on the original data [6]. where min_A and max_A are the minimum and maximum values of an attribute, A. Then Min- max normalization maps a value, v, of A to v' in the range of [$newmin_A$, $newmax_A$] by:

$$V'i= \left(\frac{(v_i - \min_A)}{(\max_A - \min_A)}\right)(\text{newmax}_A - \text{newmin}_A) + \text{newmin}_A \quad (5)$$

The third step is to determine the best cluster using the dunn index validation method for each DBSCAN and K-Means algorithm. The best clustering results are then used in each algorithm. The fourth step is the DBSCAN algorithm will input the epsilon value and the optimal minimum point is obtained from the results of the Dunn Index validation. The K-Means algorithm will cluster according to the optimal number of clusters and then search outliers for each cluster by using the outlier score formula and find out whether the outliers are global or collective outliers. The fifth step is to analyze whether the outliers have the same data or points in the two algorithms. The last step was to analyze the data outliers in the two algorithms to find out information found on customer data that detected outliers.

4. Results and Analysis

Data of RFM Models in Table 2 have to normalized by using Min-Max method using equation 5 with range 0-1 and the results shown in Table 3 .This study utilizes dunn index method in searching the optimal number of cluster both K-Means and DBSCAN using (2), (3) and (4). Table 4 is the Dunn Index value in K-Means. Based on Table 4, the number of optimal clusters is 2 with a dunn index value of 1.31. In this study, the experiments on the number of clusters in K-Means were carried out from clusters 2-9 because the dunn index value produced in this study was the more number of clusters, the smaller the dunn index value.

Table 2. Data of RFM Models

No.Customer	R	F	M (Rp)
1	201	1	385.000
2	174	4	2.244.000
3	4	5	4.225.000
4	121	3	116.000
5	148	2	193.000
6	202	2	2.058.000
7	3	2	180.000
8	59	7	1.248.000
9	243	1	100.000
10	17	19	4.230.000
11	86	5	1.675.000
12	121	3	175.000
...	201	1	385.000
1866	121	7	1.155.000

Table 3. Normalized Data

No.Customer	R_N	F_N	M_N
1	0.7390	0.0000	0.0007
2	0.6397	0.0144	0.0041
3	0.0147	0.0192	0.0078
4	0.4449	0.0096	0.0002
5	0.5441	0.0048	0.0004
6	0.7426	0.0048	0.0038
7	0.0110	0.0048	0.0003
8	0.2169	0.0288	0.0023
9	0.8934	0.0000	0.0002
10	0.0625	0.0865	0.0078
11	0.3162	0.0192	0.0031
12	0.4449	0.0096	0.0003
...			
1866	0.4449	0.0288	0.0021

Table 5 is the dunn index value in the dbscan algorithm. This study examines epsilon values from 1 to 6 and the minimum value points from 0.1 to 1.0. Based on the epsilon and minimum point value tested, the highest dunn index value is 1.02 with the optimal number of clusters is 2. So, Based on the Dunn index value, the number of optimal clusters generated in both algorithms is two clusters.

The next step after finding the optimal number of clusters is to determine the outliers in both algorithms. In the DBSCAN algorithm, the data detected as outliers amounted to 37 data outside of the data in cluster 1 and cluster 2. In DBSCAN algorithm, the amount of data in cluster 1 was 800 and cluster 2 was 1030. The data in cluster 1, cluster 2 and outliers in

DBSCAN algorithm are shown in Table 6. The third column in Table 6 is the outliers produced by DBSCAN algorithm and the outliers.

Table 4. The Optimal Cluster Determination using Dunn Index (K-Means)

No.	Number of K-Means Cluster	DI of K-Means
1	2	1.31
2	3	0.60
3	4	0.51
4	5	0.44
5	6	0.35
6	7	0.25
7	8	0.25
8	9	0.20

Table 5. The Optimal Cluster Determination using Dunn Index (DBSCAN)

No.	No of DBSCAN Cluster	Eps	Mintps	DI of DBSCAN
1	8	0.1	10	0.21
2	4	0.2	10	0.21
3	2	0.3	10	1.02
4	2	0.4	10	1.02
5	2	0.5	10	1.02
6	3	0.1	5	0.86
7	2	0.4	3	1.02
8	2	0.5	3	1.02

Table 6. Cluster 1, Cluster 2 and Outlier in DBSCAN Algorithm

Cluster 1	Cluster 2	Outliers	Cluster 1	Cluster 2	Outliers
0	2	20	47	40	960
1	6	22	48	42	963
3	7	25	51	44	993
4	11	27	52	45	1033
5	12	88	53	46	1135
8	13	251	54	49	1151
9	17	269	56	50	1155
10	18	275	57	55	1162
14	21	305	58	59	1163
15	23	351	69	60	1174
16	26	482	71	61	1183
19	29	718	75	62	1352
24	30	722	81	63	1375
28	31	731	85	64	1419
32	33	813	86	65	1559
35	34	857	87	66	1588
37	36	863	...	...	1767
41	38	886	1866	578	1830
43	39	956			

The cluster results in K-means algorithm are 2 with the number of data in cluster 1 is 1065 and cluster 2 is 801. Outliers in K-means are found objects in datasets that are much different from the rest of the objects in the data set [15]. Sitanggang and Baehaki [15] identified outliers in K-means to be 2, namely global and collective outliers. Global outliers are outliers that occur when an object deviates from the data set. Collective outliers are outliers that occur when a cluster [1], [15]. The outliers in this study are global data because there is no number of data found in the cluster below 1% of the total data. The data that detects global outliers are 25 from cluster 1. As for the collective outliers as outliers who have the amount of data in the cluster below 1%. The 1% value is based on Sitanggang and Baehaki research. The determination of outliers in K-Means with a global outlier is identified as an object which is far away from the centroid on all cluster. This study have identified each cluster to analyze global outlier using outlier score in (6) [1]. The outlier score used in this K-means aims to looked the outlier score have generated by each data that detected by the outlier [15]. It means is the higher the outlier score indicates that the data is far from the centroid of cluster.

$$\text{Outlier Score} = \frac{dist(o,c_o)}{i_{c_o}} \quad (6)$$

Where *o* is an object in the dataset, c_o is nearest centroid or center ti the object *o*, dist (o,c) is distance between the object o to its nearest centroid c_o and i_{c_o} is average distance from c_o to the object assigned to *o*.

Outlier scores on the K-means algorithm are used from 2.0 to 9.4 based on the score outliers that have been generated as shown in Table 7. The number of outliers in K-means in cluster 1 is 26 and the number of outliers in cluster 2 is 37 datasets. In this study, the number of outliers in K-Means all of 63 data. The Outlier score in K-means algorithm shown in Table 7. Table 7 is the outlier score in cluster 1 and cluster 2. The second column in Table 7 are the outlier's data on cluster 1 which has same outliers with DBSCAN algorithm that marked bold. Based on the third column in Table 6 and the second column in Table 7, it can be seen that there are some of the same data marked in bold. The same data in each table is 26 and is shown in Table 8 in the fourth column. Twenty six data is customer data which has the same point in both algorithms.

Table 7.Outlier score in K-Means

No.	Cluster 1 (C1)	Score of Outlier C1	Cluster 2 (C2)	Score of Outlier C2
1	22	2,32	15	2,09
2	25	2,01	189	2,14
3	88	3,78	238	2,14
4	251	5,22	267	2,16
5	269	2,08	272	2,24
6	275	3,14	322	2,09
7	351	2,11	329	2,09
8	482	4,19	451	2,19
9	718	3,97	523	2,14
10	722	5,45	534	2,16
11	813	2,19	569	2,06
12	857	6,06	597	2,19
13	886	8,83	643	2,01
14	956	3,84	671	2,06
15	993	2,26	680	2,01
16	1033	3,88	683	2,04
17	1135	2,75	686	2,22
18	1151	4,61	761	2,19
19	1155	3,23	768	2,01
20	1163	9,42	842	2,14
21	1183	7	878	2,19
22	1419	2,6	890	2,22
23	1559	5,03	1035	2,19
24	1588	3,29	1191	2,22
25	1767	3,86	1200	2,24
26	1830	8,46	1236	2,24
27	-	-	1323	2,19
28	-	-	1335	2,14
29	-	-	1344	2,11
30	-	-	1399	2,09
31	-	-	1467	2,16
32	-	-	1556	2,11
33	-	-	1577	2,04
34	-	-	1580	2,16
35	-	-	1589	2,14
36	-	-	1624	2,04
37	-	-	1793	2,11

Table 8. Dataset having the same outlier both in DBSCAN and K-Means

No.	Dataset K-Means	Dataset DBSCAN	The Same Outliers (DBSCAN and K-Means)
1	15	20	22
2	189	27	25
3	238	305	88
4	267	731	251
5	272	863	269
6	322	960	275
7	329	963	351
8	451	1151	482
9	523	1162	718
10	534	1174	722
11	569	1352	813
12	597	1375	857
13	643	-	886
14	671	-	956
15	680	-	993
16	683	-	1033
17	686	-	1135
18	761	-	1151
19	768	-	1155
20	842	-	1163
21	878	-	1183
22	890	-	1419
23	1035	-	1559
24	1191	-	1588
25	1200	-	1767
26	1236	-	1830
27	1323	-	-
28	1335	-	-
29	1344	-	-
30	1399	-	-
31	1467	-	-
32	1556	-	-
33	1577	-	-
34	1580	-	-
35	1589	-	-
36	1624	-	-
37	1793	-	-

The last step of this study after finding outlier data in both algorithms is finding important information from customer data by looking at the Recency, monetary value and frequency of each customer data. There are 4 symbols found in this study, namely R ↓ F ↓ M ↓, R ↓ F ↑ M ↓, R ↓ F ↓ M ↑, and R ↓ F ↑, M ↑. Symbol ↓ is a low value compared to the average value, the symbol ↑ is a high value compared to the average value. In this study found the R value of the subset was obtained between 0.0-0.2243, F value between 0.106-1.0 and M value between 0.004-1. This means that this customer data has a very high R value from the average of the other datasets which is below 0.3589, different F values and heights from the average average

data set that is below 0.0227 and a low M value and height is also different from the average of other datasets which are below and above 0.0068.

Based on Table 1 shown that symbol R ↓ F ↓ M ↓ is a New Customer who has buying behaviour with a low number of transactions and still low money. The R ↓ F ↑ M ↓ symbol is a Prospect Customer who has a high purchasing behaviour with a high number of transactions but the money is still low. The R ↓ F ↓ M ↑ symbol is a New Customer with a low number of transactions but a high amount of money is spent. The symbol R ↓ F ↑, M ↑ is Loyal Customer with buying behaviour with a high number of transactions and the amount of money spent is also high.

5. Conclusion and Further Research

This study found outliers in DBSCAN are 37 and K-Means are 63 (26 in cluster 1 and 37 in cluster 2). The found outlier has some data or object which is equal to 67 percent, which is the result of 37 outliers in DBSCAN and 26 at K-means.This study also found that the outliers produced in K-Means were global outliers. This is because the data found is part of the data set in the data that has the amount of data above 1% of the entire data. In addition, the outliers found at the same K-means as DBSCAN are outliers in cluster 1 of K-means. This requires further research to find each global outlier in K-means as an outlier in DBSCAN.

The Outliers that were found consisted of 3 charactristic of clusters, namely Prospect Customers, New Customers, and Loyal Customers. It was concluded that most of the data in this study included the characteristics of Lost Customers, namely customers who had not purchased for a long time and made a purchase with a low number of transactions and also low money. Therefore, this company needs to make a strategy so that customers can make payments and remain loyal to the company.

References

[1] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. San Francisco, CA, itd: Morgan Kaufmann. 2012. 745: 2.
[2] Dursun A, Caber M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives.* 2016; 18: 153–60.
[3] He X, Li C. The Research and Application of Customer Segmentation on E-commerce Websites. 2016;
[4] Hosseini SMS, Maleki A, Gholamian MR. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*. 2010; 37(7): 5259–5264.
[5] Kandeil DA, Saad AA, Youssef SM. *A two-phase clustering analysis for B2B customer segmentation*. Proceedings-2014 International Conference on Intelligent Networking and Collaborative Systems, IEEE INCoS 2014. 2014; 221–228.
[6] Khajvand M, Zolfaghar K, Ashoori S, Alizadeh S. *Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study*. Procedia Computer Science. 2011;3:57–63.
[7] Alamsyah A, Nurriz B, Carlo AM. Monte Carlo. *Simulation and Clustering for Customer Segmentation in Business Organization*. Conference: 2017 3rd International Conference on Science and Technology -Computer(ICST). 2017.
[8] Dudik JM, Kurosu A, Coyle JL, Sejdić E. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. *Computers in Biology and Medicine.* 2015; 59: 10–18.
[9] Ester M, Kriegel HP, Sander J, Xu X. *Density-Based Algoritm for Discovering Clusters in Large Spatiasl Database with Noise. Comprehensive Chemometrics*. KDD-96 Proceedings. 1996; 2: 635–654.
[10] Hermawati R, Sitanggang IS. *Web-Based Clustering Application Using Shiny Framework and DBSCAN Algorithm for Hotspots Data in Peatland in Sumatra*. Procedia Environmental Sciences. 2016; 33: 317–323.
[11] Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies.* 2013; 1(6): 2321–7782.
[12] Sethi C, Mishra G. A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset. *International Journal of Scientific & Engineering Research*. 2013; 4(6): 1559–1567.
[13] Holmbom AH, Eklund T. Customer Portfolio Analysis using the SOM. 2008; (2007): 412–422.
[14] Berry MJA, Linoff GS. AM. second edition. Customer Relation Management, Concept and Technologies. 2008.

[15] Sitanggang IS, Baehaki DAM. *Global and collective outliers detection on hotspot data as forest fires* indicator in Riau Province, Indonesia. ICSDM 2015-Proceedings 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. 2015; 66–70.
[16] Bellatreche L, Cuzzocrea A, Song IY. Advances in data warehousing and OLAP in the big Data Era. *Information Systems*. 2015; 53: 39–40.
[17] Dunn JC. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*. 1974; 4(1): 95–104.
[18] Halkidi M. On Clustering Validation Techniques. 2001; 107–45.
[19] Çelik M, Dadaşer-Çelik F, Dokuz AŞ. *Anomaly detection in temperature data using DBSCAN algorithm*. INISTA 2011-2011 International Symposium on INnovations in Intelligent SysTems and Applications. 2011; (June 2014): 91–95.
[20] Parvaneh A, Abbasimehr H, Tarokh MJ. Integrating AHP and data mining for effective retailer segmentation based on retailer lifetime value. *Journal of Optimization in Industrial Engineering*. 2012; 5(11): 25–31.
[21] Buttle F, Stan M. Customer Relationship Management. Third edit. Butterworth-Heinemann; 2015.
[22] Hamzehei A, Fathian M, Farvares H, Gholamian MR. A new methodology to study customer electrocardiogram using RFM analysis and clustering. 2011.
[23] Çelik M, Dadaşer-Çelik F, Dokuz AŞ. *Anomaly detection in temperature data using DBSCAN algorithm*. INISTA 2011-2011 International Symposium on INnovations in Intelligent SysTems and Applications. 2011: 91–95.
[24] Liping Z, Song D, Shiyue L. *Analysis of power consumer behaviour based the complementation of K-means and DBSCAN*. 2017;(1).
[25] Ghalenooie MB, Sarvestani HK. *Evaluating Human Factors in Customer Relationship Management Case Study: Private Banks of Shiraz City*. Procedia Economics and Finance. 2016; 36(16): 363–373.
[26] Vendramin L, Campello RJGB, Hruschka ER. *On the Comparison of Relative Clustering Validity Criteria*. In: Apte C, Park H, Wang K, Zaki MJ, editors. Proceedings of the 2009 SIAM International Conference on Data Mining. Philadelphia, PA. Society for Industrial and Applied Mathematics. 2009: 733–44.