Comparative Unsupervised Clustering Approaches for Customer Segmentation

Asmin Alev Aktaş
Research and Innovation
Ata Technology Platforms
Istanbul, Turkey
asmina@atp.com.tr

Okan Tunalı
Research and Innovation
Ata Technology Platforms
Istanbul, Turkey
okant@atp.com.tr

Ahmet Tuğrul Bayrak
Research and Innovation
Ata Technology Platforms
Istanbul, Turkey
tugrulb@atp.com.tr

2021 2nd International Conference on Computing and Data Science (CDS) | 978-1-6654-0428-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDS52072.2021.00097

***Abstract*—Machine learning-driven studies to get potent insights about customers are essential for the business world to grow as they achieve smarter in marketing and sales activities. Finding the consociate patterns of customer interaction activities leads to finding sensible segments. By this, strategists can reach out to different groups of customers with customized services, offers and plans. However, although clustering algorithms are reliable by virtue of them being competent studies, not all of them fit the studied domain. In this study, six well-known clustering algorithms with different parameters are applied to real-life customer purchase history data. The outcomes are compared, and the density distribution of data features in created clusters are visualized. Thus, it is possible to see the role of each selected feature on the differentiation of clusters. The cluster labels of data points (customers) are mapped in pairs of algorithms. As a result, the similarities and differences in clusters created by different algorithms are more straightforward to catch. Moreover, in addition to labeling data points with class labels, a hybrid approach is presented to obtain information about class label probabilities by fitting the support vector classification model. The proposed study gives promising results in understanding how different clustering algorithms fit the customer data and stands out with multi-sides evaluation and comparison experiments.**

***Index Terms*—k-means, spectral clustering, Gaussian mixture, agglomerative clustering, ward, mean-shift, support vector classifier**

I. Introduction

Understanding customer behavior and categorizing customers based on their purchasing histories is essential for the majority of commercial industries. This is even more crucial when it comes to a dynamic and constantly changing sector of fast-food. Clustering allows marketers to identify groups of customers sharing similar purchasing habits in their customer base. It reveals deep insights that can not be seen by examining all customers simultaneously. It is critical to segment customers as it allows marketers in their decision-making to better adapt their marketing efforts to various audience subsets in terms of sales, promotion, and campaign development strategies. For instance, whilst it is not possible to produce a single campaign that meets all customers' needs, more customer-specific campaigns can be produced for different groups by identifying collective patterns of customer behavior.

However, dealing with clustering algorithms has its own challenges, [1], [2], [3] when it comes to finding the most suitable clustering models for the real data as well as finding the optimum parameters. The ideal clusters may often be defined by what seems sensible given the problem domain and application. [4] finds recency, frequency, and monetary (RFM) metrics to apply the k-means algorithm for segmenting retail industry customers. A different approach can be seen in [5]. The study presents a complex network approach for data clustering. Dataset is represented as a network, taking into account different metrics for linking each pair of objects. The presented clustering approach is applied in both real-world and artificially generated data sets. One comparison based study [6] applies nine clustering methods available in the R language for systematic comparison on data assumed to be normally distributed. Another study [7] compares four clustering algorithms and finds the similarities between created clusters.

In our study, six different clustering algorithms are selected to apply to the customer data for a comparative application, and dominating features for each created cluster are visualized. The question of how features affect the creation of clusters is replied to. The distributions of features in each cluster and the similarity of clusters are shown with figures and mapping tables.

Evaluating the performance of the unsupervised algorithms is challenging as they look for previously undetected patterns. Clustering algorithms only return a label without providing a probability or confidence interval. Conversely, supervised learning methods like support vector classifier (SVC) allow the probability calibration of classifiers.

In this study, after labeling data points with specific cluster centers, SVC is fitted to the labeled data. As it is expected, the SVC learned from the clustering algorithms and overfitted. Thus, it behaved like the clustering algorithm itself with an extra feature to find the probabilities of the clusters as well. In addition to research, analysis and comparison, an innovative study is carried out to add a new function to the clustering system. With the new development, not only cluster labels but also cluster probabilities for each data point are calculated.

One key aim for our future studies is to automate the segmentation of customers. Finding the valuable patterns for customer behavior recognition and clustering customers according to these patterns can lead the study for further developments.

978-1-6654-0428-0/21/$31.00 ©2021 IEEE
DOI 10.1109/CDS52072.2021.00097

A study [8] is presented on automatic classification systems. Applying this to a dynamic and rapidly changing domain can save a lot of time and money in setting up business strategies.

The rest of the paper is organized as follows: In section II, the data used in the study and the preprocessing steps are explained. In section III, the research methodology is described in detail. In section IV, the experimental results are presented. The paper concludes with a discussion of the findings and an overview of the direction for future work.

II. Data and Preprocessing

The dataset consists of 234320 rows of anonymous customers' online purchase history records from five different fast-food restaurants for three years. Selected features are *purchase_date*, *customer_id*, *quantity*, and *unit_price* for the products purchased. For this study, instead of combining behavioral and demographic variables in cluster analysis, only purchase history data is considered. Data statistics can be followed in Table I.

TABLE I: Statistical Information of Data

stats/features	quantity	priceunit	purchase amount
count	234320.00	234320.00	234320.00
mean	1.021	8.326	8.435
std	0.297	6.504	6.689
min	1.000	0.000	0.000
max	27.000	37.500	89.850

A. Feature Extraction

The purchase amount is calculated by multiplying quantity and the unit's price and added to the data set as a feature. To retrieve the last feature set, *recency*, *frequency* and *monetary* (RFM) analysis is chosen as it stands out as a preferred method based on the literature to extract information from customer's purchases. Furthermore, the active months metric (time between first purchase to last purchase, in months) is calculated to find the *period* which is the purchase number over the active months. RFM metrics have proven to be effective predictors of a customer's willingness to engage in marketing messages and their offers [9]. Recency is the time between the last purchase and present or until the last date data is collected. Frequency is the measurement of how often a customer purchases. Lastly, monetary is the total amount of money a customer spends.

B. Outlier Elimination and Normalization

Outliers are observations or suspicious measures because they are either much smaller or much larger than the vast majority of the observations [10]. For our analysis or statistical tests, it is more effective to remove the outliers from the data as part of data preprocessing. Any outlier in data may give biased or invalid results that can impact the performance of the machine learning methods. To find the customers period of the visit, only customers who have had more than two visits (minimum of three visits) are considered. The ones who do not match the requirements are eliminated. For the rest of the features, Z-score is used to remove outlier values. The values that are above three standard deviations far from the mean are discarded. After these steps, 1034 customers left and the data became more consistent.

Log scaling is done for normalization. It calculates the log of values to squeeze a wide range into a narrow range. It is useful [11] when different features vary in different ranges and the gap between feature values are tremendous. For our data set, while some customers have only a few purchases and a small amount of monetary, regular customers have a large scale of purchases over the three years. Additionally, having data in a shorter range makes it easier to read and looks tidier for visualization.

III. Methods

For the first part of the study, six different clustering algorithms are applied to the same data, and the distributions of the features around the cluster centers are visualized. It is possible to see different algorithms having similar cluster characteristics, and the same features are distributed in similar ranges. However, because of labeling with different numbers, whilst one algorithm labeled a similar cluster as 1, another might label it as 0. It is hard to capture all the similarities from the figure. To see the similarities and distinctions clearly, a mapping algorithm is developed, and pair tables are created.

The algorithms that are applied:

- MiniBatchKMeans
- SpectralClustering
- GaussianMixture
- AgglomerativeClustering
- MeanShift
- Ward

After comparing all the algorithms, it is detected that MiniBatchKMeans (MB), Spectral Clustering (SC), and GaussianMixture (GM) have at least one similar cluster. The details will be shared in the experiment section.

To create a system where the cluster labels and the probabilities of them are obtained, SVC is applied. First of all, the data is labeled with each of the algorithms that have been used in the previous part of the study. Later, SVC is applied and as is expected, it fits almost perfectly for all of the algorithms. This means it over-fitted and worked exactly the same with the clustering algorithms that fitted. Thus, whenever a new data point is checked to find which cluster it might belong to, the SVC model performs the same with the clustering algorithm that it has been trained with. The SVC model then finds the data points' probabilities of belonging to each cluster along with their cluster labels.

For the second part of the study, both SVC results are shared, which are trained on datasets labeled with SpectralClustering (SC) and Agglomerative Clustering (AC) cluster centers outcomes. The main parameters for SC are the number of clusters. The proximity that the nearest_neighbors is used for this study to construct the affinity matrix by calculating a graph of the closest neighbors. Using SC to find customer groups from transaction data is also presented in the study [12] and a two-level subspace weighting spectral clustering (TSW) algorithm is displayed.

The parameters for AC are 'average' for linkage, 'euclidean' for affinity and the rest is used from the previous default parameter set. AC is a type of hierarchic clustering method that is a well-fit algorithm for the pattern and cluster mining on text data [13], [14]. It is difficult to predict if AC is suited to the data set used for this study as it places all but nine customers into one group. However, SVC fits perfectly and performs like AC after being trained with the dataset labeled with AC.

In our study, the concept of within-cluster sum of squares (WCSS) is used to decide the number of clusters. This is derived using the concept of minimizing WCSS. As the number of clusters increases, the WCSS keeps decreasing. The change in the difference of WCSS is calculated:

$$0.160, 0.116, 0.0748, 0.0198, 0.040, 0.045, 0.001, -0.003 \tag{1}$$

The difference in chances dramatically decreases, starting from k=3. To keep the consistency with the previous part, the dimension of the projection subspace is chosen to be three for SC and AC applications. From the business side, this number may vary depending on the budget and the number of campaigns to be produced to reach out to different groups. For this study, three is a convenient number to see the feature distributions within the clusters in a clear visual to analyze the dominant features in shaping the clusters.

The SVC module supports different kernels for classification problems. For this study, after labeling data with three cluster centers, SVC is applied using three different kernels (linear, poly, RBF), and the performances are calculated. It is observed that for each kernel, SVC fits perfectly to the trained data. However, a linear kernel is used as it fits slightly better than other kernels.

Why finding probabilities is important is a question related to the business problem. If customers are not required to cluster strictly, then the ones found with low probabilities are also considered in the same segments. If not, there might be different business plans for customers who do not strictly belong to any clusters.

IV. Experiments

A. Comparison of Clustering Algorithms

A controlled experiment is done to compare the approaches of different clustering algorithms from python libraries. Judging the performance of a clustering algorithm is only possible in the context. For an unsupervised technique, it is difficult to say one algorithm is better than the others. The more fruitful algorithm for the data set can give the most consistent results from the business perspective. This is why seeing the feature distributions, as well as each feature's role in creating the clusters, is important. For example, if a business case is required to cluster customers by their monetary, then the algorithm that creates the clusters mostly based on monetary would be a good choice. This is also important in defining the customers with low engagement. For instance, if a cluster includes customers whose recency values are higher than usual, it can be considered as a warning in defining the group of people before they leave. Finding the common pattern of customers who are about to leave is an important problem as [15] study presents, finding these critical groups before they churn saves more money and effort than getting them back after losing them. Thus, the visualization of features probability density plots are essential for the study and the developed code[1] can be found in GitHub.

In the beginning, default parameters are used. These parameters are called by the methods, and only if there are some specific parameters needed, they are defined in the object creation step for each algorithm separately.

Default parameters:

default_base = {'quantile': .3, 'eps': .3, 'damping': .9, 'preference': -200, 'n_neighbors': 10, 'n_clusters': 3, 'min_samples': 20, 'xi': 0.05, 'min_cluster_size': 0.1

Different from this parameter set, for Spectral Clustering eigen_solver='arpack' and affinity="nearest_neighbors" are used. For Gaussian Mixture, covariance_type='full' is selected as each component is considered to have its own general covariance matrix. For Aglaromerative Clustering, the linkage is chosen to be "average". The criterion of linkage is used to determine what distance is used between the observed data points. Lastly, for MeanShift (MS), 'quantile', which is used for bandwidth estimation, is tried with different numbers. It is noticed that '0.18' is a good value to get three clusters. Selecting bigger than 0.18 resulted in less number of clusters being created. In contrast, smaller kernel bandwidth resulted in more clusters.

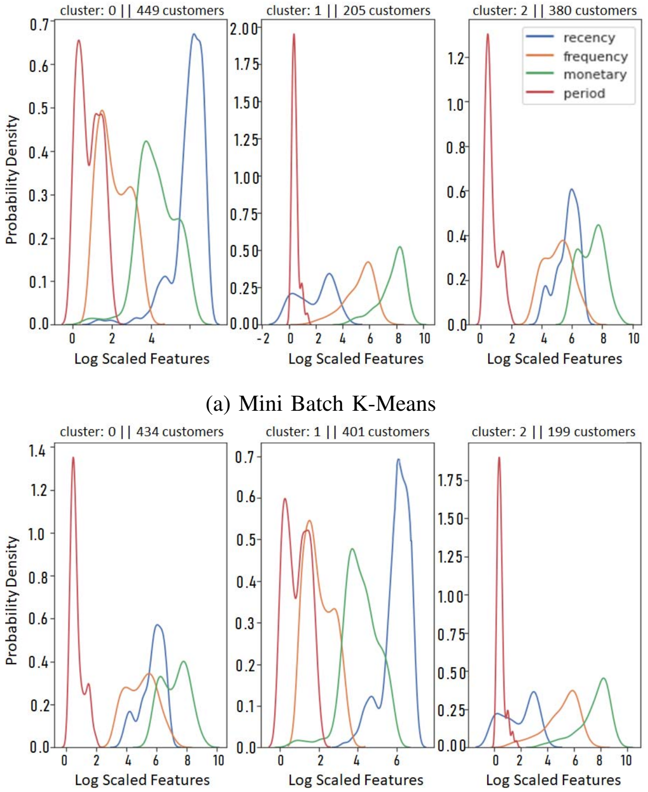
Each of the algorithms offered a different approach to the difficulty of discovering natural groups in data. However, some algorithms are observed to create similar cluster characteristics. This is understood as they have similar probability density distributions of features.

The visual is a density plot showing logged features values on the x-axis and the probability density of feature values on the y-axis. Features are shown in different colors. Blue is used for recency, green for monetary, red for period and orange is used to show frequency.

Fig. 1 displays the feature distributions of MiniBatchK-Means on Fig. 1a and SC algorithm on Fig. 1b. For SP, the affinity parameter is also tried with 'rbf' and the result does not change considerably. Feature distributions and cluster volumes remained similar.

For Fig. 1a, the class that MB labels with 2 have similar characteristics with the class of the Spectral Clustering Algorithm with the label 1 in Fig. 1b. For the rest, it is possible to see that similar clusters are found:

[1] https://github.com/asminalev/ClusteringVisualisation



(a) Mini Batch K-Means

(b) Spectral Clustering

Fig. 1: MiniBatchKMeans and SpectralClustering

0, 1, 2 —-Labels from the MiniBatchKMeans
1, 2, 0 —-The corresponding labels from Spectral Clustering

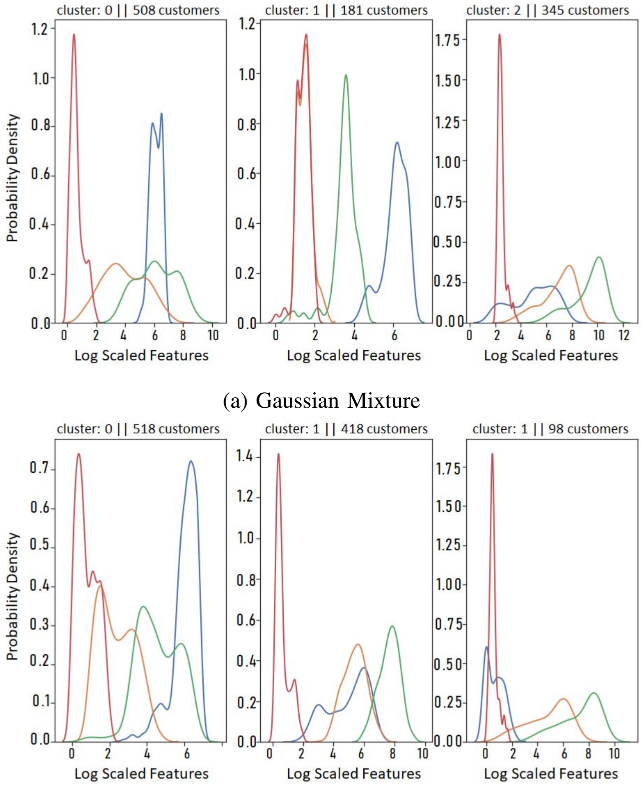
To see all the similarities in numbers, a mapping algorithm is applied, and the results can be seen in Table II.

TABLE II: MiniBatchKMeans and SpectralClustering Points Map

Mini Batch K-means	SpectralClustering		
	0	1	2
0	40.0	401.0	8.0
1	14.0	0.0	191.0
2	380.0	0.0	0.0

On the table, 401 data points are labeled with 0 by the first algorithm and labeled with 1 by the second algorithm. The 191 data points are labeled with one by the first algorithm and they are labeled with 2 by the second algorithm. For the last cluster, 380 data points are labeled with 2 by MB and 0 by SC algorithms. The table is consistent. This shows that both algorithms agree on most of the data points being in the same class.

For the next pairs, Gaussian Mixture and Ward are compared in Fig. 2. The cluster labeled with 2 has a similar distribution for features except for recency (shown in blue).



(a) Gaussian Mixture

(b) Ward

Fig. 2: GaussianMixture and Ward

However, their volumes are different. While the first algorithm assigns 345 data points to class 2, the second algorithm only assigns 98 points in class 2. The rest of the clusters seems to have different distributions and volumes. Table III shows these observations with numbers to see it clear.

TABLE III: Gaussian Mixture and Ward Centres Points Map

Gaussian Mixture	Ward Clustering		
	0	1	2
0	298.0	210.0	0.0
1	181.0	0.0	0.0
2	39.0	208.0	98.0

As we see from Table III, the data points that Gaussian Mixture is labeled with 0 are labeled with 0 or 1 by the second algorithm. It is hard to say that two clusters map each other. Another algorithm used is Agglomerative; the results of this algorithm are presented in Fig 3. It labeled almost all the data points with the same cluster label.

The volumes of the clusters are not balanced. To see the effects of the different connectivity parameters, the connectivity matrix itself and kneighbors_graph are used. However, the result does not change significantly. In agglomerative clustering, the connectivity matrix is used as the constraint when finding the clusters to merge [16]. It defines which samples are considered connected and creates a larger cluster by following a given structure.

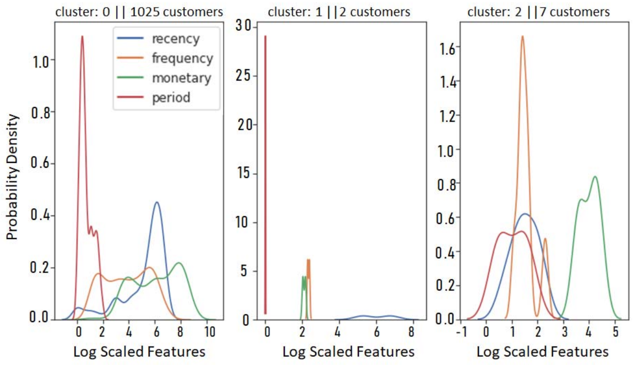


Fig. 3: Agglomerative

Next visual shows the feature spectral distributions for the MeanShift algorithm in Fig. 4.

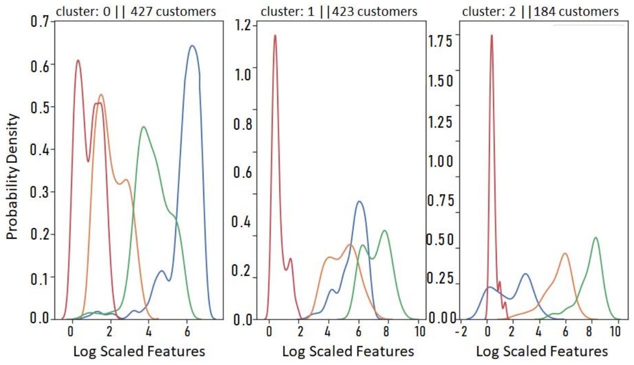


Fig. 4: MeanShift

MS needs the longest time to run. It takes five seconds to run while all other algorithms give results within a second. Considering the data not being enormous, it might not be the best option for large data sets. Moreover, the clusters are quite similar to SC clusters by both distributions and volumes.

Finally, we observe that one of the most common clusters is the one that MB labeled as 1, SC and GM labeled as 0, despite some little differences in the number of data points. Another general observation is that recency distributes in different ranges in most of the clusters. It shows that recency is an important metric in defining different patterns.

B. SVC Application on Labelled Data

In the previous part, after comparing the clustering algorithms, data is labeled with each of the algorithms separately. Here the results of SC and AC are presented. SP and AC are chosen to present to show that it does not matter how the customers are distributed to the different clusters. SVC always learns from the clustering algorithm perfectly.

TABLE IV: Accuracy of SVC models after fitting on labeled data by SC and AC

Algorithm	Performance Metric	Cluster Label		
		0	1	2
SC	precision	0.96	1.00	0.98
	recall	1.00	1.00	1.00
	f1 score	1.00	0.97	0.98
	support	49	92	66
AC	precision	0.99	0.00	1.00
	recall	1.00	0.00	0.50
	f1-score	1.00	0.00	0.67
	support	204	1	2

The labeled data sets are split into train and test parts. Then, test parts are hidden from the models. To interpret classes probabilistically, the output of an SVC probability calibration is used. Here besides predicting the class labels, the probability of the respective labels is also obtained by performing classification. Probabilities for each data point are expected to give confidence in the prediction. The calibration module can be used to calibrate the probabilities of a particular model better. It also makes it possible to add support for probability estimation. This application is made to add a new function to the clustering algorithms to return the probability value for each data point that is decided to belong to a certain cluster.

Three different kernels are used for the kernel parameter of SVC. Performances results for each of them vary in a similar range of numbers between 0.96 to 1.00. For demonstration, a linear kernel is used. The precision, recall, f1 score and support values for SVC results after applying to the labeled data sets labeled by SC and AC algorithms can be seen from Table IV. The motivation here is to over-fit the SVC on the train data.

Moreover, the aim of the study is not only to see the performances of different kernels but to find the SVC model that fits perfectly to the clustering algorithm labeling results. By overfitting SVC, it starts to function like the clustering algorithm with an extra feature to find probabilities as well.

From the table, we see that SVC perfectly fits the SC results. For AC, SVC learns cluster 0 characteristics almost perfectly. However, since the training set does not include any sample from cluster 1, SVC does not learn that pattern and can not find it, which is expected when we think there is only one customer assigned to cluster 2. Furthermore, for cluster 2, because the training set includes samples from this cluster, SVC learns it with 0.50 recall. The test set had two customers from this class and it labeled one of them correctly, which is a high success despite 0.50 because it learns with only five samples.

In the following Fig. 5, test data that is labeled by SC but hidden in the training process of SVC is used to predict cluster labels and respective probabilities. Most of them, except six data points, are labeled with more than 80% probabilities. The rest has a probability between 50% to around 75%. It means that except for six customers, the rest are strongly sharing a similar pattern with their groups.

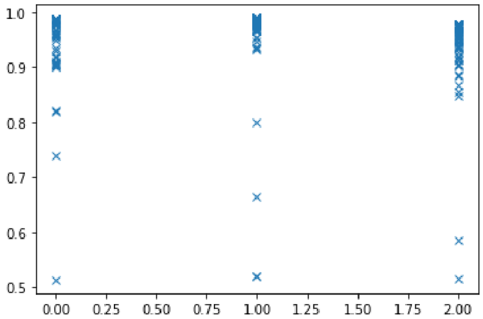


Fig. 5: SVC Results for SC Algorithm

V. Conclusion

Customer segmentation has almost unlimited potential that can lead companies to more effective ways in creating strategies of reaching potential customers, marketing products or coming out with new customized solutions. Knowing the customers' purchasing behavior is key in the decision-making process of determining product pricing properly, developing special marketing campaigns, designing distribution strategies in an optimal way, prioritizing new product development strategies, and managing advertising campaigns.

Finding the best-suited algorithms to cluster the customers is a key problem in finding a realistic approach. For the clustering algorithms, it is not possible to favor one over another. With this study, we show the similarities and differences between different methods. Comparing different approaches light the way to understanding the best-suited methods and parameters for the studied domain.

After comparisons and labeling data, finding the probabilities of each data point by using SVC, adding a new function to the clustering algorithm, through this, the cluster labels and the probabilities for each cluster are found.

This is an innovative work to modify well-known clustering algorithms with a new ability to solve important business cases, such as creating flexible and adjustable strategies and campaigns. If the customers are required to be clustered in a strict way according to the specific feature, only those who have high probabilities can be in those clusters. If the business offer is not so specific to a group, the customers who are not found to belong to any of the groups strongly can also be considered.

Overall, we see that among all of the clusters, the recency metric has a vital role in creating the clusters. It appears to have different pick points in each of the three clusters. In contrast, the period is observed to range in similar values for almost all of the clusters. It can be read that it does not matter how periods change when finding different patterns. It also shows that spectral clustering has a high potential to be appropriate for this data as it has the most common clusters with other algorithms. It can be understood that SC finds some dominant and agreeable cluster patterns.

MiniBatchKMeans and Gaussian Mixture are the other options that give balanced and sensible results. The proposed study gives promising results and stands out opportunities for further development. Although the data studied comes from the fast-food industry, the work is extendable to any business that deals directly with customer purchases. For future work, customer clustering using demographic information such as age, gender, etc., will be studied and explored. Furthermore, creating ensemble models to combine behavioral and demographic clustering and combining different approaches from different clustering models will add further value to the next steps.

Finally, automated clustering is aimed to be studied in the future. Since the domain is dynamic and rapidly changing, adjusting the groups and finding new patterns will help the business strategists catch up with trends and understand customers ahead of time.

References

[1] A. K. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, pp. 264–323, 1999.

[2] S. Bano and N. Khan, "A survey of data clustering methods," *International Journal of Advanced Science and Technology*, vol. 113, 04 2018.

[3] A. Nagpal, A. Jatain, and D. Gaur, "Review based on data clustering algorithms," 04 2013, pp. 298–303.

[4] O. Doğan, E. Ayçin, and Z. Bulut, "Customer segmentation by using rfm model and clustering methods: A case study in retail industry," *International Journal of Contemporary Economics and Administrative Sciences*, vol. 8, no. 1, pp. 1–19, Jun. 2018.

[5] G. F. de Arruda, L. da Fontoura Costa, and F. A. Rodrigues, "A complex networks approach for data clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 23, pp. 6174 – 6183, 2012.

[6] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. da F Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PLOS ONE*, vol. 14, no. 1, pp. 1–34, January 2019.

[7] A. A. Aktaş, A. T. Bayrak, O. Susuz, and O. Tunalı, "An application of unsupervised clustering approaches in customer segmentation," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–6.

[8] S. Theodoridis and K. Koutroumbas, "Chapter 13 - clustering algorithms ii: Hierarchical algorithms," in *Pattern Recognition (Fourth Edition)*, fourth edition ed., S. Theodoridis and K. Koutroumbas, Eds. Boston: Academic Press, 2009, pp. 653 – 700.

[9] A. Dursun and M. Caber, "Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis," *Tourism Management Perspectives*, vol. 18, pp. 153–160, 2016.

[10] D. Cousineau and S. Chartier, "Outliers detection and treatment: A review," *International Journal of Psychological Research*, vol. 3, 01 2010.

[11] S. Dehaene, V. Izard, E. Spelke, and P. Pica, "Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures," *Science*, vol. 320, no. 5880, pp. 1217–1220, 2008.

[12] X. Chen, W. Sun, B. Wang, Z. Li, X. Wang, and Y. Ye, "Spectral clustering of customer transaction data with a two-level subspace weighting method," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3230–3241, 2019.

[13] D. Agnihotri, K. Verma, and P. Tripathi, "Pattern and cluster mining on text data," in *2014 Fourth International Conference on Communication Systems and Network Technologies*, 2014, pp. 428–432.

[14] Y. Yaari, "Segmentation of Expository Texts by Hierarchical Agglomerative Clustering," Tech. Rep. cmp-lg/9709015, Jul 1999. [Online]. Available: http://cds.cern.ch/record/394095

[15] A. T. Bayrak, A. A. Aktaş, O. Susuz, and O. Tunalı, "Churn prediction with sequential data using long short term memory," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–4.

[16] J. Li, "Agglomerative connectivity constrained clustering for image segmentation," *Statistical Analysis and Data Mining*, vol. 4, pp. 84–99, 02 2011.