[1]Lulu Yu

The Application of K-means Clustering Algorithm in the Evaluation of E-Commerce Websites

JES Journal of Electrical Systems

Abstract: - When dealing with large amounts of high-dimensional transaction data, clustering approaches often struggle with challenges including elasticity, weak processing capabilities of high-dimensional data, sensitivity to data sequence across time, independence of parameters, and ability to manage noise. These problems may prevent the methods from providing accurate predictions. Experiments conducted with data samples collected from 300 different mobile phones purchased on Taobao yielded the following results. K-means beats Single-pass in evaluating e-commerce transactions because of its higher intra-class dissimilarity and inter-class similarity. K-means clustering is an approach to the organization of massive datasets that is both effective and flexible. The outcome of the clustering algorithm is sensitive not only to the total number of clusters but also to how they were initially arranged. Because of this, it is simple to demonstrate that the results of clustering are best optimized locally. For this reason, continuing research into the elements that influence the number of clusters produced by this method as well as the starting locations of the clustering center is a crucial endeavor.

Keywords: K-Means Clustering; E-Commerce; Clustering Analysis

1. Introduction

E-commerce has been experiencing a meteoric rise in prevalence as a direct result of the rapid development of online and computer network technology. E-commerce, at its core, is a business strategy that enables the trading of products and services, as well as the collecting of rent and other payments, over the Internet. E-commerce was initially developed by the Japanese in the 1990s. As a direct consequence of this, routine purchases made online generate a mountain of transaction data. During the course of a transaction, the server is able to automatically collect and store this information in a transaction database so that it can be accessed at a later time. As a direct consequence of this development, normal company operations can continue unabated. Access to and analysis of a customer's whole purchase history can be consolidated into a single record in the transaction database for the purpose of streamlining the process (1). Additional information on clients can be stored in the record of purchases, which has a large amount of available space for this purpose. The use of data mining techniques in the transaction database has the potential to discover substantial economic value that has not yet been realized. For instance, we may analyze the purchasing patterns of individuals by determining the connections between the numerous products that they transport from the store to their homes. This enables store owners to invest more money into advertising the products that their clients are truly interested in purchasing (2,3). If we are able to centralize the things that are most frequently exchanged, we will be able to uncover the linkages that exist between these items, which will allow sellers to engage in targeted marketing. In the meanwhile, if groups of customers who have similar preferences can be identified, adverts can be crafted specifically for each distinct grouping of consumers. There is the potential for increased revenues from product sales (4,5). In recent years, clustering strategies like level-based, increment-based, density-based, and grid-based techniques have been increasingly popular in the field of research on online shopping. On the other hand, the fact that it does not take into consideration all of the associations that already exist between the newly produced items is likely the cause of its poor clustering efficiency. On the other hand, density-based clustering is able to make use of sampling technology and partition in order to effectively manage the vast datasets. Using sampling technology, this technique is effective for managing high-dimensional data; but there will always be some degree of sample error (6,7). Every day, it creates reams of transaction data, which is enough information to cram hundreds of records into a database. Because of the difficulties involved in transitioning from a transaction database to a high-dimensional database, the vast majority of clustering techniques are completely useless (8,9).

[1] Faculty of Business,City University of Macau, Macau 999078,China

School of Economics and Management, Xi 'an Mingde Institute of Technology, Xi' an, Shaanxi, 710124, China;

E-mail: rainbowyll@126.com (corresponding author).

2. Literature Review:

It is impossible to use algorithms effectively (10–12). The invention of the wavelet probably occurred in the 1930s. Despite their seemingly unconnected topics, several concepts have been offered that are eerily similar to one another. Morlet and Grossmann did not officially name the wavelet until 1984. In 1985, Mallat discovered the theoretical similarities, which is also how wavelet theory received its many uses (13,14). A fundamental unit of Wavelet theory is the mother wavelet. It must fulfil certain requirements before it can act as the model wavelet. The zero-frequency component of the mother wavelet is the physical meaning of the previous two equations. In other words, the mother wavelet's average on time must be zero, and it must have a shock wave (15,16).

To develop the wavelet further, we begin with the mother wavelet. A new function is formed that takes on the values of the equation's parameters when they are changed. The function that first defined wavelets is known as the mother wavelet function. There is no overlap between the two modes of operation. Therefore, by modifying the scale factors, we can obtain wavelets of varying frequencies. In addition, the translation factors, represented by b in the equation, are considered. Filling the time axis with wavelets is necessary for time domain analysis using the wavelet transform. We may now do studies of data in the time domain. By translating the centres of all the wavelet functions away from the time axis zero in either direction with the translation factor b, the desired function shifts can be implemented. The basic component of a wavelet transform is the utilization of the idea of a wavelet function. There are many connections between many small pieces of the original function or signal. Disintegration of the wavelet may lead to a gradual weakening of its strength, as well as pieces with different scales. Wavelet transformation or decomposition was the initial name for this method of decomposition (17–19). Wavelet reversal, also known as wavelet reconstruction, is the process of reassembling the original signal from its component waves.

The aforementioned clustering approach has been criticised for its lack of flexibility. Issues including weak independence of parameters, insufficient ability to handle noise, and sensitivity to the temporal sequence of data are common when working with massive amounts of high-dimensional transaction data. This algorithm outperforms its predecessors in dealing with data in high dimensions. K-means clustering is a flexible and economical algorithm that maintains its effectiveness even when the number of data points increases. This technique will look for k-division as part of its clustering process. This division yields the most accurate square error function calculation. Clustering effects are substantial when the final group is concentrated and the disparities between the groups are clear [13-16].

First, we discover the centres of the K initialised groups, then we move the objects to those centres, and finally, we merge the groups back together. Last but not least, repeat the preceding procedures an unlimited number of times until convergence is achieved [17]. To draw these conclusions, run some tests in which you collect data from all 300 of Taobao's mobile devices. K-means clustering is the best approach for high-dimensional data. When there are distinct differences between groups, it is also capable of demonstrating great clustering effects, expressing very high efficiency and considerable elasticity.

3. Methodology:

3.1. Data Collection

Taobao is now one of the most well-known online marketplaces in the world, having swiftly spread to new countries since its inception. As a result, researchers are free to use Taobao as a data collection tool. We look at data collected from purchases made on mobile phones and tablets. Three hundred different mobile phone products were used as samples for this research [19].

The K-means approach is very effective at clustering numerical values. Data preprocessing, as described in this paper, is performed before the experiment to transform qualitative information into quantitative figures. For instance, the original seller registration information is in the format yyyy-mm-dd. After that, you may figure out how many days are necessary to register by deducting the total time spent registering from the seller's initial registration time. When referring to the type of phone, a value of 1 indicates a smart phone, whereas a number of 0 indicates a non-smart phone. Since the data in the brand column is purely textual, we can display it with tags rather than the attribute's actual name. The processed data is tallied and presented in [11] table.

3.2. Neural Network

In wavelet neural networks, the wavelet transform is utilised to express the functions of network neurons. This brings together the concepts of wavelet decomposition and neural networks in a novel network architecture.

In wavelet theory, an instrument known as a projection quantity is utilised for the purpose of describing functions that are defined on L2(R). Additionally, the projection quantity was found for the wavelets and the functions. The development of discrete wavelet decomposition is underpinned by the same fundamental idea. In addition to this, it selects the wavelet function that is able to carry out its operations in both the time domain and the frequency domain on a regular basis. They are able to give the wavelet function time-frequency localization qualities by carrying out the steps outlined above. Because wavelets have a time-frequency property, the intercepted discrete wavelet decomposition can be utilised to approximate any function. In most cases, it will select wavelets that have distributions in the time-frequency domain that meet the conditions that are appropriate for this approximation function. However, the type of wavelet function that is used in the training of neural networks differs according on the application of the wavelet neural network that is being considered. It is not limited to things that can be counted on the hands of a single person. As a result, wavelets can be utilised in a wide variety of contexts where judgement is necessary. The fundamental building block of a wavelet is known as a neuron.

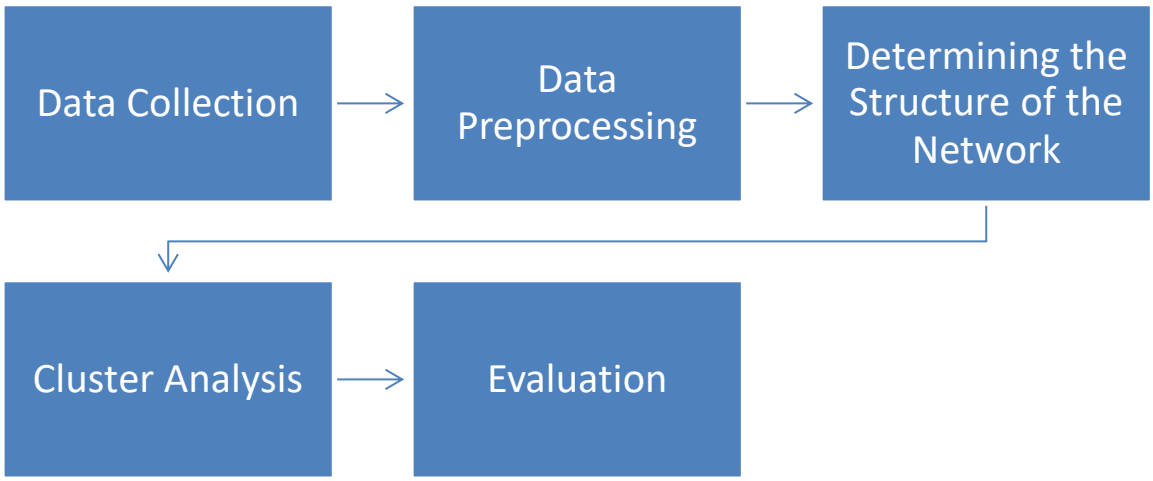

Fig. 1. The Proposed Approach's Procedure

Nonlinear regression and the parametric nonlinear function family are two applications of wavelet. The network might pick up these details during training as well. Back-propagation learning is commonly used in wavelet neural networks to optimize network parameters. Parameters representing the wavelet neural network's weights were selected when a best match was found for the input or output data. As a result, less room for the study's inaccuracy is created. The dilation (Di) and translation (-ti) operations of the wavelet operator produce the straight line in Fig. 2. A single neuron in a network is depicted by this line. The following equation describes the wavelet neural network:

$$\boldsymbol{f(x) = \sum_i w_i \alpha_i \varphi \left(\frac{x - t_i}{\alpha_i}\right) + C^T x + b}$$

(1)

Trace the wavelet back to its origin. The best space basis is usually taken into account. The most popular option is a radial, smooth wavelet that considers both the time domain and the frequency domain.

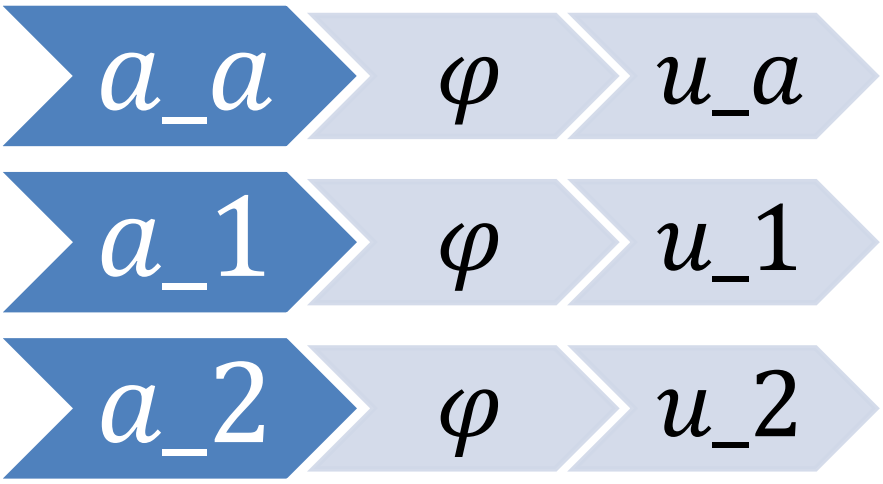

Fig. 2. Wavelet Neural Network

3.3. Clustering Analysis

As is the case with clustering, data items can be grouped together into clusters if they have a high degree of similarity among themselves but only a low degree of similarity with objects that belong to other groups. The concept of cluster analysis may be traced back to the beginnings of many different scientific subfields. An improved understanding of a company's clientele can be achieved through the application of cluster analysis, which divides customers into various groups based on the similarities they share. Clustering is a method that can be used to infer plant and animal classifications or to group genes with related functions. This helps researchers gain a deeper understanding of the underlying group structure. On the other hand, clustering is a method that may be utilised to classify different kinds of writing found on the internet.

The triple stands for the clustering characteristics, with N being the total number of text vectors that are part of the clustering class.

Table 1. Data Formats of E-Commerce Transactions of Mobile Phones

Brand	Sales	Total Sales	Price	Product Time	Screen Color	Smart	After Sale
Samsung	12	15	2500	150	50	1	0
Nokia	6	9	1922	100	1500	0	0
Sony	5	7	2010	50	670	0	2
Blackberry	4	6	3850	25	100	1	0

$$\bar{x} = LS/N \tag{2}$$

The augmented heuristic method known as K-means clustering has been in use ever since it was developed in the 1950s. The first stage in iterative enhanced heuristics is to partition the entering data points into initialised groups. The next step is to determine the mean of the entire group as a whole. The next step is to transport the things to the centres that are located in the most immediate proximity to them geographically. After that, start the process over from the beginning and keep going until you reach convergence. The K-means method uses a framework that is determined by the center-of-mass distance in order to classify different sets of data. When clustering, it is best to have a high degree of similarity across the groups being compared, but only a low degree of similarity inside each group being compared. The average value of an object inside a group can be used to indicate the degree to which the group is cohesive. The processing steps of the K-means clustering method are detailed in this paragraph. In the first step of this process, we take into account an overall average based on all of the categories by selecting K items at random. Each item is given a cluster to which it most closely belongs based on its initial distance from the mean value of the group. This distance is calculated using the data from all of the items. Calculations need to be made to

determine the new data set's average value. Keep doing what you've been doing up until the point where the criterion function converges. To summarise, the role entails the following responsibilities:

$$J_c(m) = \sum_{j=1}^{k} \sum_{x \in C} |Xi - Zj|^2 \quad (3)$$

Meanwhile, and both are multidimensional data sets. Figure 3 depicts the general strategy of the K-means algorithm and may be found here. The following must be remembered: one) the K-means clustering coefficient. The Kmeans value technique then sorts data by user, with the group centres indicated by the items' means. (2) The group's numbers, designated by K, serve as the input. There are n distinct items in the number set D. (3) The output is a collection of K distinct classes.

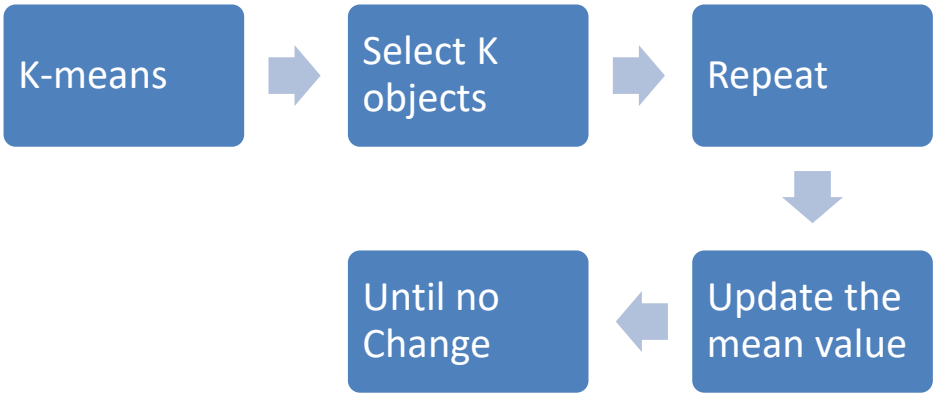

Fig. 3. K-Means Algorithm Flowchart

3.4. Experimental Analysis

We designed a battery of numerical tests to put our proposed approach to assessing e-commerce transactions through its paces. The process of excessive experimentation is shown in Figure 1.

3.5. Obtaining the Data Source

Taobao is China's largest online shopping platform, with almost 500 million users. There are more than 800 million everyday items available for buy online. The Taobao data used in this study is very typical. We ran our experiments with data collected from online purchases made from mobile devices, which largely includes the following three features. Price, ringtone, time to market, smart phone status, screen colour, manufacturer, and so on are just few of the mobile phone product attributes. The second type of consideration is seller characteristics such the seller's good feedback rating, shop registration date, seller credit, and after-sale support. The sales status of an item, including its current sales volume, sales stage, and cumulative sales status, is an example of the third characteristic. A total of 300 examples, one from each mobile phone product listing on Taobao, are used in the paper to illustrate the selection process [11].

3.5.1 Data pretreatment

The K-means method is very effective when it comes to the clustering of numerical data. The information is organised in advance in order to facilitate its conversion from textual to numerical form during the course of the experiment. The format of yyyy-mm-dd, for example, is frequently used to convey the seller's registration time because it is the format that is used to collect the original data. It is possible to determine the registration day by taking the whole time data and subtracting the time the vendor initially registered from that total. A conventional phone is represented by the value 0, whereas a smart phone is represented by the value 1. The first digit, "0," indicates a worldwide warranty, the second digit, "1," indicates three guarantees, and the third digit, "2," indicates all other types of phone after-sale service. Because brand is not a number, the column for phone brands can display the brands as tags rather than giving them a specific name for an attribute because brand is not a number. Table 1 [11] displays the outcomes of the processing that was performed.

3.5.2 Experiment

E-commerce transaction data from 300 distinct mobile devices can be clustered using the K-means method, which can then be used to locate the clustering centre and the object sets of clustering data. This is possible because of the design of the algorithm. It is possible to create new clusters based on data with objects selected at random. The next step is to sort the items into categories based on the average value of the categories. At long last, we get the chance to tweak the group average as the object average is calculated for each subgroup. The preceding steps must be repeated until the group size is steady. Figure 4 depicts the fundamental operations of a wavelet neural network during feature extraction. The following prerequisites are necessary to assure accuracy before proceeding:

To begin, let's look at the network's preferences. The problem of choosing how many wavelets to utilise can be approached using the typical model selection criterion employed in statistics. Such metrics can include, but are not limited to, Akaike's information criterion (AIC), the Schwarz bayesian information criterion (BIC), and the final prediction error criterion (FPEC). In most cases, overfitting can be avoided in sequence prediction by simply setting the wavelet to 1.

Parameters for starting the procedure, number two. For each wavelet, it is possible to ignore the training sample space's input sample size and scale level. When making a choice between the two, the rule of thumb is often used. The level of the scale was typically set at 4. It is common practise to use an input sample size that is equal to the number of variables plus two for every tiny wavelet.

The third limitation is the schooling system. The wavelet chosen by the network is dependent on its input and output data samples, as well as its initial parameters. Limits on the number of repetitions and conditions for stopping are examples of parameters. The first wavelet is then generated just after the regressive wavelet selection process is completed. Figure 4 depicts the flow of data through a wavelet neural network.

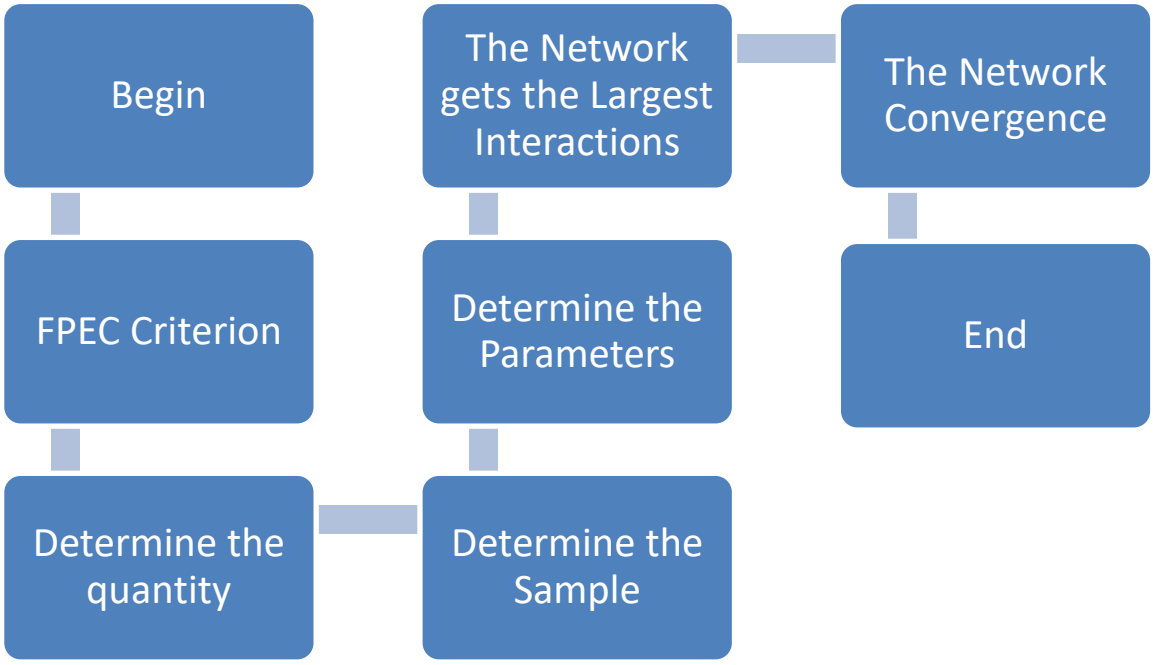

Fig. 4. Wavelet Neural Network

The importance of the order in which you present your models cannot be overstated. In the aforementioned formula, K represents the processing time delay, and na the number of input parameters to the clustering neural network. Both values for this parameter must be nonnegative, with 1 being the smallest acceptable value. The common prediction model can help with the level-selection problems. In terms of visual appeal, plots of autocorrelation and partial autocorrelation functions for subjective observations stand out. However, these functions are unnecessary for building a wavelet neural network. Prediction model; for each layer, we employ a unique method to establish

the loss function. Then, we determined the best possible value for na by applying Akaike's Final Prediction Error Criterion (FPEC).

4. Result and Discussions

In this investigation, we use data collected from 300 different phones that were purchased on Taobao as our experimental participants. Since its inception, Taobao has experienced phenomenal growth, emerging as not only one of the world's electronic trading platforms but also one of the most significant platforms for conducting e-commerce transactions. As a consequence of this, the information that can be found on this website will most likely be included in the sample set that is utilized for the study. In the meantime, the data on internet phone sales may be segmented into three primary groups for the purpose of conducting experiments. These categories are the product attribute, the seller attribute, and the sale status. The impact of clustering is measured by comparing the average dissimilarity between classes and the average similarity within classes. The following are two instances of these different parameters:

$$H = -\frac{1}{nlogc}\sum_{i=1}^{K}\sum_{i=1}^{K} uijloguij \tag{4}$$

where H is a measure that indicates the degree to which different classes are related to one another. The capacity to discern between the two groups becomes more difficult as k gets smaller, and the opposite is also true.

$$F = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n} u_i j^2 \tag{5}$$

Various values of K for the input group size and the dataset D were tested, and the outcomes were compared to those obtained by employing the Single-pass clustering method. Each clustering strategy is tested for 20 rounds, and then the 10 results with the best clustering effects are selected. The findings are shown in Tables 3 and 4. The data was analyzed in the simulation using K-means clustering and a single-pass approach. The quality of this literature was determined by how different its classes were from one another and how similar its members were to one another.

To determine whether or not the proposed approach is useful for studying e-commerce transactions, we conduct an experiment. We utilise the K-means technique to evaluate the data and a wavelet neural network to extract features for the experiment. Table 2 summarises the findings. K-means clustering yields outcomes that are more distinct to one another and more similar to one another. Class similarity has been measured and found to vary from 63.48 percent to 987.1 percent. The 65% cutoff marks the divide between the two classes. Therefore, it is reasonable to conclude that K-means clustering is applicable to the study of e-commerce transaction data. All three of these factors may be responsible for the observed phenomenon. The K-means clustering technique may efficiently handle high-dimensional data and large datasets by overcoming their inherent limitations in terms of scalability and performance. But when confronted with a mountain of data on online purchases, things change. Data linked to e-commerce transactions, which follow a distinct temporal order, can be satisfactorily clustered using the K-means clustering method because this method is insensitive to the order in which the data is given. Finally, this will mitigate the influence introduced by the data's chronological order [6] when analysing online purchases. This is because K-means clustering relies on a large number of strongly independent parameters.

Table 2. K-mean Clustering for 10 times

Running Times	Dissimilarity	Similarity
1	0.6537	0.6544
2	0.7687	0.7654
3	0.3645	0.8655
4	0.5785	0.7585
5	0.7287	0.6758
6	0.7565	0.5656
7	0.6443	0.7657
8	0.5436	0.8655
9	0.6543	0.8675

10	0.5356	0.7658

In the second experiment, we extract features with a wavelet neural network and use the Kmeans method to analyse the data. We cluster the data using the single-pass clustering technique, which we utilise to perform the experiment. Figure 1 illustrates the comprehensive experimental approach that was taken. The results of this experiment demonstrate that the method in question is effective. A single pass is all that is required to complete the clustering. Figure 1 illustrates the comprehensive experimental approach that was taken. Following ten iterations of putting the clustering algorithms through their paces, we chose the top ten findings based on how effectively they clustered the data, as shown in Table 3. The experimental parameters consist of the critical distance, denoted by d, and the cluster number, denoted by K. We determine that the average degree of proximity is 56.28 percent. The average degree of dissimilarity, which comes in at just 24.29%, is likewise significantly lower than it was in Table 3. In comparison to the Single-pass method, the K-means method performed significantly better when working with the data from online shopping. These are useful for not only one but also two different reasons. This is one of the defining features of this type of data. The second problem is that it is difficult to learn what values should be plugged into the clustering process. This is a significant obstacle. The technique has a problem with a lack of independent parameter control.

Table 3. Single Class Algorithm for 10 Times

Running Times	**Dissimilarity**	**Similarity**
1	0.6587	0.6424
2	0.7883	0.7954
3	0.3445	0.7855
4	0.5535	0.4985
5	0.7127	0.5758
6	0.6565	0.5356
7	0.5443	0.6657
8	0.4836	0.7655
9	0.6143	0.5675
Mean	0.6356	0.7858

The last test compared the suggested method to existing methods in terms of how long it takes to examine e-commerce transactions. Over-experimentation, depicted in Figure 1, entails extracting features with a wavelet neural network and evaluating data with a K-means algorithm. Figure 1 illustrates the procedure. In Section C, we go into depth into the parameters. In Table 4, we see the average amount of time spent training for each of the 20 rounds. Using the results of this experiment, we can assess how well the method described in this paper works. Twenty such instances occurred! The number of clusters (represented by K) and their separation distance (represented by r) are the experimental parameters. The experiment's outcomes are tabulated in Table 4. In Table 4, T represents the entire amount of training time. The average duration of a training session is 12.38 and 8.54 seconds, the research shows. This represents a 4 second decrease in the indicated training time. The K-means algorithm's reliance on input parameters and its indifference to the passage of time are primary contributors to these results. This technique not only clusters the data, but also searches for the K divisions that yield the smallest value for the square error function, ultimately leading to the best answer.

Table 4. Result of the Training Time

Running Time	**Single-Pass**	**K-means**
1	14.54	12.01
2	10.34	10.54
3	12.64	6.65
4	11.54	6.54
5	12.43	7.54
6	11.53	8.56
7	10.46	7.43

Both the business landscape and consumer shopping habits have been dramatically altered by the stratospheric rise of internet retail. As more people turn to online marketplaces like Taobao to fulfil their buying demands, the necessity of e-commerce site success study and evaluation methodology increases. K-means clustering is a very new technique developed for this very purpose.

In light of the research described in the introduction, the K-means clustering method is highlighted as a useful method for assessing data related to e-commerce transactions. In this analysis, clusters are generated using an algorithm that considers data on items, vendors, and transactions. This methodology is helpful for the evaluation of e-commerce websites because it allows researchers and analysts to reveal hidden relationships and patterns in the data.

When high-dimensional data is included, the K-means algorithm excels where the single-pass approach fails. The results make this quite clear. Because doing business online generates such massive amounts of data that needs to be processed and analysed quickly, this revelation is ground-breaking. The K-means algorithm is an option worth considering when doing an examination of e-commerce websites due to its impressive ability to group and categorise data points based on similarities. Given the frequent need to segment markets or classify goods, this is crucial information to have at one's disposal. The data points can be intuitively sorted into relevant groups.

In addition to shedding light on the efficacy of the K-means algorithm, this research also reveals how effectively it can adjust to novel conditions. The technique's adaptability and scalability make it a promising tool for processing large data sets. Particularly useful in e-commerce, where website data can vary considerably in product variety, consumer behaviour, and sales patterns, the K-means algorithm's flexibility ensures that meaningful clusters may be produced despite being confronted with complex and varied datasets. This is especially useful in the realm of online commerce, where data can show a wide range of product varieties, customer habits, and sales trends. This is doable because the seemingly infinite number of options can be sorted into meaningful groups.

This finding is more consequential than simply increasing algorithms' efficiency. Understanding the interplay between similarity and difference among internet shoppers is made easier by this study. Companies looking to improve their online presence, product offerings, and marketing tactics will find this data to be of the utmost value. E-commerce enterprises can improve user experience, sales, and customer satisfaction by gaining a deeper insight of the interconnections across different groups and the degree to which their members are similar to one another.

If the e-commerce industry were to implement the K-means clustering algorithm, it might significantly alter consumers' traditional patterns of behaviour when it comes to making purchases over the internet. You can picture a brick-and-mortar company with an online component utilizing this algorithm to divide its customer base into smaller subsets so that it may send more relevant ads to each subset. With a detailed customer database, a store may cater its product recommendations, promotions, and more to each individual customer. The result has been an increase in sales and interest from customers.

The K-means algorithm excels in analyzing e-commerce transactions because it disregards the passage of time and is not excessively concerned with the sequence in which data items are presented. This is more important than ever in the modern era, as people's preferences are always changing. Merchants may benefit from this attribute if they were able to track their consumers' shopping habits and provide recommendations based on that data. This allows companies to maintain their products current and competitive in a rapidly evolving industry.

By comparing the K-means algorithm to the standard approaches, the study emphasizes the significance of maintaining a state of flux in the ways in which e-commerce data is analysed. The preferences of modern Internet consumers are always shifting, so it is imperative that organisations employ flexible approaches. Because of its superiority in training time and accuracy, the study's findings support the use of the K-means algorithm and encourage additional investigation into its potential.

This study explains how companies may utilize data analytics to stay ahead of the competition as the proliferation of online shopping transforms economies around the world. Strategic examination of e-commerce sites using K-means clustering can aid decision-making, consumer engagement, and the growth of the online marketplace. This novel strategy aids businesses in adapting to the changing digital economy and better serving their clients.

5. Conclusion

Low elasticity, poor handling of high-dimensional data, sensitivity to temporal sequence, lack of parameter independence, and ineffective noise management are only some of the issues that arise when working with a huge amount of high-dimensional transaction data using traditional clustering techniques. In this study, we introduce K-means clustering with a partitioning methodology. Data is divided into K initialised subsets, centres are determined for each, and objects are rearranged so that their centres are as close as feasible to their own; this allows us to reconstruct the original groups. After that, repeat the stages that came before it until you reach a point of convergence. K-means clustering algorithm is more efficient than conventional clustering algorithms, has superior performance when dealing with high-dimensional data, and exhibits high levels of intra-class dissimilarity and inter-class similarity when assessing e-commerce transactions. If you conduct When working with a large dataset, the K-means clustering technique is not only incredibly efficient but also very versatile. Both of these qualities contribute to its widespread use.

Acknowledgments

An Analysis on the Development of Shaanxi New Energy Industry, Item No. 2022HZ0506

References

[1] Deng Y, Gao Q. RETRACTED ARTICLE: A study on e-commerce customer segmentation management based on improved K-means algorithm. Inf Syst E-Bus Manag. 2020 Dec 1;18(4):497–510.

[2] Mathivanan NMN, Ghani NAMd, Janor RM. Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products. In: 2019 IEEE Conference on Big Data and Analytics (ICBDA). 2019. p. 1–4.

[3] Rahardja U, Hariguna T, Baihaqi WM. Opinion Mining on E-Commerce Data Using Sentiment Analysis and K-Medoid Clustering. In: 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media). 2019. p. 168–70.

[4] Agrawal A, Kaur P, Singh M. Customer Segmentation Model using K-means Clustering on E-commerce. In: 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). 2023. p. 1–6.

[5] Punhani R, Arora VPS, Sabitha S, Kumar Shukla V. Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce. In: 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). 2021. p. 149–54.

[6] Yadav V, Shukla R, Tripathi A, Maurya A. A New Approach for Movie Recommender System using K-means Clustering and PCA. J Sci Ind Res. 2021 Nov 2;80(02):159–65.

[7] Wang W. Application of E-Commerce Recommendation Algorithm in Consumer Preference Prediction. J Cases Inf Technol JCIT. 2022 Feb 21;24(5):1–28.

[8] Sarker KU, Saqib M, Hasan R, Mahmood S, Hussain S, Abbas A, et al. A Ranking Learning Model by K-Means Clustering Technique for Web Scraped Movie Data. Computers. 2022 Nov;11(11):158. Xiahou X, Harada Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. J Theor Appl Electron Commer Res. 2022 Jun;17(2):458–75.

[9] Singh H, Kaur P. An Effective Clustering-Based Web Page Recommendation Framework for E-Commerce Websites. SN Comput Sci. 2021 Jun 15;2(4):339.

[10] Valdiviezo-Diaz P. Partitional clustering based on PCA method for segmentation of products. In: 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). 2021. p. 1–4.

[11] Kumar MS, Prabhu J. A hybrid model collaborative movie recommendation system using K-means clustering with ant colony optimisation. Int J Internet Technol Secur Trans. 2020 Jan;10(3):337–54.

[12] Hariguna T, Baihaqi WM, Nurwanti A. Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier Algorithm. Int J Inform Inf Syst. 2019 Sep 1;2(2):48–55.

[13] Raja DRK, Kumar GH, Basha SM, Ahmed and ST. Recommendations based on Integrated Matrix Time Decomposition and Clustering Optimization. Int J Perform Eng. 2022 Apr 30;18(4):298.

[14] Arul V, Kumar A, Agarwal A. Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering. In: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). 2021. p. 1602–4.

[15] Pan H, Yang X. Fast clustering algorithm of commodity association big data sparse network. Int J Syst Assur Eng Manag. 2021 Aug 1;12(4):667–74.

[16] Mulyawan B, Christanti MV, Wenas R. Recommendation Product Based on Customer Categorization with K-Means Clustering Method. IOP Conf Ser Mater Sci Eng. 2019 Apr;508(1):012123.

[17] Rajput L, Singh SN. Customer Segmentation of E-commerce data using K-means Clustering Algorithm. In: 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence). 2023. p. 658–64.

[18] Kumaresh S, Haran R, Jarret MM. Analytics of e-Commerce Platforms Based on User-Experience (UX). In: Peng SL, Hsieh SY, Gopalakrishnan S, Duraisamy B, editors. Intelligent Computing and Innovation on Data Science. Singapore: Springer Nature; 2021. p. 309–18. (Lecture Notes in Networks and Systems).

Lulu Yu, Female, Han, Place of birth: Xi'an, Date of birth: April 1984, Education: Graduate, Degree: Master, Title: Associate Professor, Research Interests: Practical Economics.