An Application of Customer Embedding for Clustering

Conference Paper · November 2022

DOI: 10.1109/ICDMW58026.2022.00019

An application of Customer Embedding for Clustering

Ahmet Tuğrul Bayrak
Data Science & Innovation
Ata Technology Platforms
İstanbul, Turkey
tugrulb@atp.com.tr

***Abstract*—Effective and powerful strategic planning in a competitive business environment brings businesses to the fore. It is important for the growth of the business to move the customer to the center by acting more intelligently in the planning of marketing and sales activities. In order to find customer behavior patterns, the use of clustering models from machine learning algorithms can yield effective results. In this study, traditional customer clustering methods are enriched by using customer representations as features. To be able to achieve that, a natural language processing method, word embedding, is applied to customers. By using the powerful mechanism of word embedding methods, a customer space is created where the customers are represented based on the products they have bought. It is observed that appending customer embeddings for customer clustering have a positive effect and the results seem promising for further studies.**

***Index Terms*—customer clustering, word embedding methods, customer embedding, feature engineering, k-means algorithm, rfm analysis**

I. Introduction

Customer clustering is a process to split customers into consistent groups which have similar characteristics such as shopping habits, lifestyle, food choices, etc. [1]. Moreover, customer clustering is one of the most principle strategic planning and marketing approaches where the grouping of customers is executed under different categories such as purchasing capacity and the interest to buy. The clustering is performed according to the similarity in customers in numerous dimensions related to the products under consideration. The more accurately and properly the clusters applied for targeting customers by a particular organization, the more successful the organization is in the industry. The main objective of customer clustering is precisely estimating the needs of a customer and therefore improving profitability by acquiring or producing products in the required amount at the time for the right customer at optimum cost.

Commercial enterprises benefit from customer purchase history data in order to understand customer behavior and categorize customers on this axis. Dealing with clustering algorithms in the literature also has its own difficulties [2], [3]. Especially, it is quite challenging to find the most suitable clustering models for real data and to decide on the optimum parameters necessary to maximize the success of these models. In addition, working with high-dimensional data can create difficulties in clustering as in every field [4]. In marketers' decision-making processes, an effective grouping of customers is crucial for targeting marketing efforts, developing advertising, promotion and campaign strategies, and for better adapting these strategies to various target audience subsets [5]. In the process of finding the best segmentation algorithm, the suitability of a particular model often depends on the domain of the problem. Considering the application area and the sector, it is appropriate to use the algorithm or systems in which the algorithms work together, which seems logical. This is the most vivid criterion in the evaluation of unsupervised learning models. Unlike supervised learning algorithms, it is not always possible to evaluate unsupervised learning models with certain metrics and calculations. Therefore, the algorithms that produce the most convenient and effective results are accepted as the most suitable for that field.

There are numerous studies in the literature that use the customer recency, frequency and monetary (RFM) method to segment retail industry customers using a K-means clustering algorithm for customer clustering. The expansion of RFM values; is the time elapsed since the last time of arrival, the frequency, and finally the monetary return. In other approaches, a complex network approach is presented for data clustering, where the dataset is represented as a network, taking into account different criteria for linking each pair of objects [6], [7]. The presented clustering approach has been applied both in the real world and in artificially generated datasets. In another comparison-based study, Rodriguez applies nine clustering methods available in R to make a systematic comparison of data that is assumed to be normally distributed [8]. Customer embedding also started to be applied in recent years. According to Xie et al. apply an unsupervised deep embedding for customer segmentation [9]. Likewise, Jagabathula et al. also use an embedding method for clustering [10].

Word embedding, a natural language processing (NLP) topic, calculates the similarity of words based on their frequency and order. With this mechanism, in recent years, different entities have started to be embedded beside words [11]. In particular, item2vec, where the similarities of the products are determined, is widely used in recommendation systems [12].

In our study, customers in a dataset are embedded using a word embedding mechanism. Customer embedding vectors have also been added as a feature in the clustering of customers

2375-9259/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDMW58026.2022.00019

that are better represented by embedding. In this way, success in clusters with RFM was compared.

The remainder of the study is as follows: in section II the details of the method and data used in the study are provided. In section III and section IV, the experiments and their results are discussed, and in section V and section VI, the model outputs are evaluated and the work done is summarized.

II. Data and Preprocessing

In the study, an open dataset that is available on Kaggle is used[1]. The dataset consists of e-commerce companies online transactions. The dataset consists of a total of 1014269 numbers of rows. There are six columns in the dataset. The details of the features;

- invoice_no: Invoice number (Nominal). A 6-digit integral number is uniquely assigned to each transaction
- stock_code: Product code (Nominal). A 5-digit integral number is uniquely assigned to each distinct product
- description: Product name (Nominal)
- quantity: The quantities of each product per transaction and (Numeric)
- invoice_date: Invoice date and time (Numeric). The day and time when a transaction is generated
- unit_price: Unit price (Numeric). Product price per unit
- customer_id: Customer number (Nominal). A 5-digit integral number is uniquely assigned to each customer
- country: Country name (Nominal). The name of the country where a customer resides

There are some preprocess steps applied to the dataset. The outlier values are eliminated.

$$\Delta = 1.5 * (Q_3 - Q_1)) \tag{1}$$

Q_1, Q_3: The average values of each partition when divided by the mean value for an attribute.

$$outlier = data[i] < Q_1 - \Delta \cup data[i] > Q_3 - \Delta \tag{2}$$

Limits can be determined with the value Δ, which represents the weighted difference between quarters. The outlier values are then calculated as follows.

After the outliers are removed, min-max normalization is applied to the features in the dataset. The number of rows after the preprocessing step is 1012837.

III. Method

In this section, we give details about the word embedding methods first. After that, we explain the customer embeddings are created. Finally, we mentioned how the clustering algorithm is built.

[1]https://www.kaggle.com/code/ilkeryildiz/rfm-clustering-of-customers-using-k-means/data

A. Word Embedding

Word embedding is a powerful mechanism. Its basic usage is to represent words in a text corpus. A neural network is run behind and a word space is created as a result based on the similarity of words created with numerous different dimensions [13], [14]. This word space can be used to find the most similar words for a given one.

Even if word embedding is an NLP topic, it has been applied in various areas. Instead of words, an item, a customer, an event, or more generally an entity can be embedded after being provided in a valid format [15]. In a case when a group of the product is fed into a word-embedding mechanism, it may recommend similar items, which is also quite popular in recent years [12].

There are also some other methods for representing words as numbers such as one-hot encoding and term frequency-inverse document frequency (TF-IDF) [16]. However, these methods can not capture the semantic relation between words. Besides the sparsity of the word matrixes are quite high.

B. Customer Embedding

In the study, customers are embedded using their generated customer ids instead of words using the same logic (Figure 1). Customers, who have the same items at a similar time, are assumed to be similar. The customers are grouped by the item and customer id lists are created. The customer lists are considered sentences whereas customer ids are considered words like in a sentence. The idea behind this is that, if the customers are near these lists, it means that they are similar. It is an analogous logic when the words are near to each other many times, they are semantically close. Apart from this, another embedding logic is executed for the customer. For that embedding, customers are grouped by their countries. By this embedding, it is going to answer the question of whether the customers are similar when they buy products at similar times in the same country.

The embedding vectors of customers are added as features to the dataset. It might be considered the same as the one-hot encoding of categories. However, the difference is that a neural network is run in the background which can detect and reason patterns. Besides, the time of the transaction is another factor for embedding. The order of customers in a list (corresponding to a sentence in the NLP sense) is formed based on that. That differentiates between using an embedding model and one-hot encoding even if they derive from the same feature.

For embedding, one of the best embedding models [17], the FastText model is created and a customer space is established. The embedding model is built using the parameters in Table I.

C. Customer Clustering Model

In previous sections, the preprocessing steps and feature engineering are explained. After creating the features, the clustering model is built. The clustering algorithm applied is K-means [18] for the study.

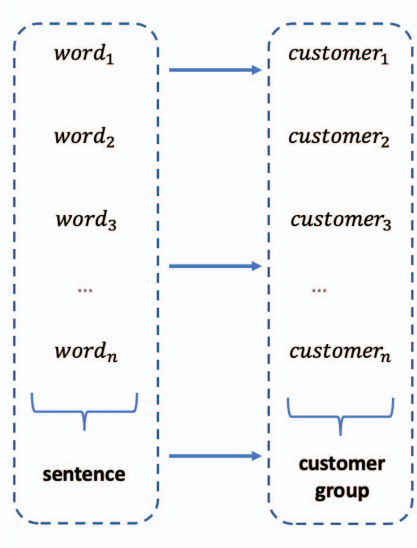

Fig. 1: Sentence-Customer group

TABLE I: EMBEDDING MODEL PARAMETERS

Model	Parameters
embedding_size	20
window_size	40
min_word	5
iter	100

The algorithm is called K-means due to the fact that the letter k represents the number of clusters chosen. An observation is assigned to a particular cluster for which its distance to the cluster means is the smallest. The principal function of the algorithm involves finding the k-means. First, an initial set of means is defined and then subsequent classification is based on their distances to the centers [19]. Next, the clusters' mean is computed again and then reclassification is done based on the new set of means. This is repeated until cluster means don't change much between successive iterations [20]. Finally, the means of the clusters are once again calculated and then all the cases are assigned to the permanent clusters.

The number of clusters is decided by using the elbow method [21] in Figure 2.

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)} \tag{3}$$

s_i: Silhouette coefficient
a_i: Average distance to objects in current cluster
b_i: Objects in the cluster closest to the current cluster. average distance

As it can be seen, s_i varies between -1 and +1. Ideally, the expected a_i would be close to 0, and in this case It means that a_i approaches 1. Therefore, it can be thought that the closer s_i is to +1, the better the clustering.

TABLE II: SILHOUETTE SCORES

Clusters	Silhouette Score (1)	Silhouette Score (2)	Silhouette Score (3)
2	0.69185	0.69784	0.69002
3	0.64355	**0.66847**	0.64176
4	0.60675	0.60843	0.60603
5	0.61608	0.61809	0.615438
6	0.54737	0.54789	0.54472
7	0.55326	0.55384	0.54886
8	0.54659	0.54702	0.54550
9	0.49207	0.49247	0.49011

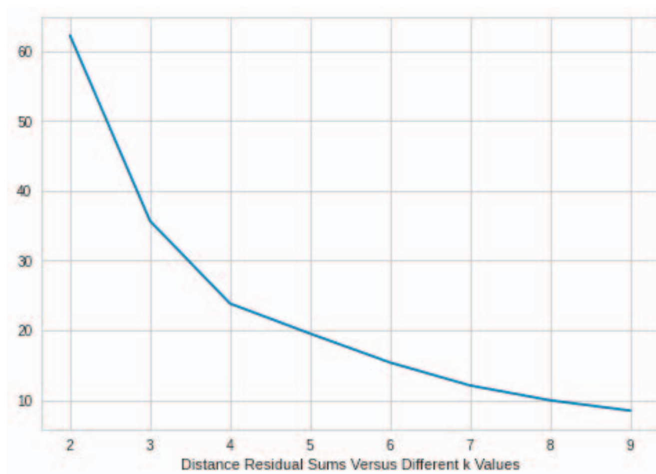

Fig. 2: Elbow method for Optimum number of clusters

IV. Experiments

For the study, the clustering model is built for a different set of features to be able to see their effects. There are two sets of features. The main features are recency, frequency and monetary. Additional features are appended to see their influence on the result. The first feature set appended is the embedding vectors of the customer having the same items at similar times. Besides, the second set consists of the embedding vectors that the customers in the same country and buy products at similar times.

In the study, the models are trained in Python programming language using sklearn and gensim libraries.

V. Observations

In the experiment sections, clustering models with different features are compared. The results can be observed in Table II. Silhouette Score (1), Silhouette Score (2) and Silhouette Score

(3) correspond to the clustering models with rfm features, rfm & customer-product embedding features and rfm & customer-country embedding features. In the table, it is seen that customer embedding features have an impact on the model's success. There are two different embedding vectors are created based on the products and countries. Customer-country embedding has slightly worse results than the rfm features. Moreover, customer-product embedding features have improved the result marginally. Customer-country embedding is too general and it is thought that it does not represent the customers well enough. However, customer-product embedding has more specific patterns and can represent customers better. Therefore, customer-product embedding has a positive effect on customer clustering in the study. In Figure 3, the output of three clusters is visualized where customer embedding vectors are added as features. The number of clusters is as follows; cluster 1: 3535, cluster 2: 1532, cluster 3: 796.

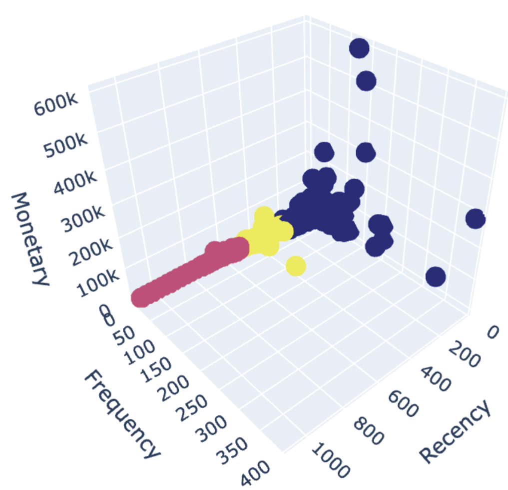

Fig. 3: Customer-Product embedding outputs

VI. Conculution

In today's world where the competition is quite high, customer clustering is one of the main concepts in almost any field in the industry. Companies use customer clustering to be able to group their customers and analyze them better. In this way, they can make their predictions more robust. There are numerous different methods to cluster customers. In this study, open and online retail data is used. RFM features are created and customers are grouped based on those features. The clustering method applied is k-means.

As being a different approach, customers are embedded based on the product they buy and the countries they are from. Customer ids are embedded using FastText which is an NLP method in traditional use. After creating a customer space, customers can be represented as n-dimensional vectors. These vectors are added as features and their effect is examined on the clustering. The results show that the representations of customers have a positive impact on customer clustering. In the future, for customer embedding, transformers are planned to use and compare the existing methods.

References

[1] D. K. Dey, S. Garla, and G. Chakraborty, "Comparison of probabilistic-d and k-means clustering in segment profiles for b2b markets," 2011.

[2] P. Agarwal, M. Alam, and R. Biswas, "Issues,challenges and tools of clustering algorithms," *ArXiv*, vol. abs/1110.2610, 2011.

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

[4] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," *Univ. Minnesota Supercomp. Inst. Res. Rep.*, vol. 213, 01 2003.

[5] J.-J. Jonker, N. Piersma, and D. Van den Poel, "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability," *Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report*, 01 2002.

[6] F. A. Guilherme F.de Arruda, Luciano da Fontoura Costa, "A complex networks approach for data clustering."

[7] 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.

[8] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, F. A. Rodrigues, and L. d. F. Costa, "Clustering algorithms: A comparative approach," 2016.

[9] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," 2015. [Online]. Available: https://arxiv.org/abs/1511.06335

[10] S. Jagabathula and Subramanian, *Lakshminarayanan and Venkataraman*. Ashwin, A Model-Based Embedding Technique for Segmenting Customers (January 20, 2017.

[11] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "Starspace: Embed all the things!" 2017. [Online]. Available: https://arxiv.org/abs/1709.03856

[12] M. G. Ozsoy, "From word embeddings to item recommendation," 2016. [Online]. Available: https://arxiv.org/abs/1601.01356

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *JOURNAL OF MACHINE LEARNING RESEARCH*, vol. 3, pp. 1137–1155, 2003.

[14] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[15] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "Starspace: Embed all the things!" 2017. [Online]. Available: https://arxiv.org/abs/1709.03856

[16] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, 2018.

[17] M. Toshevska, F. Stojanovska, and J. Kalajdjieski, "Comparative analysis of word embeddings for capturing word similarities," in *6th International Conference on Natural Language Processing (NATP 2020)*. Aircc Publishing Corporation, apr 2020.

[18] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67.

[19] S. Dasgupta and Y. Freund, "Random projection trees for vector quantization," 2008. [Online]. Available: https://arxiv.org/abs/0805.1390

[20] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is np-hard," *Theoretical Computer Science*, vol. 442, pp. 13–21, 2012, special Issue on the Workshop on Algorithms and Computation (WALCOM 2009). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304397510003269

[21] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 533–538.