Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Performance analysis of soft clustering approaches for e-commerce customer segmentation

Elektronik ticaret müşteri segmentasyonu yumuşak kümeleme yaklaşımlarının performans analizi

Hüseyin Fidan[1,*]

[1]*Burdur Mehmet Akif Ersoy Üniversitesi, Endüstri Mühendisliği Bölümü,15030, Burdur Türkiye*

Abstract

Soft clustering, which is one of the most important research areas of unsupervised machine learning, is preferred because it provides more suitable results for real-life applications. The fundamental idea of concept is that an item can belong to multiple clusters. Fuzzy-based approaches are generally applied for analysis. Especially due to the insufficient data, soft clustering algorithms have poor performance. In this study, a new soft clustering method based on Grey System Theory has been developed. The method and other soft clustering approaches used in the literature were applied for customer segmentation to a dataset containing customer transaction data of an e-commerce company. According to the results, it was determined that the developed method has given more successful results in small datasets compared to other soft clustering algorithms.

Keywords: Soft clustering, Overlapping clustering, Grey system theory, Soft grey relational clustering, Customer segmentation

Öz

Denetimsiz makine öğrenmesinin en önemli araştırma alanlarından biri olan yumuşak kümeleme, gerçek yaşam uygulamaları için daha uygun sonuçlar sağlaması nedeniyle tercih edilmektedir. Kavramın temel fikri, bir öğenin birden fazla kümeye ait olabilmesidir. Analiz için genellikle bulanık tabanlı yaklaşımlar uygulanmaktadır. Özellikle veri yetersizliği nedeniyle yumuşak kümeleme algoritmaları düşük performans göstermektedir. Bu çalışmada, Gri Sistem Teorisine dayalı yeni bir yumuşak kümeleme yöntemi geliştirilmiştir. Yöntem ve literatürde kullanılan diğer yumuşak kümeleme yaklaşımları, bir e-ticaret şirketine ait müşteri işlem verilerini içeren veri setine müşteri segmentasyonu için uygulanmıştır. Sonuçlara göre geliştirilen yöntemin diğer yumuşak kümeleme algoritmalarına kıyasla küçük veri setlerinde başarılı sonuçlar verdiği belirlenmiştir.

Anahtar kelimeler: Yumuşak kümeleme, Örtüşen kümeleme, Gri sistem teorisi, Yumuşak gri ilişkisel kümeleme, Müşteri segmentasyonu

1 Introduction

One of the main goals of companies for profit maximization is to increase their customer value [1]. Customer Relationship Management (CRM), that is one of the strategic research areas, aims to gain new customers, increase purchasing and customer loyalty. Especially with the increase in e-commerce transactions all over the world, this aim has become a critical issue for companies. Accessing real transactional data with e-commerce has enabled more realistic analyses and thus customer profiles have identified more accurately [2]. Customer segmentation that is defined as dividing customers into groups according to their similar characteristics [3], is seen as one of the most common tools for understanding customer structure [4]. In order to better understand and interpret the customer structure, customers are divided into categories according to their characteristics and transactions [5]. Customer segmentation has been applied in various research areas such as finance [6], e-commerce [7], industry [8], and information technologies [9]. In the studies, RFM (Recency-Frequency-Monetary) is commonly used for grouping customers [10, 11, 12] and customer scores are calculated with transactional data, and customers are grouped according to the scores [13]. However, this method has some drawbacks such as determination of threshold points according to scores [14], causing unsuccessful forecasting results [15] and differences in RFM values according to sectors [13]. Although different models have been developed to expand RFM capabilities such as LRFM [16] and LRFMP [17], the process in question is basically clustering, which is an unsupervised machine learning technique [4]. In this context, clustering is seen as the most appropriate method to be used for customer segmentation [5, 18, 19].

Clustering is an unsupervised machine learning technique and defined as identifying the most proper group for data items [20]. The concept is based on the idea that an item with more similar features will definitely be found in a dataset [21]. In other words, clustering is the grouping of items in a dataset according to their similar properties. There are two approaches depending on the number of clusters an item can be in. The basic clustering idea, called

* Sorumlu yazar / Corresponding author, e-posta / e-mail: hfidan@mehmetakif.edu.tr (H. Fidan)
Geliş / Received: 14.12.2024 Kabul / Accepted: 18.12.2025 Yayımlanma / Published: 15.01.2026
doi: 10.28948/ngumuh.1601524

hard clustering, is that an item can belong to only one cluster [4, 20, 21]. The second one is soft clustering that each item can be located in more than one cluster [22]. In these studies, various clustering algorithms are used such as K-Means [23, 24], K-Medoids [25], EM (Expectation-Maximization) [26], SOM (Self Organizing Map) [16, 27], FCM (Fuzzy C-Means) [28, 29]. Hard clustering methods are generally preferred in customer segmentation, and analyses are carried out with the K-Means algorithm [1]. However, determining only one cluster for customers will not be sufficient to characterize customer profiles. For example, labelling a movie with just "adventure" will not give the audience enough information about the movie. The film can be labelled as both "adventure" and "romantic". The same is true for data that is frequently encountered in social life, such as news, articles, photographs, videos, and books. Similarly in customer segmentation, hard clustering will determine a customer only in one group such as "low login". But, the consumer may be one who logs into the system less but spends a lot. In this case, an exact inference cannot be made about the consumer. However, with soft clustering, a customer may be included in the cluster of "low login", as well as in cluster of "high spenders". Soft clustering will give more detailed information about the customers and will have an important role in customer segmentation. It is clear that further increasing of soft clustering approaches will make positive contributions to customer segmentation studies. The reasons why soft clustering is rarely used are that soft approaches have some deficiencies such as complicated implementations [30], high sensitivity for outliers [31], long processing time [32], difficulty in finding optimal cluster centers [33, 34], low performance [35] and challenges in interpreting the results [36]. In this context, soft clustering is recognized as a research area where new models and algorithms need to be developed [28, 37]. On the other hand, in recent years, it has been suggested that Grey System Theory (GST) can provide successful results in the case of insufficient data [38, 39]. In several studies, GST-based Grey Relational Clustering (GRC) method was developed for hard clustering, and successful results were obtained [33, 36, 39, 40]. But, GST, which can provide a solution to the problems of insufficient data and small datasets, is not included in soft clustering studies. The development of GST-based soft clustering algorithm will make positive contributions to soft clustering literature in analysis of small datasets and insufficient data. The main purpose of the study is to develop a new GST based soft clustering algorithm and determine its performance for customer segmentation of small datasets. Thus, the GST based algorithm to be developed will provide an alternative solution to the problems experienced due to insufficient data in the analysis of small data sets with FCM and EM based algorithms.

In this study, a new GST-based soft clustering algorithm named Soft-G was developed and applied to customer transaction data obtained from an e-commerce company. In order to observe the performance of the algorithms in terms of the amount of data, six datasets consisting of 10, 50, 100, 500, 1000, and 5130 customers were created. Besides the Soft-G, the most preferred soft clustering algorithms FCM, Type2-FCM (FCM-T2), and EM for customer segmentation were used in the analysis. Clustering performances were compared with the internal validation measures Partition Coefficients (PC), Partition Entropy (PE), Fuzzy Silhouette Coefficients (FSC) and Xie-Beni Index (XB).

The rest of the article is organized as follows. In the second section, the clustering problem, soft clustering algorithms, and related literature are introduced. Section 3 includes the explanations of the datasets, research model, theory of Grey Relational Analysis and performance metrics. The experimental results and comparisons are given in Section 4. The last section contains conclusions and suggestions.

2 Background

2.1 Clustering problem

Clustering, one of the unsupervised machine learning methods, is the grouping of items in a dataset according to their feature-based similarities. The concept is based on the assumption that there may be items with more similar features than others in a dataset [21]. So, the clustering is defined as identifying the most proper groups for data items [20]. Distance criteria such as Cosine, Manhattan, Gini, and Euclidean are used to determine the similarity levels between items. Items with low distance criteria are those that are close to each other in terms of features and can be included in the same cluster. In the clustering literature, two types of methods are used, as hard and soft, depending on an item can belong to one or more clusters [22, 28, 41].

2.2 Soft clustering

More than one term can be used to describe objects or events in social life. A similar situation can be thought for clustering analyses. Determining more than one label to express the attributes of an item means that the items can be included in more than one cluster. This approach, called soft clustering, is seen as a more realistic for clustering applications [28, 41, 42]. More exhaustive and more exclusive clusters can be realized for representing data by using soft approaches.

The idea that an item may be in more than one cluster was first expressed as a mathematical model by Jardine and Sibson [43]. Considering more than one group for items as fuzzy, Ruspini [44-46] carried out the first studies that included fuzzy theory in grouping items. The first clustering method based on fuzzy theory was developed on the basis of the K-means algorithm [47, 48]. The clustering approach became widespread in the field of unsupervised machine learning with the name FCM, after Bezdek's work [49] on pattern recognition on uncertain membership. The studies of Ruspini [44], Dunn [47] and Bezdek [49] constitute a starting point for soft clustering methods. In this context, most of the methods developed on the subject in the literature use the fuzzy theory and are based on the FCM algorithm.

2.2.1 Fuzzy C-Means clustering

FCM is the most widely applied soft clustering algorithm. As seen in Figure 1, each item has relational values called membership functions with all cluster centers. Membership values are between [0,1] and their sum equal to 1.

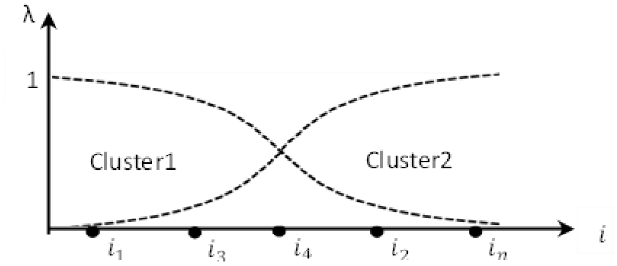

Figure 1. Fuzzy clustering

The distance of the items is calculated by Euclidean metric. It is aimed to minimize the weighted sums of the membership values calculated by the distances. It is seen in Equation 1 that N items and K clusters are created using fuzzy approach.

$$\min J = \sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_{i,k}^{m} |x_i - \mu_k|^2 \tag{1}$$

where $\lambda_{i,k}$ is membership value of item i for cluster k and $\lambda_{i,k} \in [0,1]$. m is fuzzifier parameter and $m \in (1, \infty)$. As the value of m increases, the fuzzyness increases, as m gets closer to 1, the FCM acts as K-means [50]. In the literature, it is usually taken as $m = 2$ for better results [28, 49, 50]. x_i is the i^{th} item, μ_k is center of cluster k and $|x_i - \mu_k|$ represents the Euclidean distance between item i and center of cluster k. $\lambda_{i,k}$ is shown in Equation 2.

$$\lambda_{i,k} = \frac{1}{\sum_{j=1}^{K} \left(\frac{d(x_i,\mu_k)}{d(x_j,\mu_k)}\right)^{\frac{2}{m-1}}} \tag{2}$$

$\lambda_{i,k} = 1$ means that i is in cluster k and cannot be in any other one. $\lambda_{i,k} = 0$ means that item i is not in k. $\lambda_{i,k}$ is $(0 < \lambda_{i,k} < 1)$ and represents the closeness of the item i to cluster k. Considering that the sum of all membership values of an item should be equal to 1 [49], the membership value of an item being greater than 0 means that the item belongs to at least two clusters.

Although many fuzzy-based clustering methods have been developed in the literature, almost all of them are based on FCM [28]. FCM algorithm, which has attracted a lot of attention in the clustering literature since its first appearance, is widely used for datasets having uncertainty. It is a powerful algorithm that offers flexible usage, but it has some shortcomings in practice. It is stated that there are problems such as the necessity of determining parameter values before the analysis [28], the analysis results being affected by the parameter values, and unsuccessful results with a limited amount of data [51]. Especially in the case of insufficient data, FCM fails in clustering analyses due to the inability to produce sufficient membership functions [36, 51].

2.2.2 Type2 Fuzzy C-Means

FCM-T2, which is the extended form of the FCM algorithm, is a fuzzy-based soft clustering algorithm developed by Rhee and Hang [52]. In the traditional FCM algorithm called Type1, each item is assigned a membership function according to the distance values from the cluster centers. Items with high membership values have a dominant effect in determining the center. This situation creates problems in the selection of cluster centers in noisy datasets and leads to a decrease in clustering performance. It is assumed that high membership values have low uncertainty and low membership values have large uncertainty [52]. So, the problem will be eliminated by applying fuzziness again to Type1 FCM membership values. Thus, compared to FCM, cluster centers will have a better representation in FCM-T2 and more successful results will be obtained in the case of noisy data [29]. The expression of Type2 membership values (a_{ik}) as Type1 membership values (λ_{ik}) is given in Equation 3.

$$a_{ik} = \lambda_{ik} - \frac{1-\lambda_{ik}}{2} \tag{3}$$

$$v_k = \frac{\sum_{i=1}^{n} a_{ik}^{m} x_i}{\sum_{i=1}^{n} a_{ik}^{m}} \tag{4}$$

Determining cluster centers by using the Type2 membership values is shown in Equation 4. v_k is the center of cluster k.

2.2.3 Expectation maximization

It is a Gaussian Mixture Model (GMM)-based clustering algorithm introduced by Dempster, Laird and Rubin [53]. The EM algorithm developed to determine the maximum similarity in case of missing data is a probability-based soft clustering method [53]. The algorithm assumes that data points are generated from a mixture of probability distributions and it has two stages [54]. In the first step of the algorithm called Expectation, missing data is estimated. In the second stage called Maximization, the parameters are estimated according to the generated data. These steps are repeated until the estimated parameters approach the previous parameter values [53].

$$p(x) = \sum_{k=1}^{K} \pi_k \, N(x_n; \mu_k, \sigma_k^2, b_k) \tag{5}$$

For a single variable x, Gaussian Mixture distribution shown in Equation 5, can be expressed as a linear stacked form of Gaussian density values K. The component mixture N that calculated according to each k value in the equation, has the values mean μ_k, variance σ_k^2 and bias b_k. π_k represents the mixing coefficients, $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ [55].

One of the most successful methods of estimating parameters is Maximum Likelihood Estimation (MLE). MLE and Bayesian theorem are used for parameter estimation of the EM algorithm [55]. According to Bayes' theorem, the posterior probability of the distribution in Equation 5 can be calculated as in Equation 6.

For the Expectation step, the posterior probabilities are calculated by Equation 7. For maximizing likelihood (i.e., for the Maximization step), the derivatives of Equation 7 with respect to π , μ, σ^2 and b and b are taken and equated to 0. Thus, according to the estimated data in the Expectation step, parameter values are calculated in the Maximization step. The processes are repeated until these parameter values are closest to each other [56].

$$p(k|x) = \frac{\pi_k \, N(x_n;\mu_k,\sigma_k^2,b_k)}{\sum_{k=1}^{K} \pi_k \, N(x_n;\mu_k,\sigma_k^2,b_k)} \tag{6}$$

$$\ln p(x|\pi,\mu,\sigma^2) = \sum_{n=1}^{N} \ln(\sum_{k=1}^{K} \pi_k \, N(x_n|\mu_k,\sigma_k^2,b_k)) \tag{7}$$

2.3 Related works on customer segmentation

Customer segmentation is a concept used in the field of marketing. It was first introduced by Smith [57] in his 1956 paper titled "Product differentiation and market segmentation as an alternative marketing strategy". The paper proposed the market segmentation approach, which involves dividing a heterogeneous market into more homogeneous segments, along with a product differentiation strategy. E-commerce systems enable interactive relationships with customers by using web infrastructure and personalization technologies [58]. In the book named "Electronic Commerce: A Managerial Perspective," Turban et al. [59] also examined the structure and purchasing behaviour of customers. Emphasizing the importance of customer segmentation within the scope of improving customer relations, the authors stated that it would be appropriate to group e-commerce customers into 3 groups: "impulsive deals," "patient deals," and "analytic deals". Ozer [60] divided music service customers into 6 groups by applying clustering analysis to the survey data he collected from 160 people. Bhatnagar and Ghose [61], who divided online consumers into groups according to their purchasing behaviour, conducted a survey on 1330 people. The researchers determined that the customers would be grouped into 3 groups according to the Bayesian Information Criteria (BIC) value and identified customer groups using the latent class model they developed. Rho et al. [62] argued that using e-commerce data instead of surveys in customer segmentation would yield more realistic results. The analysis, which they conducted by using K-means algorithm, resulted in the formation of 3 clusters based on demographic characteristics and transaction data. Tsai and Chiu [13] remarked that traditional market segmentation applications are based on customer demographic characteristics and lifestyles. So, this approach has some problems and the segmentations are unreliable. The researchers compared the algorithm's performances and stated that the developed algorithm's performance was higher and that better segmentations could be achieved using fuzzy algorithms. Chan [27], who divided online auction customers into 3 groups, "impulsive deals," "patient deals," and "analytic deals," using the approach, applied the neural network-based Self-Organizing Maps (SOM) algorithm for clustering. In a study in which the clustering of 3419 e-commerce customers according to their purchasing behaviour was carried out with K-Means and SOM algorithms, it was determined that the K-means algorithm has more stability [63]. Performing cluster analysis using the same algorithms, Hong and Kim [64] suggest that although there are basically 3 groups of customers: high propensity to buy (segment A), distrustful customer (segment B), and customers who rarely visit (segment C), customers can also be among these segments. According to this approach, customers with both A and B segment features will be in the AB segment. Thus, segmentation was carried out with 5 clusters as A, AB, B, BC and C. Kansal et al. [65] created 5 clusters using K-means, Hierarchical and Mean-Shift algorithms, compared their performance with the Silhouette index and determined that the performance of K-means and Hierarchical clustering algorithms was high. In a recent study, Koul and Philip [66] determined that the K-means algorithm gave more successful results than the hierarchical and density-based clustering algorithms, according to the results of the analysis performed using hard clustering algorithms.

Representing customers who make transactions in e-commerce with only one group is not a realistic approach. In this context, soft clustering based on the idea that a customer can be in more than one cluster will yield more realistic results. In fact, this idea was also emphasized by researchers in the early years of e-commerce. Basak and Kumar [67] suggested that a precise model of a customer cannot be created exactly, but different group membership values can be determined using fuzzy approaches with historical transaction data. Their study, which was carried out only in a theoretical framework, suggests that customer segmentation should be analysed with soft clustering methods. In another study, customer segmentation for an online music service was carried out with fuzzy clustering [60]. Sicilia and Garcia [58] stated that the definition of a valued customer is a fuzzy concept, and they proposed the fuzzy theory for segmentation of online customers according to their calculated values and developed a mathematical model. Shin and Sohn [68] compared K-Means, SOM and FCM algorithms using stock trading customers data, found that FCM gave the most robust results. Kaur et al. [69] determined that Type2 fuzzy algorithms give healthier results in clustering noisy datasets due to their approach to detecting cluster centers. It was also determined by Gosain and Dahia [29] that FCM-T2 gave more successful results compared to FCM. Wu and Chou [11] developed a new soft clustering method that determines the differences with the cut-off value by creating a score vector with customer information. According to the method, if the cut-off value is high, the probability of the customer to be in the multi-group increases. Although it is stated that the method gives more successful results than the hard approaches, the main deficiencies of the study are that there is no model for determining the cut-off value and that the customer differences are determined according to the highest customer score. Hiziroglu [5] made a systematic review of 40 studies including soft approaches on customer segmentation, determined that 20 of these studies were based on SOM and 6 of them were based on FCM algorithm. It is surprising that the hard clustering algorithm SOM is used much more instead of the FCM algorithm for the soft approach in customer segmentation. There are other researchers who make similar findings. Khalili-Damghani, Abdi and Abolmakarem [70] argued that K-Means is mostly used in customer segmentation, performed by developing a hybrid model that includes K-Means, decision tree and rule mining. In another study, a hybrid model was developed with K-Means-based k-overlapping and k-harmonic algorithms, and also claimed that the outlier problem in soft clustering is a serious problem and they achieved successful results with the hybrid approach [30]. According to Yoseph, Malim and Almalaily [71], soft approaches were given little place in customer segmentation, compared the performances of EM and FCM algorithms. The results of the study revealed that the EM algorithm was more successful. Ferraro and Giordani [26] found that EM produced similar clustering results with fuzzy approaches in clustering.

In recent years, the literature has placed increasing emphasis on fuzzy clustering, particularly highlighting the limitations of FCM. To address these shortcomings, researchers have explored hybrid approaches, aiming to enhance the performance of FCM. Although these developments undoubtedly contribute to the literature, they also prove that FCM-based clustering approaches are inadequate in some areas. Drawing attention to the low performance of FCM in noisy datasets, Wu and Zhang [72] developed the adaptive weighted fuzzy clustering method. Yu et al. [73] claimed that FCM did not give meaningful results in fuzzy boundary data. The authors used FCM together with Rough-set theory to remove the uncertainty in the boundary values and also stated that the performance increased. Yue et al. [74] emphasized that the high level of uncertainty in real transaction data reduces the clustering performance and suggested the harmonic difference method for the solution. Considering the performance of fuzzy-based clustering algorithms decreases in case of insufficient data, Alekhya and Sasikumar [75] used the FCM and Support Vector Machine algorithm as a hybrid. Researchers found an increase in FCM performance, and recommended that the method could even be used in the clustering of niche diseases. Oskouei et al. [76] remarked that the initial cluster center selection in the FCM algorithm is a factor affecting the success of the FCM, and recommended using the FCM and Feature-Weight method together. In this way, they calculated 3% increase in FCM performance.

Recently, methods such as Deep Learning (DL), Genetic Algorithm (GA) and Artificial Neural Network (ANN) have been applied in soft clustering studies [10, 77-79]. However, these methods are costly applications because they require long computational time, large amount of data, high performance peripherals and hard implementations [77]. Shimura, Li, and Fukumoto [79] emphasized that there is a lack of consideration of clustering infrastructures in DL approach analyses. According to Menzies et al. [80], emphasizing that DL methods are successful in analyses with a large amount of data, DL gives unsuccessful results when the amount of data decreases. According to the research, an improved K-means algorithm will be less costly and more successful than DL methods in clustering analysis. Similarly, methods based on GA, ANN and association rules require much more data, long computational time, capacity hardware and analysis results have low stability [10]. For this reason, the decrease in the amount of data for these approaches leads to their failure.

When the studies on customer segmentation are examined, it is seen that the studies are generally carried out with hard clustering approaches, and with the spread of e-commerce, survey methods are abandoned and analyses are carried out with e-commerce transaction data. In addition, it is seen that there is no clarity in the number of customer groups in the studies, and accordingly the number of clusters varies between 2 and 6. The most important shortcoming about the methods used in the studies is that soft approaches are not sufficiently included in applications. It is seen that studies using soft approaches are limited to K-Means and Fuzzy based algorithms and new methods with high performance have not been developed sufficiently. Especially the fact that fuzzy-based clustering algorithms present different accuracy values due to data-based problems negatively affects the selection of appropriate algorithms. Peters et al. [28] argue that there is a need for basic approaches that deal with uncertainty as a whole (holistically) rather than approaches involving hybrid and complex algorithms in soft clustering applications. In this context, the use of basic soft approaches in the development of low-cost, high-performance and easy-to-implement soft clustering algorithms will make significant contributions to the literature.

In some studies, it is seen that GST-based approaches are applied in machine learning in case of uncertainty caused by insufficient data. It is emphasized that successful results were obtained in hard clustering by using GRC [33, 36, 40]. It is seen that recent GST-based studies are generally carried out in areas such as optimization [81, 82], multi-criteria decision making [83, 84], image segmentation [85, 86], clustering [36, 87]. Although there are a few studies in the field of clustering, no GST-based study on soft clustering has been found. The GRC method, which provides clustering analysis using similarity values between items, also has the ability to perform analyses with a small number of items. In this context, GST-based approach will be a suitable option for soft clustering of customer segmentation.

3 Material and method

3.1 Dataset

The dataset to be used for the analysis was obtained from an actively operating e-commerce company. The company, which has been operating since 2020, has 38426 registered customers as of October 2024. The number of customers who have purchased from the company at least once is 5130. The dataset contains 7 variables that are given in Table 1.

Table 1. Dataset variables

Variables	Explanations
Var1	The count of all purchases
Var2	How many days ago the last purchase was
Var3	Number of years the customer was registered
Var4	The count of customer's logins
Var5	Number of customer's product returns
Var6	Number of comments (Positive-Negative)
Var7	Number of deletions from cart

In order to evaluate the performance of the algorithms according to the amount of data, 6 datasets with customer numbers of 10, 50, 100, 500, 1000 and 5130 were created. Each item in the dataset represents a customer. Explanations about the created datasets are presented in Table 2.

Table 2. Datasets

Dataset	Item count	Explanations
Ds1	10	Top 10 items of the dataset
Ds2	50	Top 50 items of the dataset
Ds3	100	Top 100 items of the dataset
Ds4	500	Top 500 items of the dataset
Ds5	1000	Top 1000 items of the dataset
Ds6	5130	Entire dataset

3.2 Grey relational analysis

Grey Relational Analysis (GRA) that is a decision-making method based on Grey System Theory (GST) determines the similar items with a reference [38, 39]. The similarity values which vary between 0 and 1 are calculated by item criteria. The item with the highest value indicates the highest similarity with the reference. So, items can be clustered using relational similarity degrees [39].

First step of GRA analysis is the building of decision matrix. Decision matrix X given in Equation 8 consist of alternatives (or items) in the dataset ($a = 1,2,3,\dots m$) and the criteria (or features) ($c = 1,2,3,\dots n$) of each alternative. For example X_{2n} represents the n^{th} criteria of the second item.

$$X = \begin{matrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{matrix} \tag{8}$$

Normalization is applied to minimize the differences. In the study, Min-Max normalization was preferred to avoid negatively affecting the data distribution. A reference series X_{0c} that defines the decision criteria is added to X as an item (alternative). The reference series is shown in Equation 9 and it has all criteria of X. In clustering, the reference series is not added, and normalization is applied by considering the first item as the reference series.

$$X_{0c} = X_{01}, X_{02}, \dots X_{0n} \tag{9}$$

If the maximum value of the criteria contributes the decision positively, the utility based Min-Max normalization (Equation 10) is used. Thus, the normalized matrix given in Equation 11 is obtained.

$$X^*_{ac} = \frac{X_{ac} - minX_{ac}}{\mathrm{max}X\,ac - minX_{ac}} \tag{10}$$

$$X^* = \begin{matrix} X^*_{01} & X^*_{02} & \cdots & X^*_{0n} \\ X^*_{11} & X^*_{12} & \cdots & X^*_{1n} \\ X^*_{21} & X^*_{22} & \cdots & X^*_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ X^*_{m1} & X^*_{m2} & \cdots & X^*_{mn} \end{matrix} \tag{11}$$

The normalization process is not mandatory for GST-based analyses. However, in some studies, it is emphasized that the application of normalization eliminates the discrete values and thus increases the analysis performance [36, 40].

Absolute differences are applied to find the difference values between each criterion and the items. Equation 12 is applied to X^* to find absolute differences and the matrix of absolute differences shown in Equation 13 is obtained.

$$\Delta_{ac} = |X^*_{0c} - X^*_{ac}| \tag{12}$$

$$\Delta = \begin{matrix} \Delta_{01} & \Delta_{02} & \cdots & \Delta_{0n} \\ \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1n} \\ \Delta_{21} & \Delta_{22} & \cdots & \Delta_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_{m1} & \Delta_{m2} & \cdots & \Delta_{mn} \end{matrix} \tag{13}$$

As an example, to obtain the Δ_{21} value, the difference between the criterion value of the reference series and the value of the 2nd alternative's 1st criterion in the normalization matrix is calculated ($\Delta_{21} = |X^*_{01} - X^*_{21}|$). Δ_{max} and Δ_{min} values are determined in Δ. By using these values, the matrix of grey relational coefficients (γ) is built by Equation 14. γ is shown in Equation 15.

$$\gamma_{ac} = \frac{\Delta_{min} + \rho\Delta_{max}}{\Delta_{ac} + \rho\Delta_{max}} \tag{14}$$

$$\gamma = \begin{matrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mn} \end{matrix} \tag{15}$$

The parameter ρ in Equation 14 is known as the distinguishing parameter, which can either collapse or expand the coefficient matrix γ, and ρ is within the range of [0,1]. Changing the value of ρ affects the values of the matrix but not the item-criteria relationship. In the literature, $\rho = 0.5$ is commonly used [88].

Equation 16 demonstrates that the average value of the criteria for each item in γ. δ_a represents the gray relational degrees of the item a. Equation 17 shows the series of gray relational degrees for all items (δ), which represents the value of relationship between the reference series and alternatives. $\delta_a = 1$ means the highest relational degree between a and the reference.

$$\delta_a = \sum_{c=1}^{n} \gamma_{ac} \tag{16}$$

$$\delta = \begin{matrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_m \end{matrix} \tag{17}$$

The relational degrees calculated with the first reference, as shown in Equation 17, are added to the G similarity matrix. The process of calculating the relational degrees is repeated with the second item as the reference. The relational degrees obtained with the second reference are also added to the G and the next item is selected as the reference. This process is repeated until all items in the dataset have been selected as the reference. As a result, the calculated relational degrees for all items are included in the G. The maximum value in the G matrix indicates the two items with the highest similarity (Equation 18). In this way, the items that will form the first cluster are determined.

$$cluster = \max G \tag{18}$$

The center of the first cluster is determined by taking the average values of the two items in the new set, based on their properties. Then, both of these items are removed from the dataset and their averages are added as a new item. A new decision matrix is created with this updated dataset, and the process is repeated to determine the next cluster. This process is iterated until the desired number of clusters is obtained in the dataset. After completing the entire process, the relational degree values for all sets are found for each item in the dataset, as shown in Algorithm 1 and Algorithm 2. The threshold value for cluster membership is considered to be 0.5 and above. Values close to 1 indicate a high degree of relational similarity [88].

GRA method has been applied in various fields and has recently been utilized in data mining and machine learning [33, 36, 40]. Grey Relational Clustering (GRC), which is based on GRA method, was introduced by Deng [38] and later improved for clustering by Jin [39]. In literature, GRC has made significant contributions to clustering applications by offering a new perspective to solve data sparsity-related problems [40]. Studies using the GRC method mainly focus on hard clustering, and a soft clustering method based on GST has not yet been found in the literature.

3.3 Research Model

The model developed for soft clustering based on GST is given in Figure 2. The research model consists of three sequential steps as preparing dataset, building decision matrix and soft clustering. Soft clustering stage contains the developed algorithm Soft-G that has two GST based processes. The first step of the Soft-G is the GRC analysis to find initial clusters, and the second one is the GRA analysis for finding multi-clusters.

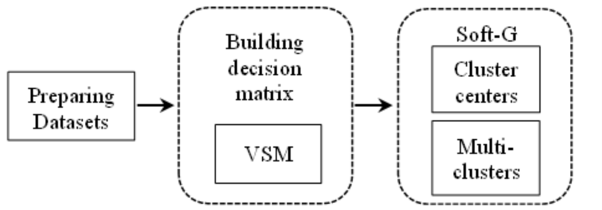

Figure 2. Research model

3.3.1 Preparing datasets

While creating the 6 datasets to be used in the analysis, only consumers who made a shopping transaction were selected. Since there was no missing value problem in the created datasets, data cleaning process was not applied. Since the analysis data were obtained from real transaction data, no intervention was made to the datasets with the idea of presenting the current situation exactly. In this context, the datasets were analyzed with real transaction values, and outliers were not excluded from the analysis.

3.3.2 Building decision matrix

Vector Space Model (VSM) is a common method used for text representation of documents. With VSM, vector representation of documents can be made according to the number of words in each document [33]. In this way, VSM, which is created in a matrix structure, provides convenience in calculating similarities between documents based on word frequencies [33, 36]. Since the clustering process in the Soft-G algorithm will be realized according to the similarity values between the transaction values of the each consumer, it would be appropriate to create the decision matrix based on VSM. The VSM model is created in a matrix format based on the transactional data values of each customer in the datasets. Based on the VSM, the decision matrix created for Ds1 is given in Table 3.

Table 3. Decision matrix for Ds1

Customers	Var 1	Var2	Var3	Var4	Var5	Var6	Var7
Cust1	9	35	3	134	0	0	7
Cust2	12	17	1	45	0	5	2
Cust3	14	16	2	63	0	2	12
Cust4	5	23	3	91	0	1	3
Cust5	7	8	4	218	0	0	4
Cust6	124	48	5	147	0	12	87
Cust7	82	25	3	105	0	1	7
Cust8	43	15	4	84	0	0	12
Cust9	44	14	2	71	0	0	31
Cust10	55	33	1	67	0	6	28

3.3.3 Finding cluster centers

A GST-based Soft-G algorithm has two stages. In the first stage, the centers of clusters were carried out by GRA.

Thus, the initial clusters are determined and items are placed only in one cluster. The method is given in Algorithm 1.

Algorithm 1 is applied separately for each dataset in the analysis. The algorithm detects the items with a high degree of similarity in the dataset and combines them by taking the average. These merging operations are repeated until the desired number of clusters is achieved. If the number of clusters is selected as 1, the algorithm continues to combine until all items are gathered under a single cluster. So, if the number of clusters is uncertain, cluster_count=1 is selected. Both the elements and centers of the clusters formed by the specified number of clusters can be determined by Algorithm 1.

Algorithm 1: Finding cluster centers

```
    Input: Dataset (in VSM form), number of cluster (cluster_counts)
    Output: Clusters

    Find_centers (dataset D, cluster_counts)
    do{
      set d= item count (D);
      set dm=build decision matrix(D)
      normalization (dm,utility based)
      set i=0 as reference // i=0 for first customer
     do {
          calculate the absolute differences (Δ_i)
          calculate the Grey relational coefficients (γ_i)
          calculate the Grey relational degrees (δ_i)
          add (δ_i , G) // G has relational degrees
          i++ //for next customer
        } while (i<d) // until the last customer
     find the max degree in G as j
     set item_1 and item_2
     set new_cus=mean(item_1,item_2)
     remove(item_1 and item_2)
     add(new_cus , D)
    }while (item count (D) > cluster_counts)
return (D)
```

3.3.4 Finding multi-clusters

After the cluster centers and the items to be included in them are determined, the relationship levels of each item with the cluster centers are found. The second stage of Soft-G is the finding multi-clusters by using Algorithm 2.

The relational similarity degrees (S) of the customers in the dataset (D) with each cluster produced in Algorithm 1 are calculated by the GRA analysis given in Algorithm 2. S is a matrix with dimensions (number of customers x number of clusters) showing the cluster similarity of the items. High S value indicates high similarity.

Algorithm 2: Finding multi-cluster

```
Input: Dataset (D), clusters (C)
Output: Relational similarities of cluster items (S)

Find_multi (dataset D, centers C)
     set d=item count (D) // number of customers
     set i=0 as reference // i represents the customers
     do {
       set dm=build decision matrix(C , D[i])
       normalization (dm,utility based)
       calculate the absolute differences (Δ_i)
       calculate the Grey relational coefficients (γ_i)
       calculate the Grey relational degrees (δ_i)
       add (δ_i , S) // S is similarity matrix of all
                  customers with clusters
       i++
    }while (i < d) // until the last customer in D
return(S)
```

3.4 Validation Indexes

According to the dataset being analysed, two types of validation indices are generally used: external and internal. External validation is used when the groups of items are known; otherwise internal validation indices are used. Index values reveal the clustering performance, which means the success of clustering algorithms. Partition Coefficients (PC), Partition Entropy (PE), Fuzzy Silhouette Coefficients (FSC), and Xie-Beni (XB) are frequently used indices for internal validation in soft clustering [89, 90]. These indices have been developed based on fuzzy membership values. The sum of the membership values formed as the number of clusters of an item is equal to 1. However, in EM and Sof-G, the sum of the membership values of an item is not equal to 1.

PC, PE, FSC and XB indexes perform calculations using the values of a matrix formed by membership functions in the range of $[0-1]$. In order to determine the EM and Soft-G clustering performances with these indexes, it is necessary to convert the EM probability values and Soft-G relational degree values to $[0-1]$ range. In order to increase the comparison and interpretation possibilities of the analyses, the process of converting the values in a dataset to a desired range is a widely used method in statistics [91]. Since there is no change in the proportional values after this normalization process, the effect of each feature will not change [92]. According to Equation 19, normalization is achieved by dividing the membership value of each item by the sum of all cluster membership values. Thus, the sum of probability values (for EM) and relational degree values (for Soft-G), which show the relations of an item with other clusters, will add up to 1.

$$u_i = \frac{v_i}{\sum_{c=1}^{k} v_c} \tag{19}$$

In the formula given in Equation 19, k shows the number of clusters, u_i indicates the calculated membership value of the item for cluster i, v_i the relationship level of the item in cluster i. $\sum_{c=1}^{k} v_c$ shows the sum of the relationship values of the all cluster items. So it can be written as $\sum_{c=1}^{k} u_{ic} = 1$

3.4.1 Partition Coefficients (PC)

PC index is a criterion used to determine the similarity levels of items within the scope of fuzzy-sets [93]. As shown in Equation 20, the PC index value is obtained by dividing the sum of squares of the membership values of each item in each cluster by the total number of items [94].

$$PC = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^2 \quad (20)$$

where n represents the number of item and k represents the number of cluster. u_{ij} is membership function of cluster i of item j and PC value is $0 \leq PC \leq 1$. $max\ (PC)$ indicates the highest clustering performance and PC value is not higher than 1. $PC = 0$ means that all items have equal membership values, which indicates an unsuccessful clustering [90].

3.4.2 *Partition Entropy (PE)*

PE is another index for determining performances of soft clustering algorithms. It indicates the differences degree of the clusters [93] and is given in Equation 21.

$$PE = -\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} \log_2 u_{ij} \quad (21)$$

Lower values of PE indicate higher clustering performance, where $0 \leq PE \leq \log_2 k$. $PE = 0$ indicates that all items are in one cluster, while $PE = \log_2 k$ indicates that all cluster sizes are equal. However, PE and PC values alone are not sufficient to determine clustering performance. These values are evaluated together, and the performance of the cluster is determined [89].

3.4.3 *Fuzzy Silhouette Coefficients (FSC)*

FSC is an adapted version of the silhouette index for soft clustering and is also called the Fukuyama-Sugeno index [95]. It determines the degree of similarity of an item to its own cluster compared to other clusters by using membership values. FSC is calculated by using Equation 22.

$$FSC = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^m \left\|x_j - a_i\right\|^2 - \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^m \left\|a_i - \bar{a}\right\|^2 \quad (22)$$

where $\bar{a} = \sum_{i=1}^{k} \frac{a_i}{c}$. FSC takes into account both compactness and separation values in clusters and $-1 \leq FSC \leq +1$. High FSC values mean high clustering performance [90, 95].

3.4.4 *Xie-Beni index (XB)*

XB is another widely used criterion for internal validation of soft clustering. XB index, which can calculate both compactness and separation values like FSC, is as seen in Equation 23 [96].

$$XB = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^m \left\|x_j - a_i\right\|^2}{n\ min_{ij} \left\|a_i - a_j\right\|^2} \quad (23)$$

XB index takes values between 0 and infinity, with a lower value indicating a higher cluster performance [89]. In other words, the clustering is considered more successful when the XB index is closer to 0.

4 Experiment and results

4.1 *Experimental settings*

In the experimental analysis, datasets containing customer transaction data from an e-commerce company were prepared as presented in Table 2. Customer segmentation was performed using commonly employed soft clustering algorithms: FCM, FCM-T2, EM, and Soft-G. FCM, FCM-T2, and EM were implemented using RStudio. The implementations of GRC and GRA for Soft-G were coded in C#, while SqlServer 2008 R2 was employed for database operations.

In this analysis, the number of clusters was selected as k=3, k=4, k=5, and k=6 in accordance with the literature [27, 60, 61, 62, 64, 65]. However, hard clustering algorithms like K-Means, K-Medoids, Hierarchical, and SOM, commonly used in customer segmentation, were not included in this study. RStudio was employed to compute the validation index values for FSC, PE, PC, and XB methods.

4.2 *Clustering results*

Algorithms used in the analysis were applied to Ds1, Ds2, Ds3, Ds4, Ds5 and Ds6 according to the number of clusters k=3, k=4, k=5 and k=6. FCM, FCM-T2, EM and Soft-G membership values used for clustering Ds1, k=3 were calculated for each item and membership values were given in Appendix A. The clustering results of Ds1 are given in Table 4.

Based on the clusters presented in Table 4, the results of FCM and EM clustering indicate that unlike FCM-T2 and Soft-G, there are no items assigned to multiple clusters. As can be seen from the membership values in Appendix A FCM and EM algorithms exhibited hard clustering behaviour. This means that no item could be associated with two or more clusters. However, FCM-T2 suggests that Cust6 is located in both Cluster2 and Cluster3. Notably, Soft-G exhibits different clustering results compared to the others. Soft-G assigned the Cust3 to both Cluster2 and Cluster3. Upon reviewing Table 4, it is evident that clusters are generally formed in an unsimilar manner. However, Soft-G and FCM-T2 detect more overlapping items, whereas FCM and EM do not identify any overlapping items.

As seen in Table 4, Cust6 and Cust3 exhibit inter-cluster overlapping characteristics. In other words, according to the FCM-T2 results, Cust6, and according to the Soft-G results, Cust3, are customers positioned near cluster boundaries and demonstrating overlapping behaviours. Based on the values presented in Table 3, Cust6 appears to be a loyal customer with high shopping frequency and login activity, as well as a long-standing presence on the platform. The relatively high number of reviews and cart removal actions indicates active engagement with the platform. In this context, Cust6 can be classified as a high-value customer. In contrast, Cust3 is characterized by low shopping and login activity, a relatively low number of reviews, and a high rate of cart removals. Therefore, Cust3 can be considered a risky customer for the platform. FCM-T2 clustering results indicate that Cust6, and Soft-G clustering results indicate that Cust3, belong to customer segments whose consumption behaviours are modifiable and open to managerial intervention.

Table 4. Clustering results of items (for Ds1, k=3)

Clusters	FCM	FCM-T2	EM	Soft-G
Cluster 1	Cust1, Cust2, Cust3, Cust4, Cust7, Cust8, Cust9, Cust10	Cust1, Cust2, Cust4, Cust8, Cust9, Cust10	Cust1, Cust2, Cust3, Cust4, Cust7, Cust8, Cust9, Cust10	Cust6
Cluster 2	Cust6	Cust6, Cust7	Cust5	Cust1, Cust2, Cust3, Cust4, Cust5, Cust7, Cust10
Cluster 3	Cust5	Cust3, Cust5, Cust6	Cust6	Cust3, Cust8, Cust9

4.3 *Comparing results*

Internal validation metrics are commonly employed to assess the clustering performance of datasets not having cluster information [89, 90]. In this study, the dataset does not provide any information about the clusters to which the items belong. Hence, PC, PE, FSC, and XB internal validation indexes were employed to evaluate the clustering performance. The cluster validation values for the algorithms are presented in Table 5.

Based on the validation index values in Table 5, it is evident that algorithm performances are lower in datasets Ds1, Ds2, and Ds3, which have a smaller number of items, compared to Ds4, Ds5, and Ds6 datasets. Choosing a successful algorithm is particularly challenging process in Ds1. As shown in Figure 3 and Figure 4, the algorithm with the highest performance varies depending on the PE and PC values and the number of clusters. For instance, at k=3, EM has the highest PC value but not the lowest PE value. Similar patterns can be observed for k=5 and k=6. For this reason, it can be said that a successful clustering algorithm for Ds1 cannot be determined by PE and PC that should be evaluated together.

Table 5. Validation values of the clustering

Data sets	Algorithms	k=3				k=4				k=5				k=6			
		FSC	PE	PC	XB	FSC	PE	PC	XB	FSC	PE	PC	XB	FSC	PE	PC	XB
Ds1 (10 items)	FCM	0.55	**0.56**	0.80	1.3	0.09	0.23	0.93	25.3	0.26	**0.29**	0.91	102	0.22	**0.26**	0.92	83.4
	FCM-T2	0.38	1.34	0.44	**0.5**	0.47	1.99	0.25	4.1	**0.48**	2.32	0.2	**2.2**	0.36	2.58	0.16	77.4
	EM	**0.59**	8.48	**0.99**	1.4	0.34	**0.00**	**0.99**	5.6	0.34	3.43	**0.99**	67.5	0.33	2.99	**0.99**	65.5
	Soft-G	0.52	1.48	0.38	7.4	**0.51**	1.87	0.58	**3.4**	**0.48**	2.39	0.95	52.1	**0.46**	2.84	**0.99**	**62.5**
Ds2 (50 items)	FCM	0.14	0.29	0.90	15.9	0.13	0.49	0.82	15.4	0.50	1.51	0.48	29.7	0.30	0.55	0.82	19.1
	FCM-T2	0.08	1.37	0.42	**1.3**	0.57	2	0.25	1.6	0.57	2.32	0.2	4.7	0.30	2.58	0.16	5.5
	EM	0.32	**0.07**	**0.97**	2.3	0.32	**0.13**	**0.94**	4.1	0.28	**0.11**	**0.95**	85.1	0.46	**0.05**	0.96	7.4
	Soft-G	**0.44**	0.49	0.54	**1.3**	**0.58**	0.74	0.61	**1.4**	**0.66**	1.2	0.88	**4.3**	**0.52**	0.96	**0.97**	**4.7**
Ds3 (100 items)	FCM	0.63	0.87	0.67	0.4	0.57	1.16	0.57	2.4	0.60	1.28	0.56	1.6	0.64	1.35	0.56	5.9
	FCM-T2	0.47	1.3	0.46	**1.3**	0.24	2	0.25	**1.4**	0.07	2.3	0.2	3.7	0.44	2.58	0.16	5.2
	EM	0.25	**0.02**	**0.99**	21.1	0.07	**0.02**	**0.99**	50.5	0.35	**0.22**	0.91	6.8	0.01	**0.11**	**0.95**	120
	Soft-G	**0.73**	0.96	0.41	**1.3**	**0.65**	1.12	0.79	2.8	**0.62**	1.26	**0.94**	**1.3**	**0.67**	1.32	**0.95**	**3.7**
Ds4 (500 items)	FCM	**0.76**	0.61	0.75	**0.1**	**0.76**	0.77	**0.72**	**0.1**	0.66	**0.99**	**0.64**	**0.1**	**0.70**	1.08	0.63	**0.5**
	FCM-T2	0.35	1.31	0.45	0.9	0.17	1.87	0.29	1.9	0.09	2.32	0.2	4.1	0.25	2.58	0.16	5.7
	EM	0.27	**0.13**	**0.94**	14.4	0.30	**0.76**	0.65	912	0.25	1.16	0.53	7522	0.27	**0.69**	**0.70**	45816
	Soft-G	0.65	1.58	0.68	0.2	0.68	1.68	0.55	0.2	**0.69**	2.15	0.53	0.3	**0.70**	2.75	0.50	0.8
Ds5 (1000 items)	FCM	**0.82**	0.61	0.77	**0.1**	**0.71**	0.85	0.69	**0.1**	**0.73**	0.99	0.65	**0.1**	**0.73**	1.1	**0.63**	**0.2**
	FCM-T2	0.43	1.33	0.45	1.4	0.25	1.9	0.28	17.9	0.61	2.32	0.2	6.8	0.66	2.5	0.16	2.5
	EM	0.33	**0.32**	**0.85**	7.03	0.22	**0.52**	**0.78**	77.9	0.02	**0.61**	**0.74**	47.68	0.41	**1**	0.60	35318
	Soft-G	0.62	1.28	0.67	2.4	0.59	1.42	0.5	6.1	0.61	1.83	0.48	8.6	0.63	1.96	0.54	9.7
Ds6 (5130 items)	FCM	**0.87**	0.54	0.81	**0.2**	**0.69**	0.84	0.68	**0.1**	0.74	0.94	0.67	**0.1**	0.71	0.99	0.66	0.3
	FCM-T2	0.68	1.3	0.46	**0.2**	0.63	1.86	0.29	**0.1**	**0.75**	2.32	0.2	**0.1**	**0.75**	2.58	0.16	**0.1**
	EM	0.21	**0.21**	**0.90**	5.9	0.09	**0.34**	**0.86**	45.4	0.11	**0.31**	**0.86**	22.7	0.05	0.30	**0.87**	38.5
	Soft-G	0.61	1.48	0.23	2.4	0.61	2.51	0.54	6.18	0.55	2.81	0.43	8.6	0.58	3.96	0.31	9.72

Figure 5 and Figure 6 present the FSC and XB index values for Ds1. When clustering the Ds1 dataset with k=4, k=5, and k=6, Soft-G emerges as the most successful algorithm based on the FSC index values. The same result is observed for XB index values. It can be stated that Soft-G and FCM-T2, which have similar XB values, outperform the other algorithms. Conversely, the FCM algorithm demonstrates the lowest performance in clustering of Ds1, as indicated by the FSC and XB values.

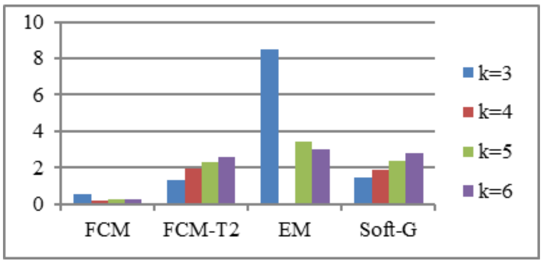

Figure 3. PE values of algorithms for Ds1

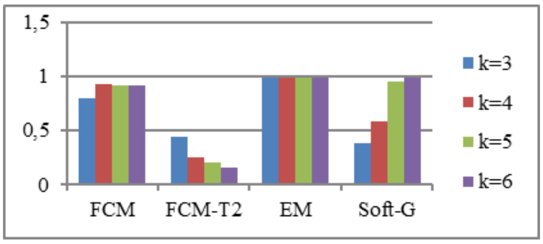

Figure 4. PC values of algorithms for Ds1

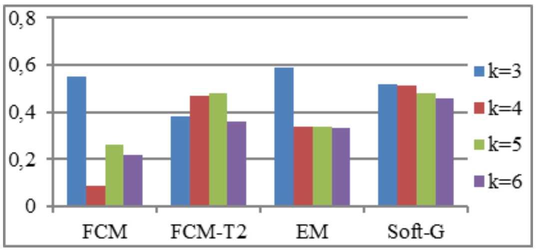

Figure 5. FSC values of algorithms for Ds1

The clustering of Ds2 and Ds3 datasets, which contain a higher number of items compared to Ds1, yielded clearer performance results. Based on the PE and PC values, the EM algorithm demonstrated high performance. However, when considering the FSC and XB indexes, both EM and FCM emerged as the least successful algorithms. On the other hand, Soft-G algorithm exhibited higher performance than other algorithms in Ds2 and Ds3 according to FSC and XB.

It is noteworthy that fuzzy-based algorithms have demonstrated higher performance in Ds4 than in Ds1, Ds2 and Ds3. Particularly, the FSC and XB values indicate that the FCM and Soft-G algorithms yield favourable results in Ds4. In Ds5 and Ds6 datasets, it was observed that FCM and FCM-T2 outperformed other algorithms. In this context it can be stated that fuzzy-based algorithms are more successful in datasets having more data. According to XB values in Ds4, Ds5 and Ds6, the performance of the EM algorithm is relatively lower compared to the others. Similarly, FSC values in these datasets show that EM is not a successful algorithm. But, PC and PE values indicate EM has better clustering results. This finding reveals confusion on EM is successful or not. We can explain this contradiction by the different methodologies of validity indexes. While XB and FSC have a methodology based on the compactness-separation approach, PE-PC indexes have a calculation method based on the average of membership values. For example, the average of items with high membership values provides a high PC value. However, a high PC value does not mean that the compactness value of items within the cluster will be high. Similarly, items with low PC values may have close values to each other and their compactness may be high. While the arithmetic mean-based PE and PC indices provide measures for assessing the degree of fuzziness in clustering, the XB and FSC indices, which are based on the differences among cluster elements, evaluate intra-cluster compactness and inter-cluster separation. In this context, the XB and FSC index values ensure more reliable results for assessing the performance of soft clustering methods. Considering that the quality of a cluster should be evaluated by the closeness of items within the cluster (compactness) and the difference of the cluster from other clusters (separation), it can be said that the PE-PC approach will be insufficient in deciding on cluster quality. In other words, it would be appropriate to take XB and FSC values into consideration when evaluating the algorithms. In this context, it cannot be said that EM, which has low FSC and high XB values in all data sets, has high clustering performance by only looking at PE-PC values.

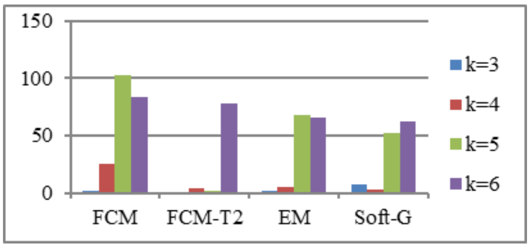

Figure 6. XB values of algorithms for Ds1

Based on the study findings, it is evident that the number of clusters has an impact on algorithm performances in cluster analysis. The results indicate that, across all datasets, clustering with k=3 generally yield higher FSC values and lower XB values. Conversely, as the number of clusters increases to k=5 and k=6, FSC values tend to decrease while XB values increase. In essence, the analysis suggests that clustering performance declines as the number of clusters to be formed increases. Consequently, choosing less number of clusters will be suitable for high-performance soft clustering analyses.

According to the experimental results, it has been observed that the algorithm performances are low in the datasets with a low number of items, and the algorithms have higher performances with the increase in the amount of data. These findings are similar to the soft clustering literature. It has been observed that the Soft-G gives successful results with a small amount of data. It has been determined that the Soft-G is more successful than other algorithms in Ds1, Ds2, Ds3 and Ds4. In addition, due to the rising amount of data, no significant decrement or increment was observed in the performance of Soft-G. So, it can be stated that Soft-G is a stable algorithm. On the other hand, there were significant increases in the performance of fuzzy-based algorithms due to increase in data amount. In particular, the FCM algorithm exhibited significant performance improvements in Ds4, Ds5 and Ds6.

Another finding within the scope of the study is that the EM algorithm does not perform as well as stated in the literature. Contrary to the findings of [65], it cannot be said that EM exhibits high performance in small datasets. Unlike [26] and [71], the results of this study show that EM does not produce clustering results similar to FCM. FSC and XB show that EM has poor performance in any dataset. In this context, the EM algorithm is not recommended for soft clustering in this article.

It has been noted that FCM and FCM-T2 algorithms exhibit lower performances in small datasets. However, upon analysis, it was discovered that FCM-T2 yielded more successful outcomes than FCM in Ds1, Ds2, and Ds3 datasets. This observation can be attributed to the fact that small datasets have high outlier value problems. FCM-T2, being less sensitive to discrete values compared to FCM, demonstrates greater success in such scenarios. This finding aligns with the findings of [29] and [69], supports for their findings.

5 Conclusion and suggestions

Soft clustering is an unsupervised machine learning method to use exploring the customer structure. Most of the studies on customer segmentation preferring soft approaches use the FCM algorithm. Since FCM based algorithms have some deficiencies, new methods are developed on the basis of fuzzy. It is generally accepted in the literature that fuzzy-based approaches fail in small datasets and are more affected by discrete values. On the other hand, GST enables the execution of meaningful and reliable relational analyses even with a limited amount of data. Owing to this characteristic, GST-based methods produce effective and consistent results particularly in the analysis of small-scale datasets. In this study, the Soft-G algorithm, which was developed for soft clustering purposes based on GST, demonstrates superior performance compared to other algorithms in clustering small datasets and yields successful outcomes.

A dataset comprising data from 5130 customers was created using variables selected from an e-commerce company based on the RFM approach. Subsequently, six datasets were formed with 10, 50, 100, 500, 1000, and 5130 items, respectively. Cluster analysis was conducted on these datasets using the FCM, FCM-T2, EM, and Soft-G algorithms, and the results were evaluated using the FSC, PE, PC, and XB validation indices. The analysis results revealed that the Soft-G algorithm exhibited superior performance compared to other algorithms in clustering the smaller datasets (Ds1, Ds2, and Ds3). While FCM yielded unsuccessful results in Ds1, Ds2, and Ds3, it showed successful outcomes in clustering Ds4, Ds5, and Ds6. Consequently, for datasets smaller than 500 items, the Soft-G algorithm is recommended, whereas FCM is suitable for datasets larger than 500 items. Additionally, in larger datasets, both FCM and FCM-T2 demonstrated successful results. Therefore, the choice between FCM and FCM-T2 is less critical for larger datasets. However, when opting for a fuzzy-based algorithm instead of Soft-G in small datasets, FCM-T2 is a more suitable choice compared to FCM. On the other hand, the study found that the EM algorithm, often used in soft clustering in the literature, achieved higher success rates than fuzzy-based algorithms in small datasets but exhibited lower performance than Soft-G. Accordingly, the study suggests Soft-G algorithm instead of EM for soft clustering. Furthermore, since increasing the number of clusters leads to a decrease in algorithm performance, aiming for fewer clusters is recommended to create more successful clusters regardless of the dataset's size for customer segmentation.

The study includes the analysis of only one dataset due to the use of real e-commerce transactional data. In addition, comparisons are limited to soft clustering algorithms commonly used in the literature. Determining the performances of the developed method in different datasets and comparing the results with different algorithms will expand the application area of the method. In addition, study on performance analysis of datasets having outlier problem will be a valuable article that reveals the success of the method.

Conflict of interest

The author declares that there is no conflict of interest.

Similarity rate (iThenticate): %15

References

[1] S. Peker and Ö. Kart, Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review. Journal of Data Information and Management, 5, 1–21, 2023. https://doi.org/10.1007/s42488-023-00085-x.

[2] A. Gomes and M. T. Meisen, A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and E-Business Management, 21, 527–570, 2023. https://doi.org/10.1007/s10257-023-00640-4.

[3] T. P. Beane and D. M. Ennis, Market segmentation: A review. European Journal of Marketing, 21(5), 20–42, 1987. https://doi.org/10.1108/EUM0000000004695.

[4] S. Das and J. Nayak, Customer Segmentation via Data Mining Techniques: State-of-the-Art Review. In: J. Nayak, H. Behera, B. Naik, S. Vimal and D. Pelusi (Eds.), Computational Intelligence in Data Mining Smart Innovation Systems and Technologies, Vol 281, pp. 489-507, Singapore, 2022. https://doi.org/10.1007/978-981-16-9447-9_38

[5] A. Hiziroglu, Soft computing applications in customer segmentation: State-of-art review and critique. Expert Systems with Applications, 40 (16), 6491–6507, 2013. https://doi.org/10.1016/j.eswa.2013.05.052.
[6] M. Alborzi and M. Khanbabaei, Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. International Journal of Business Information Systems, 23 (1), 1-22, 2016. http://dx.doi.org/10.1504/IJBIS.2016.078020.
[7] J. Zhou, J. Wei and B. Xu, Customer segmentation by web content mining. Journal of Retailing and Consumer Services, 61, 102588, 1-10, 2021. https://doi.org/10.1016/j.jretconser.2021.102588.
[8] Y. Zong and H. Xing, Customer stratification theory and value evaluation-analysis based on improved RFM model. Journal of Intelligent and Fuzzy Systems, 40, 4155–4167, 2021. https://doi.org/10.3233/JIFS-200737.
[9] H. Zare and S. Emadi, Determination of customer satisfaction using Improved K-means algorithm. Soft Computing, 24, 16947–16965, 2020. https://doi.org/10.1007/s00500-020-04988-4.
[10] C. H. Cheng and Y. S. Chen, Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications, 36 (3), 4176–4184, 2009. https://doi.org/10.1016/j.eswa.2008.04.003.
[11] R. S. Wu and P. H. Chou, Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications, 10 (3), 331-341, 2011. https://doi.org/10.1016/j.elerap.2010.11.002.
[12] J. Wu, S. Li, L. Wen-Pin, T. Sang-Bing, L. Yuanyuan, Y. Liping and X. Guangshu, An empirical study on customer segmentation by purchase behaviours using a RFM model and K-Means algorithm. Mathematical Problems in Engineering, 8884227, 1-7, 2020. https://doi.org/10.1155/2020/8884227.
[13] C. Y. Tsai and C. C. Chiu, A purchase-based market segmentation methodology. Expert Systems with Applications, 27 (2), 265–276, 2004. https://doi.org/doi:10.1016/j.eswa.2004.02.005.
[14] J. Miglautsch, Thoughts on RFM scoring. Journal of Database Marketing and Customer Strategy Management, 8, 67–72, 2000. https://doi.org/10.1057/palgrave.jdm.3240019.
[15] I. C. Yeh, K. C. Yang and T. M. Ting, Knowledge discovery on RFM model using Bernoulli sequence. Expert Systems with Applications, 36 (3), 5866-5871, 2009. https://doi.org/10.1016/j.eswa.2008.07.018.
[16] H. H. Chang and S. F. Tsay, Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. Journal of Information Management, 11, 161-203, 2004. https://sid.ir/paper/624927/en.
[17] S. Peker, A. Kocyigit and P. E. Eren, LRFMP model for customer segmentation in the grocery retail industry: A case study. Marketing Intelligence and Planning, 35 (4), 544-559, 2017. https://doi.org/10.1108/MIP-11-2016-0210.
[18] M. Patel, D. Patel, J. Rana, V. Chauhan and B. R. Bhatt, Customer segmentation using machine learning. International Journal for Research in Applied Science and Engineering Technology, 8 (6), 2104-2108, 2020. https://doi.org/10.22214/ijraset.2020.6344.
[19] O. Dogan, A. Hiziroglu and O. F. Seymen, Segmentation of Retail Consumers with Soft Clustering Approach. In: C. Kahraman, S. Cevik, B. Oztaysi, I. Sari, S. Cebi and A. Tolga (Eds.), Intelligent and Fuzzy Techniques: Smart and Innovative Solutions, Springer, Cham,Vol 1197, pp. 39-46, İstanbul, 2021. https://doi.org/10.1007/978-3-030-51156-2_6.
[20] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques. Morgan Kaufmann Publications, USA, 2012.
[21] E. Alpaydin, Introduction to Machine Learning (2nd ed.). MIT Press, London, 2010.
[22] S. Mousavi, A. Esfahanipour and M. H. Z. Fazel, A modular Takagi-Sugeno-Kang (TSK) system based on a modified hybrid soft clustering for stock selection. Scientia Iranica, 28 (4), 2342-2360, 2021. https://doi.org/10.24200/sci.2019.52323.2661.
[23] N. Gautam and N. Kumar, Customer segmentation using k-means clustering for developing sustainable marketing strategies. Business Informatics, 16 (1), 72–82, 2022. https://doi.org/10.17323/2587-814X.2022.1.72.82.
[24] K. Tabianan, S. Velu and V. Ravi, K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14 (12), 7243, 2022. https://doi.org/10.3390/su14127243.
[25] Z. Wu, L. Jin, J. Zhao, L. Jing and L. Chen, Research on segmenting e-commerce customer through an improved K-medoids clustering algorithm. Computational Intelligence and Neuroscience, 930613, 2022. https://doi.org/10.1155/2022/9930613.
[26] M. B. Ferraro and P. Giordani, Soft clustering. WIREs Computational Statistics, 12, e1480, 2020. https://doi.org/10.1002/wics.1480.
[27] C. C. Chan, Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. Expert Systems with Applications, 34 (4), 2754–2762, 2008. https://doi.org/10.1016/j.eswa.2007.05.043.
[28] G. Peters, F. Crespo, P. Lingras and R. Weber, Soft clustering – Fuzzy and Rough approaches and their extensions and derivatives. International Journal of Approximate Reasoning, 54 (2), 307-322, 2013. https://doi.org/10.1016/j.ijar.2012.10.003.
[29] A. Gosain and S. Dahiya, Performance analysis of various fuzzy clustering algorithms: A review. Procedia Computer Science, 79, 100–111, 2016. https://doi.org/10.1016/j.procs.2016.03.014.

[30] S. Khanmohammadi, N. Adibeig and S. Shanehbandy, An improved overlapping K-means clustering method for medical applications. Expert Systems with Applications, 67, 12–18, 2017. https://doi.org/10.1016/j.eswa.2016.09.025.

[31] S. Askari, Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. Expert Systems with Applications, 165, 113856, 2021. https://doi.org/10.1016/j.eswa.2020.113856.

[32] R. J. Kuo, Y. L. An, H. S. Wang and W. J. Chung, Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. Expert Systems with Applications, 30 (2), 313–324, 2006. https://doi.org/10.1016/j.eswa.2005.07.036.

[33] H. Fidan and M. E Yuksel, A novel short text clustering model based on Grey System Theory. Arabian Journal for Science and Engineering, 45, 2865–2882, 2020. https://doi.org/10.1007/s13369-019-04191-0.

[34] S. E. Hashemi, F. Gholian-Jouybari and M. Hajiaghaei-Keshteli, A fuzzy C-means algorithm for optimizing data clustering. Expert Systems with Applications, 227, 120377, 2023. https://doi.org/10.1016/j.eswa.2023.120377.

[35] J. Elmi, M. Eftekhari, A. Mehrpooya and M. R. Ravari, A novel framework based on the multi-label classification for dynamic selection of classifiers. International Journal of Machine Learning and Cybernetics, 14 (2), 137–2154, 2023. https://doi.org/10.1007/s13042-022-01751-z.

[36] H. Fidan and M. E. Yuksel, A comparative study for determining Covid-19 risk levels by unsupervised machine learning methods. Expert Systems with Application, 190, 116243, 2022. https://doi.org/10.1016/j.eswa.2021.116243.

[37] A. Sharma, S. K. Singh, E. Badwal, S. Kumar, B. B. Gupta, V. Arya, K. T. Chui and D. Santaniello, Fuzzy Based Clustering of Consumers' Big Data in Industrial Applications. IEEE International Conference on Consumer Electronics (ICCE), pp. 01-03, Las Vegas, NV, USA, 2023. https://doi.org/10.1109/ICCE56470.2023.10043451.

[38] J. L. Deng, Control problems of grey systems. System and Control Letters, 1 (5), 288-294, 1982. https://doi.org/10.1016/S0167-6911(82)80025-X.

[39] X. Jin, Grey relational clustering method and its application. Journal Grey System, 3, 181–188, 1993.

[40] W. H. Wu, C. T. Lin, K. H. Peng and C. C. Huang, Applying hierarchical grey relation clustering analysis to geographical information systems—A case study of the hospitals in Taipei city. Expert Systems with Application, 39 (8), 7247–7254, 2012. https://doi.org/10.1016/j.eswa.2012.01.052.

[41] C. L. Chen, F. S. C. Tseng and T. Liang, An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data and Knowledge Engineering, 69 (11), 1208-1226, 2010. https://doi.org/10.1016/j.datak.2010.08.003.

[42] Y. H. Chen, E. J. L. Lu and Y. W. Cheng, Categorization of multiple documents using fuzzy overlapping clustering based on formal concept analysis. International Journal of Software Engineering and Knowledge Engineering, 30 (5), 631–647, 2020. https://doi.org/10.1142/S0218194020500229.

[43] N. Jardine and R. Sibson, A model for taxonomy. Mathematical Biosciences, 2 (3-4), 465–482, 1968. https://doi.org/10.1016/0025-5564(68)90030-8.

[44] E. H. Ruspini, A new approach to clustering, Information and Control, 15 (1), 22-32, 1969. https://doi.org/10.1016/S0019-9958(69)90591-9.

[45] E. H. Ruspini, Numerical methods for fuzzy clustering. Information Science, 2 (3), 319-350, 1970. https://doi.org/10.1016/S0020-0255(70)80056-1.

[46] E. H. Ruspini, New experimental results in fuzzy clustering. Information Sciences, 6, 273-284, 1973. https://doi.org/10.1016/0020-0255(73)90043-1.

[47] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(3), 32-57, 1973. https://doi.org/10.1080/01969727308546046.

[48] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4 (1), 95-104, 1974. https://doi.org/10.1080/01969727408546059.

[49] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.

[50] K. L. Wu, Analysis of parameter selections for fuzzy c-means. Pattern Recognition, 45 (1), 407–415, 2012. https://doi.org/10.1016/j.patcog.2011.07.012.

[51] T. M. Nogueira, S. O. Rezende and H. Camargo, On the use of fuzzy rules to text document classification. In Proceedings 10th International Conference on Hybrid Intelligent Systems, pp. 19-24, Atlanta, USA, 2010. https://doi.org/10.1109/HIS.2010.5600076.

[52] F. C. H. Rhee and C. Hwang, A type-2 fuzzy C-means clustering algorithm. In Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), pp. 1926-1929, Vancouver, BC, Canada, 2001. https://doi.org/10.1109/NAFIPS.2001.944361.

[53] A. P. Dempster, N. M. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39 (1), 1–22, 1977. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[54] K. Kalti and M. A. Mahjoub, Image segmentation by gaussian mixture models and modified FCM algorithm. International Arab Journal Information Technology, 11 (1), 11-18, 2014.

[55] R.M. Prakash and R. S. S. Kumari, Spatial fuzzy C means and Expectation Maximization algorithms with bias correction for segmentation of MR brain images. Journal of Medical Systems, 41 (1), 2016. https://doi.org/10.1007/s10916-016-0662-7.

[56] C. Carson, S. Belongie, H. Greenspan and J. Malik, Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (8), 1026–1038, 2002. https://doi.org/10.1109/TPAMI.2002.1023800.

[57] W. R. Smith, Product differentiation and market segmentation as an alternative marketing strategy. Journal of Marketing, 21, 3–8, 1956. https://doi.org/10.2307/1247695.

[58] M. A. Sicilia and E. Garcia, On fuzziness in relationship value segmentation: Applications to personalized e-commerce. ACM SIGecom Exchanges, 4 (2), 1-10, 2003. https://doi.org/10.1145/1120709.1120712.

[59] E. Turban, J. Lee, D. King and H.M. Chung, Electronic Commerce a Managerial Perspective. Prentice Hall International Inc., New Jersey, USA, 2000.

[60] M. Ozer, User segmentation of online music services using fuzzy clustering. Omega, 29 (2), 193-206, 2001. https://doi.org/10.1016/S0305-0483(00)00042-6.

[61] A. Bhatnagar and S. Ghose, A latent class segmentation analysis of e-shoppers. Journal of Business Research, 57 (7), 758-767, 2004. https://doi.org/10.1016/S0148-2963(02)00357-0.

[62] J. J. Rho, B. J. Moon, Y. J. Kim and D. H. Yang, Internet customer segmentation using web log data. Journal of Business and Economics Research, 2 (11), 2004. https://doi.org/10.19030/jber.v2i11.2940.

[63] H. Wang, An unsupervised purchase-based customer clustering method for e-supply chain. IEEE International Conference on Service Operations and Logistics, and Informatics, pp. 686-688, Beijing, China, 2008. https://doi.org/10.1109/SOLI.2008.4686485.

[64] T. Hong and E. Kim, Segmenting customers in online stores based on factors that affect the customer's intention to purchase. Expert Systems with Applications, 39b(2), 2127–2131, 2012. https://doi.org/10.1016/j.eswa.2011.07.114.

[65] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, Customer segmentation using K-means clustering. In International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 135-139, Belgaum, India, 2018. https://doi.org/10.1109/CTEMS.2018.8769171.

[66] S. Koul and T. M. Philip, Customer segmentation techniques on e-commerce. In International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 135-138, Greater Noida, India, 2021. https://doi.org/10.1109/icacite51222.2021.9404659.

[67] J. Basak and M. Kumar, E-Commerce and soft computing: Scopes of research. IETE Technical Review, 18 (4), 283–293, 2001. https://doi.org/10.1080/02564602.2001.11416974.

[68] H. W. Shin and S. Y. Sohn, Segmentation of stock trading customers according to potential value. Expert Systems with Applications, 27 (1), 27–33, 2004. https://doi.org/10.1016/j.eswa.2003.12.002.

[69] P. Kaur, A. K. Soni and A. Gosain, Robust intuitionistic fuzzy c-means clustering for linearly and nonlinearly separable data. In International Conference on Image Information Processing, pp. 1-6, Shimla, India, 2011. https://doi.org/10.1109/iciip.2011.6108908.

[70] K. Khalili-Damghani, F. Abdi and S. Abolmakarem, Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. Applied Soft Computing, 73, 816-828, 2018. https://doi.org/10.1016/j.asoc.2018.09.001.

[71] F. Yoseph, N. H. A. H. Malim and M. Almalaily, New behavioural segmentation methods to understand consumers in retail industry. International Journal of Computer Science and Information Technology, 11 (1), 43-61, 2019. https://doi.org/10.5121/ijcsit.2019.11104.

[72] C. Wu, and X. Zhang, A self-learning iterative weighted possibilistic fuzzy c-means clustering via adaptive fusion. Expert Systems and Application, 209, 118280, 2022. https://doi.org/10.1016/j.eswa.2022.118280.

[73] B. Yu, Z. Zheng, M. Cai, W. Pedrycz, and W. Ding, FRCM: A fuzzy rough c-means clustering method. Fuzzy Sets and Systems, 480, 108860, 2024. https://doi.org/10.1016/j.fss.2024.108860.

[74] G. Yue, Y. Qu, L. Yang, C. Shang, A. Deng, F. Chao, and Q. Shen, Fuzzy-rough intrigued harmonic discrepancy clustering. IEEE Trans. Fuzzy Systems, 2023. https://doi.org/10.1109/TFUZZ.2023.3247912.

[75] B. Alekhya, and R. Sasikumar, An ensemble approach for healthcare application and diagnosis using natural language processing. Cognitive Neurodynamics, 16, 1203–1220, 2022. https://doi.org/10.1007/s11571-021-09758-y.

[76] A. G. Oskouei, N. Samadi, and J. Tanha, Feature-weight and cluster-weight learning in fuzzy c-means method for semi-supervised clustering. Applied Soft Computing, 161, 111712, 2024. https://doi.org/10.1016/j.asoc.2024.111712.

[77] J. Xu, P. Wang, G. Tian et al., Short text clustering via convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modelling for Natural Language Processing, pp. 62–69, Denver, Colorado, USA, 2015. https://doi.org/10.3115/v1/W15-1509.

[78] J. Liu, W.C. Chang, Y. Wu and Y. Yang, Deep learning for extreme multi-label text classification. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17, pp. 115-124, Shinjuku, Tokyo, Japan, 2017. https://doi.org/10.1145/3077136.3080834.

[79] K. Shimura, J. Li and F. Fukumoto, HFT-CNN: Learning hierarchical category structure for multi-label short text categorization, In EMNLP Association for Computational Linguistics, pp. 811–816, Brussels,

Belgium, 2018. https://doi.org/10.18653/v1/D18-1093.

[80] T. Menzies, S. Majumder, and N. Balaji, 500+ Times Faster than Deep Learning: A Case Study Exploring Faster Methods for Text Mining StackOverflow. IEEE/ACM 15th International Conference on Mining Software Repositories (MSR), pp. 554-563, Gothenburg, Sweden, 2018. https://doi.org/10.1145/3196398.3196424.

[81] A. Rahui, M. Allouch, and M. Alami, Optimization of the tungsten inert gas welding parameters of mild steel thin sheets through the gray relational analysis method. International Journal of Advanced Manufacturing Technologies, 133, 4437–4452, 2024. https://doi.org/10.1007/s00170-024-13984-z.

[82] F. Abdullahu, F. Zhujani, G. Todorov, and K. Kamberov, An Experimental Analysis of Taguchi-Based Gray Relational Analysis, Weighted Gray Relational Analysis, and Data Envelopment Analysis Ranking Method Multi-Criteria Decision-Making Approaches to Multiple-Quality Characteristic Optimization in the CNC Drilling Process. Processes, 12 (6), 1212, 2024. https://doi.org/10.3390/pr12061212.

[83] G. Singh, and A. Pandey, Environmental sustainability integrated supplier selection in electric vehicle supply chains: A grey relational analysis approach. Environment Development and Sustainability, 2024. https://doi.org/10.1007/s10668-024-05294-x.

[84] Z. Kan, Y. Wei, T. Zhao, and J. Cao, Risk evaluation of submarine pipelines in FMEA by combination of gray relation projection and VIKOR method. Ocean Engineering, 302, 117695, 2024. https://doi.org/10.1016/j.oceaneng.2024.117695.

[85] Y. Wang, and H. Li, A Novel Single-Sample Retinal Vessel Segmentation Method Based on Grey Relational Analysis. Sensors, 24 (13), 4326, 2024. https://doi.org/10.3390/s24134326.

[86] C. Liu, M. Pang, and C. T. Lung, Image segmentation via dilated U-Net model and multi-scale gray correlation-based approach. Circuits Systems and Signal Processing, 43, 1697–1714, 2024. https://doi.org/10.1007/s00034-023-02532-x.

[87] M. Wang, Multi-data multiple gray clustering analysis based on layered mining for ubiquitous clouds and social internet of things. Internet Technology Letters, 6, e252, 2023. https://doi.org/10.1002/itl2.252.

[88] B.F. Yıldırım, Gri İlişkisel Analiz. In Çok Kriterli Karar Verme Yöntemleri, B.F. Yıldırım and E. Önder (Eds.), Dora Basım Yayın, pp. 229-236, Bursa, Türkiye, 2015.

[89] M. Halkidi, Y. Batistakis and M. Vazirgiannis, On clustering validation techniques. Journal of Intelligent Information Systems, 17 (2/3), 107–145, 2001. https://doi.org/10.1023/A:1012801612483.

[90] K. L. Wu and M. S. Yang, A cluster validity index for fuzzy clustering. Pattern Recognition Letters, 26 (9), 1275–1291, 2005. https://doi.org/10.1016/j.patrec.2004.11.022.

[91] D. Swift, K. Cresswell, R. Johnson, S. Stilianoudakis, and X. T. Wei, A review of normalization and differential abundance methods for microbiome counts data. Wiley Interdisciplinary Reviews-Computational Statistics, 15 (1), e1586, 2023. https://doi.org/10.1002/wics.1586.

[92] A. Elkhouly, A. M. Andrew, H. A. Rahim, N. Abdulaziz, M. F. A. Malek, and S. Siddique, Data-driven audiogram classifier using data normalization and multi-stage feature selection. Scientific Reports, 13 (1), 2023. https://doi.org/10.1038/s41598-022-25411-y.

[93] J. C. Bezdek, Cluster validity with fuzzy sets. Journal of Cybernetics, 3 (3), 58–73, 1973. https://doi.org/10.1080/01969727308546047.

[94] J. C. Bezdek and N. R. Pal, Some new indexes of cluster validity. IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics), 28 (3), 301-315, 1998. https://doi.org/10.1109/3477.678624.

[95] Y. Fukuyama and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method. In Proceeding of fifth Fuzzy Systems Symposium, pp. 247–250, 1989. https://api.semanticscholar.org/CorpusID:207926333. Accessed 12 September 2024.

[96] X. L. Xie and G. Beni, A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13 (8), 841–847, 1991. https://doi.org/10.1109/34.85677.

Appendix A

Memberships (FCM and FCM-T2), probality (EM) and relational degree (Soft-G) values for Ds1, k=3

Customers	FCM			FCM-T2			EM			Soft-G		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Cust1	0.95	0.03	0.02	0.74	0.18	0.08	1.00	0.00	0.00	0.27	0.52	0.36
Cust2	0.90	0.05	0.05	0.64	0.23	0.12	1.0	0.00	0.00	0.23	0.46	0.39
Cust3	0.49	0.12	0.39	0.19	0.28	0.53	1.0	0.00	0.00	0.22	0.57	0.53
Cust4	0.88	0.04	0.08	0.54	0.29	0.18	1.0	0.00	0.00	0.23	0.59	0.45
Cust5	0.01	0.01	0.98	0.22	0.29	0.49	0.00	1.00	0.00	0.25	0.46	0.34
Cust6	0.01	0.99	0.01	0.25	0.36	0.40	0.00	0.00	1.00	1.00	0.23	0.23
Cust7	0.61	0.24	0.15	0.23	0.47	0.30	1.00	0.00	0.00	0.25	0.60	0.42
Cust8	0.97	0.02	0.01	0.69	0.26	0.05	1.00	0.00	0.00	0.24	0.44	0.49
Cust9	0.94	0.04	0.02	0.66	0.27	0.07	1.00	0.00	0.00	0.23	0.40	0.60
Cust10	0.89	0.07	0.04	0.55	0.35	0.1	1.00	0.00	0.00	0.25	0.32	0.48