Using Self Organizing Maps and K Means Clustering Based on RFM Model for Customer Segmentation in the Online Retail Business

Rajan Vohra[1(✉)], Jankisharan Pahareeya[2], Abir Hussain[1], Fawaz Ghali[1], and Alison Lui[3]

[1] Department of Computer Science, Liverpool John Moores University, Liverpool, UK
{r.vohra, a.hussain, f.ghali}@ljmu.ac.uk
[2] Rustamji Institute of Technology, BSF Academy, Tekanpur, Gwalior, India
talkto.pahariya@gmail.com
[3] School of Law, Liverpool John Moores University, Liverpool, UK
a.lui@ljmu.ac.uk

Abstract. This work based on the research of Chen et al. who compiled sales data for a UK based online retailer for the years 2009 to 2011. While the work presented by Chen et al. used k means clustering algorithm to generate meaningful customer segments for the year 2011, this research utilised 2010 retail data to generate meaningful business intelligence based on the computed RFM values for the retail data set. We benchmarked the performance of k means and self organizing maps (SOM) clustering algorithms for the filtered target data set. Self organizing maps are utilized to provide a framework for a neural networks computation, which can be benchmarked to the simple k means algorithm used by Chen et al.

Keywords: Online retail data · RFM model · K means clustering · Self-organizing maps · Business intelligence

1 Introduction

According to the retail and e-commerce sales figures put out by e-marketer, the total online retail sales for 2019 was 106.46 billion pounds representing 22.3% of total retail sales. This is expected to grow to 139.24 billion pounds in 2023 representing 27.3% of total retail sales in the UK. The percentage of mobile commerce using smart phones expected to rise from 58.9% in 2019 to 71.2% in 2023. According to the office for National Statistics, UK, clothes or sports goods account for 60% of the goods purchased online in Great Britain in 2019. The other key goods or services are: House hold goods representing 49%, holiday accommodation with 44%, travel arrangements with 43% and tickets for events 43%. Retailers are interested in gaining business intelligence about their customers. This can represent the buying patterns, expenditure, repeat purchases, longevity of association and high profit customer segments. In addition sales pattern by region, season and time are key components of such knowledge. This enables design of suitable marketing campaigns and the discovery of new patterns in

D.-S. Huang and P. Premaratne (Eds.): ICIC 2020, LNAI 12465, pp. 484–497, 2020.
https://doi.org/10.1007/978-3-030-60796-8_42

the sales data which were previously unknown to the retailer. Chen et al. [1] have analysed the data set for an online retailer for the year 2011. Using k means clustering they have derived meaningful customer segments, and then used decision tree based rule induction to build decision rules that represent gained business intelligence. This work based on the work of Chen et al. by using simple k means clustering [12] and self-organizing maps [15] to perform clustering for the 2010 retail data set.

RFM (Recency, Frequency, and Monetary) is a model to analyse the shopping behavior of a customer. Recency represents the duration in time since the last purchase while frequency represents the number of purchases made in a given time period and monetary denotes the amount spent by a customer in the given time period which in the analysis performed is the calendar year 2010. Dogan et al. [2] used RFM computations and k means clustering to segment customers of a sports retail company based in Turkey to design a customer loyalty card system based on this analysis. The k means clustering analysis performed by Dogan et al. has used the retail data set designed by Hu & Yeh [6]. Sarvari et al. [7] used RFM analysis on a global food chain data set. They used k means clustering and association rule mining for segmenting customers and buying patterns. They highlighted the importance of assigning weights to RFM values. Yeh et al. [8] added time since first purchase to the basic RFM model which improved the predictive accuracy of the RFM model. Our analysis uses time in months to indicate the first purchase made by a customer in the time period under consideration – for year 2010 in our analysis. Wei et al. [9] have discussed comprehensively a review of the RFM model including its scoring scheme, applications especially in customer segmentation, merits and demerits, along with how RFM model can be extended to perform a more comprehensive analysis by adding other variables like churn and also incorporating call centre data. Customer segmentation using Neural networks is demonstrated for the global tourist business by Bloom [10]. Holmbom et al. [11] used self-organizing maps to cluster customers for portfolio analysis in order to determine profitability for target marketing purposes – in this study they used both demographic data as well as product profiles. Vellido et al. used self-organizing maps for segmenting online customer data [13]. Self-organizing maps can be visualized using the U matrix (unified distance matrix) which displays the Euclidean distance between neurons, according to ultsch [14]. While Kiang et al. [17] used self-organizing maps to discover interesting segments of customers in telecommunication service providers data sets. Although Chen et al. published their retail analytics paper in 2012 [1], the associated data set was uploaded on the UCI machine learning repository only in September 2019 [5]. Chen et al. have analysed the data set for an online retailer for the year 2011. We want to analysed for the online retailer for the year 2010 data which Chen et al. did not do. We also want to explore the potential of neural network that's why have chosen SOM on account of its Neural networks framework because of its robust architecture and lesser sensitivity to Noise in the input data set.

We did our study in two phases:

Phase 1: Cluster Profiling.
Phase 2: Performance benchmarking.

these two phases discus in details in research methodology Sect. 2.

The data set pertains to the operations of a small online retailer based in the UK. Chen et al. used RFM model and k means clustering to derive a new segmentation of customers of this online retailer. After profiling the clusters generated, decision tree based rule induction was used to derive decision rules representing business intelligence gained. We use k means and self organizing maps as well as RFM values computed to perform clustering to obtain new customer segments. While Chen et al. used 2011 data, we use the data set for 2010, from the same data source. We have profiled the clusters generated and compare the performance of k means and self organizing maps based on certain parameter values generated during execution. The remainder of this paper is organized as follows. Section 2 discusses the proposed research methodology while Sect. 3 shows the utilized data set, Sect. 4 describes the pre processing steps performed on the data set while Sect. 5 describes Simulation results and discussion. The final section concludes the paper.

2 Research Methodology

The original retail data set consists of 5,25,461 records for 2009–10 and 5,41,910 records for 2010–11. We determined 3940 distinct customer ids in the year 2010. There were 151 outliers which yielded a total of 3879 records in the filtered data set which was used for clustering computations. The first step is Data preparation and pre processing. The second step is to generate the target data set which has the distinct customer ids and the computed RFM values for each customer id. Once the target data set is generated we perform the cluster analysis using both the k means and self organizing maps. This analysis is performed using WEKA version 3.8.3. In order to get the target data set from the raw data set we have used the Excel data set in conjunction with MS Access. After the generation of the target data set, removal of outliers and normalization was done to obtain the final filtered data set which was the input for the two clustering algorithms- the K means and Self organizing maps.

The Computations in this study occur in two phases:

Phase 1: Cluster Profiling.
Phase 2: Performance benchmarking.

In Cluster Profiling we start by setting K = 4 for generating the cluster profiles, for both k means and self organizing maps. The Objective is to demonstrate the generation of cluster profiles for each of these techniques. The four clusters generated for each technique represent knowledge gained from the analysis- Distribution of data instances across clusters, Total monetary value represented by each cluster, Mean values of R, F, M, and FP for each cluster and mean spending per customer for the cluster. This gives a detailed profile of the generated clusters.

In Performance benchmarking, we compare the performance of K means and SOM by the following parameters: Execution Time, Number of iterations, Space complexity and Time complexity. The value of K is now varied for K = 2, 4 and 6. The corresponding values of these parameters are computed, tabulated and bar charts are drawn for Execution time and Number of iterations. These computed values benchmark the performance of these two clustering techniques. The Time and Space Complexity are computed for each.

The methodology adopted is this study different from Chen et al. due to the factors of introducing a neural networks computational framework in the form of SOM in addition to using K means clustering, Analysis of the Data set not studied by Chen et al. (2010), and implementing performance benchmarking of the two clustering techniques for the data set. The detailed steps contained in Data preparation and target data set generation are described in the next two sections.

3 The Data Set

The online retail data set was uploaded on 21st September 2019 on the UCI machine learning, repository. The original data set is processed suitably to create a Target Data set which is then analysed to generate the clusters. While the original data set contained 11 attributes we selected six attributes for starting the data preparation as shown in Table 1.

Table 1. Data attributes for our ML algorithms

Name	No. of Digits	Description
Invoice Number	6	Identifies each transaction uniquely
Item code	5	Identifies each product uniquely
Quantity	Numeric	The quantity per item purchased by a customer
Unit price	Numeric	Price per unit of an item
Invoice Date	Date	Date and Time of each transaction
Customer id	5 digit	Uniquely identifies each distinct customer

The Customer id is used instead of the Post code as Post code is subject to the data protection laws of the UK.

4 Data Pre Processing

The next step is to create a number of variable for our machine learning (ML) algorithms including the Amount which is calculated as the Quantity in to Unit Price. The Amount is computed for each distinct customer id for the country = UK. We then computed the number of distinct customer ids in our 2010 data set. Next segregate Date & Time components of the Invoice Date data so that distinct date & time values can be obtained for the transactions in the data set. Considering only the UK transactions we delete records with no customer id and also any missing records. Three aggregate variables have been created including recency (r), frequency (f) and monetary (m). These have the following interpretation:

Recency (r): Measures the recency of the transactions made by any customer value is in months.
Frequency (f): Measures the frequency of the purchases made by a customer over a time period in our case the year 2010.
Monetary (m): Measures the total amount spent by a customer across transactions over the year 2010.

First Purchase: Time in months since the beginning of 2010 when the first purchase is made by a customer. Accordingly the Target Data set consists of the following five attributes: Customer id, Recency, Frequency, Monetary and First Purchase. The work flow of our approach is shown in Fig. 1, while Algorithm 1 illustrated the proposed methodology.

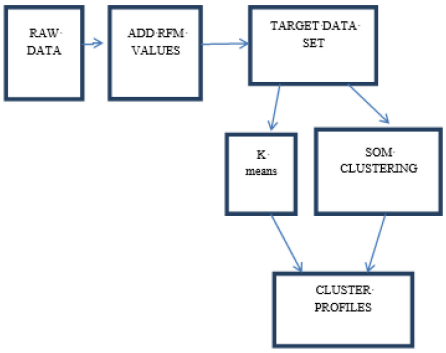

Fig. 1. The work flow

There are two computational tasks in this study – Cluster profiling and Performance benchmarking. The Cluster Profiling is done using simple K means and SOM clustering with K set at 4, and uses the target data set which consists of five attributes – namely: Cust_id, R,F,M and FP.

Performance bench marking for both the techniques is done using 4 key parameters of Execution time, Number of iterations, Space complexity and Time complexity with K varying from 2 through 4 to 6. The computed values of these computations are tabulated in Table 10.

Algorithm 1: Our proposed methodology for the analysis of online retail data.
Let X represents a set of retail data for online shopping customers where
X = {Invoice, Item code, Description, Quantity, Price, Invoice Data, Customer ID}
Let $C \subset X$, a set of customers with a number of f transactions.
$C = \{c | c \text{ has } f > 0\}$
$\forall c \in C, \exists$ r and m $\Rightarrow$ m is the monetary and r is the recency.
now add r, f and m to the X, X become X1 for online shopping customers where
X1 = {Customer ID, Monetary, Frequency, Recency, First Purchase Month}
$\forall c \in C, \exists$ outlier removal of X1.
$\forall c \in C, \exists$ normalization of X1.

Let ML to be our machine learning set
ML = {K means, SOM}
∀ml ∈ ML, find c cluster using, r, f, m and fp where fp is the first purchase

Algorithm 2: Our proposed methodology for Performance benchmarking for the K means and Self organizing maps.
Let X1 represents a set of retail data for online shopping customers where
X1 = {Customer ID, Monetary, Frequency, Recency, First Purchase Month}
For cluster Ck (k = 2, 4, 6)
calculate Ck from algorithm 1
∀ml∈ ML, find E, N, S and T

where E is Execution time
where N is No of iterations
where S is Space complexity
where T Time complexity.

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results. so we did outliers removal based on interquartile ranges. We determined 3940 distinct customer ids in the year 2010. There were 151 outliers which yielded a total of 3879 records in the filtered data set which was used for clustering computations. we also did normalization, The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Normalization is also required for some algorithms to model the data correctly. We normalized data followed by outlier removal using weka tool. Normalization and outliers removal are provided in Fig. 2. Red data point is showing outliers in Fig. 2 (Figs. 3, 4, and 5).

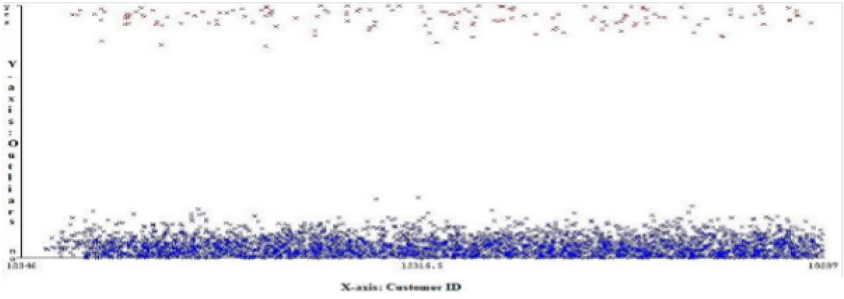

Fig. 2. Determining outliers

5 Simulation Results and Discussion

In this section, the simulation results for utilising K- mean and Self - organising map are presented. We utilised our filtered target data set for the analysis of the data. Table 2 shows the no of instances in each clusters for k means.

Table 2. K means (k = 4)

Cluster	Number of instances	%
0	748	20
1	1274	34
2	885	23
3	882	23

Table 3 shows the distribution of monetary value across the clusters for k means. Total Monetary for all clusters:285.024757.

Table 3. Total monetary value

Cluster	*Total Monetary by cluster*	*%*
0	53.807791	18.88
1	92.789915	32.56
2	69.51373	24.38
3	68.913321	24.17

Table 4 shows the RFM value computed during K means clustering.

Table 4. RFM values for K-Means clustering

Cluster	*R*	*F*	*M*	*FP*
0	0.728	0.0048	0.0719	3.28
1	0.1449	0.007	0.0728	9.61
2	0.1292	0.0373	0.0785	3.05
3	0.1483	0.0411	0.0781	2.97

Table 5 shows the mean spending per customers for each cluster for k means.

Table 5. Mean spending per customer for k-means

Cluster	*Mean spending per customer*
0	0.0719
1	0.0728
2	0.0785
3	0.0781

As seen above we get the following result:

According to the total monetary value the highest monetary value is in Cluster 1, the second highest in cluster 2 and the lowest in cluster 0. The highest mean spending per customer is in cluster 2, the second highest in cluster 3 and the lowest in cluster 0.

In this case

Cluster 0: Lowest by monetary value and mean spending per customer. Low recency and low frequency.
Cluster 1: Highest group by monetary value but the second lowest by mean spending per customer. High recency and higher frequency.
Cluster 2: The second highest group by monetary value and the highest group by mean spending per customer. high recency and higher frequency.
Cluster 3: Similar to cluster 2 in reference to monetary value and second highest group by mean spending per customer. High recency and medium frequency.

For the Self Organizing Maps we have utilised similar to K-means 4 clusters.
Table 6 shows the no of instances in each clusters for SOM.

Table 6. SOM (k = 4)

Cluster	*Instances*	*%*
0	695	18
1	966	25
2	808	21
3	1320	35

Table 7 shows the distribution of monetary value across the clusters for SOM.

Table 7. Total monetary value

Cluster	Monetary	%
0	50.648962	17.78
1	70.868476	24.86
2	105.24944	20.44
3	105.24944	36.92

Total monetary value of clusters: 285.024757.
Table 8 Shows the mean spending per customers for each cluster for SOM.

Table 8. Mean spending per customer for SOM

Cluster	*Mean spending per customer*
0	0.0729
1	0.0734
2	0.0721
3	0.0797

Table 9 shows the RFM value computed during SOM clustering.

Table 9. RFM values for SOM clustering

Cluster	*R*	*F*	*M*	*FP*
0	0.0807	0.0059	0.0729	10.61
1	0.2065	0.0124	0.0734	7.32
2	0.7098	0.0058	0.0721	3.24
3	0.11	0.0462	0.0797	2.23

According to the total monetary value per cluster the highest monetary value is in cluster 3 which also has the highest mean spending per customer. The second highest monetary value is in cluster 1 which also has the second highest mean spending per customer. The lowest monetary value is in cluster 0 which has the lowest mean spending per customer described as follows

Cluster 0: lowest group in monetary value and lowest by mean spending per customer. High recency and low frequency.
Cluster 1: The second highest group in terms of monetary value and also by mean spending per customer. High recency and higher frequency than cluster 0.
Cluster 2: The second lowest group by monetary value and the lowest by mean spending per customer. Low recency and low frequency.
Cluster 3: The highest group by monetary value and also by the mean spending per customer. High recency and high frequency.

This completes the cluster profiles for the Self organizing maps with 4 clusters.
The following cluster plots show the visual assignment of RFM for k means and som.
Frequency Plot for k means and som. X- axis represent Frequency and Y-axis represent different cluster assignment.

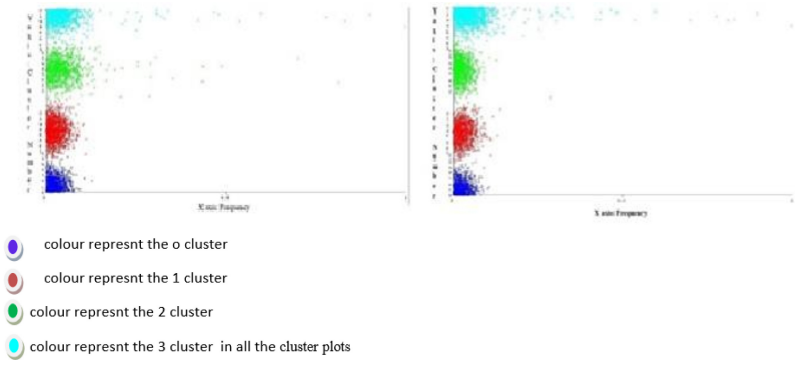

Fig. 3. Frequency plot

Monetary plot for k means and som. X- axis represent Monetary and Y- axis represent different cluster assignment.

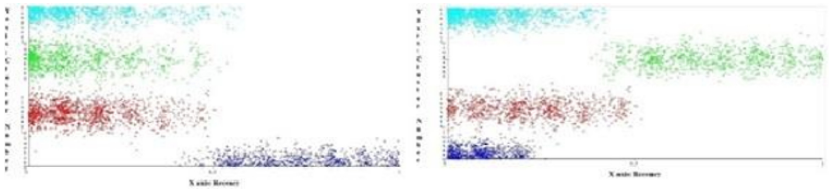

Fig. 4. Monetary plot

Recency plot for k means and SOM. X- axis represent Recency and Y- axis represent different cluster assignment.

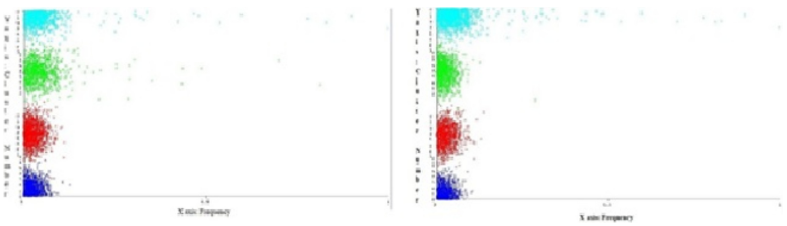

Fig. 5. Recency plot

Comparison between K means and self organizing maps (SOM).

While we have demonstrated Cluster profiling and the related computations for K = 4, in the case of both K means and self organizing maps, we now proceed to bench mark the performance of these two clustering algorithms.

For K = 2,4,6:

1. Compute Execution time and No of iterations for both K means and SOM.
2. Compute the Space and Time complexity for both K means and SOM.

The results are displayed in the form of Histograms in Figs. 6 and 7 along with Table 10.

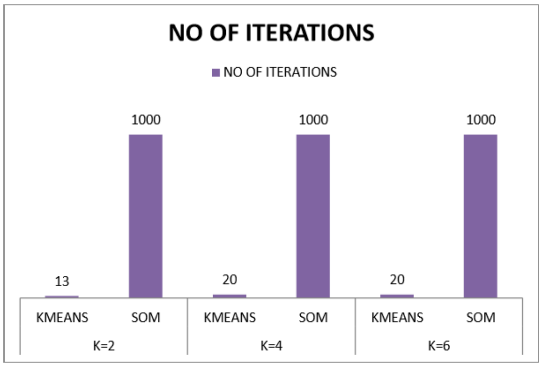

Fig. 6. SOM and k means iteration comparison

Fig. 7. SOM *and* k-means *execution time comparison*

The k means and SOM algorithm were compared on the basis of different parameters computed for the Data set and shown in the Table 10.

Linkage to Chen et al. and Differences

Chen et al. [1] analysed the same retail data set for 2011 and used k means clustering to segment the customers of the online retail store. While they tried with k = 3, 4 and 5, it concludes that the results obtained for k = 5, have a clearer understanding of the target data set than the results for k = 3 and k = 4 (Table 11).

Table 10. Comparison between K means and Self Organizing Maps

	K = 2		K = 4		K = 6	
	KMEANS	SOM	KMEANS	SOM	KMEANS	SOM
Execution Time	0.14 s	5.2 s	0.12 s	10.61 s	0.09 s	15.66 s
No of Iterations	13	1000	20	1000	20	1000
Space Complexity	O(18955)	O(7582)	O(18965)	O(7586)	O(18975)	O(7590)
Time Complexity	O(492570)	O(15156000)	O(985140)	O(30312000)	O(1477010)	O(45468000)

Table 11. K means Clustering results of *Chen et al.*

Cluster	*Instances*	*%*
1	527	14
2	636	17
3	1748	47
4	627	17
5	188	5

The relative contributions of these clusters to Monetary have also been described in the paper and there after cluster profiling in terms of r, f and m values has been done. The approach taken by this paper is to demonstrate the clustering of the customers by both K means and self organizing maps and thereafter to profile the clusters obtained in terms of r, f and m values. However it has been done for 2010. Also the performance of K means and Self Organizing Maps has been benchmarked and compared for k = 2, 4 and 6 as seen in Figs. 9 and 10 and Table 10, showing the computed values of time and space complexity for both these clustering algorithms.

Also in the case of the Self organizing maps, N = H * W, where N is the number of clusters, H is the height of the lattice and W the width of the lattice. Thus we have ensured that the comparison is done for 2, 4 and 6 clusters for both K means and Self Organizing maps. In the case of the SOM, this results in the creation of a 2*1,2*2 and 2*3 lattice, facilitating further computations. This facilitates the comparison of the two algorithms for the same parameters. Accordingly we chose K means algorithm as it is simple and popular amongst practitioners and was used by Chen et al. alto do the analysis. It allows us to compare results obtained with those of Chen et al. for the k means algorithm. In addition self organizing maps were chosen to give a Neural Networks perspective and frame work due to its robust architecture and lesser sensitivity to noise in the input data. It allows us to compare the working of the clustering algorithm using self organizing maps for the same data set. We can also gain insights on the computations done using a Neural Networks frame work and then compare the results obtained from these two key techniques.

6 Conclusions

This research paper based on the work done by Chen et al. [1] who used K means clustering to obtain a segmentation of customers for an online retailer. While the concerned data set was up loaded on the UCI Machine learning repository on the 21st of September 2019, the analysis by Chen et al. covered the retail data set for 2011. It uses the RFM model to construct a Target data set containing distinct post codes. There after decision tree based rule induction was used to obtain decision rules representing customer specific business intelligence. Clear and stable results representing the under lying data set were obtained for k = 5. In this paper we selected the data set for 2010 and then obtained the number of distinct customers who did transactions with the online retailer over the year 2010. We constructed the target data set and performed Normalization and removal of outliers. We performed K means clustering for k = 4 and then used Self organizing maps with number of clusters = 4. The clusters obtained were profiled in terms of their RFM values and the mean spending per customer. The highest and lowest monetary value clusters were identified. The K means and Self organizing maps clustering algorithms were compared for their performance for 2, 4 and 6 clusters and the results were tabulated in the histograms of Fig. 9 and 10, along with Table 10, depicting the time and space complexity for the two clustering algorithms. In reference to this research paper, further work can be done in identifying buying patterns of customers in terms of items purchased (association rules). Also a buyer loyalty program can be designed based on the buying choices made by customers and there after high value customers can be identified. The design of such a loyalty based card membership can increase the popularity and visibility of the retailer in terms of their business operations. Finally, advanced techniques of machine learning such as Deep learning can be used to design new computational architectures and obtain new results. Fuzzy learning techniques can also be used to determine which paradigm to select to obtain better and more accurate results with greater efficiency and speed.

References

1. Chen, D., Sain, S., Guo, K.: Data mining for the online retail industry: a case study of RFM model-based customer segmentation using data mining. J. Database Mark. Customer Strategy Manage. **19**(3), 197–208 (2012) https://doi.org/10.1057/dbm.2012.17
2. Dogan, O., Ayçin, E., Bulut, Z.: Customer segmentation by using rfm model and clustering methods: a case study in retail industry. Int. J. Contemp. Econ. Adm. Sci. **8**(1), 1–19 (2018)
3. https://Emarketer.com/content/uk-ecommerce-2019, read on 14 November 2019
4. https://Statista.com/statistics/275973/types-of-goods-purchased-online-in-great-britain/, read on 14 November 2019
5. https://archive.ics.uci.edu/ml/index.php, online_retail II, This is the data set used by Chen et al, which has also been used in this paper. The Data set was uploaded on 21 September 2019
6. Hu, Y.-H., Yeh, T.-W.: Discovering valuable frequent patterns based on RFM analysis without customer identification information. J. Knowl. Based Syst. **61**, 76–88 (2014). https://doi.org/10.1016/j.knosys.2014.02.009

7. Sarvari, P.A., Ustundag, A., Takci, H.: Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes. **45**(7), 1129–1157 (2016)
8. Yeh, I.C., Yang, K.J., Ting, T.M.: Knowledge discovery on RFM model using Bernoulli sequence. Expert Syst. Appl. **36**, 5866–5871 (2008)
9. Wei, J.-T., Lin, S.-Y., Hsin-Hung, W.: A review of the application of RFM Model. Afr. J. Bus. Manage. **4**(19), 4199–4206 (2010)
10. Bloom, J.Z.: Market segmentation – a neural network application. Ann. Tourism Res. **32**(1), 93–111 (2005)
11. Holmbom, A.H., Eklund, T., Back, B.: Customer portfolio analysis using the som. Int. J. Bus. Inf. Syst. **8**(4), 396–412 (2011)
12. Kohonen, T.: Self Organizing Maps. Springer Verlag, Berlin (2001)
13. Vellido, A., Lisboa, P.J.G., Meehan, K.: Segmentation of the online shopping market using Neural networks. Expert Syst. Appl. **17**(4), 303–314 (1999)
14. Ultsch, A.: Self organized feature maps for monitoring and knowledge aquisition of a chemical process. In: Gielen, S., Kappen, B. (eds.) ICANN 1993, pp. 864–867. Springer, London (1993). https://doi.org/10.1007/978-1-4471-2063-6_250
15. Miljkovic.: Brief overview of Self organizing maps. In: Proceedings of 40th International conference on information and communication technology, electronics and micro electronics (MIPRO), IEEE (2017)
16. https://www.cs.waikato.ac.nz/ml/weka/
17. Kiang, M.Y., Hu, M.Y., Fisher, D.M.: An extended self-organizing map network for market segmentation—a telecommunication example. Decis. Support Syst. **42**, 36–47 (2006)