Information Systems and e-Business Management (2020) 18:497–510
https://doi.org/10.1007/s10257-018-0381-3

ORIGINAL ARTICLE

A study on e-commerce customer segmentation management based on improved K-means algorithm

Yulin Deng[1] · **Qianying Gao**[1]

Received: 12 October 2018 / Revised: 7 November 2018 / Accepted: 14 November 2018 /
Published online: 3 December 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

With the continuous popularization of the network, the customer resources have to be valued if enterprises want to occupy a certain share in the field of e-commerce. However, the traditional clustering analysis method has obvious lag for the segmentation of e-commerce customers. Therefore, accurate and efficient customer segmentation management should be carried out for the large and complex data information of current e-commerce enterprises, so as to realize customer retention and potential customer mining and promote the efficient development of enterprises. On the basis of customer segmentation theory, for the shortcomings of traditional K-means algorithm, a new SAPK + K-means algorithm based on semi-supervised Affinity Propagation combined with classic K-means algorithm is proposed in combination with AP algorithm, which is applied to e-commerce customers for segmentation management. The results show that when the SAPK + K-means algorithm clusters the iris dataset and the ionosphere dataset, the clustering time is longer than the K-means algorithm and the AP algorithm, but the algorithm error rate in the standard data is significantly reduced and the correct number of clusters can be obtained. The main steps of SAPK + K-means algorithm applied to customer segmentation management including data acquisition, cluster analysis and analysis and evaluation of clustering results. The SAPK + K-means algorithm clusters the data information of an e-commerce customer to obtain four different customer types and proposes corresponding strategies for each type of customer. It is concluded that the SAPK + k-means algorithm can significantly improve the clustering quality of customer data information and improve the effectiveness of activities of e-commerce enterprises.

Keywords K-means algorithm · SAPK + K-means algorithm · e-commerce · Customer segmentation management

✉ Qianying Gao
gqygjj@sina.com

Yulin Deng
yulin_d@hhu.edu.cn

[1] Business School of HoHai University, Nanjing, China

1 Introduction

With the rapid development of computer network technology, the arrival of the information age has caused a huge change in the way of market competition. The internet-based business model is not only faster and more convenient in terms of time and space, but also, to a large extent, provides efficient performance for enterprises to acquire customer resources and market information (Kuo et al. 2016). In the face of product competition, enterprises should not only consider the ever-changing local market, but also rationally and effectively use and mine customer resources to achieve targeted measures for different customers (De Roos et al. 2009). The key to enterprise development is to start with the analysis of customers' needs and use customer segmentation as the means to mine and analyze various consumer groups in the system, so as to provide different types of customers with distinctive marketing methods and improve their satisfaction and loyalty to maintain the core competitiveness of the market (Hiziroglu 2013). With the continuous enrichment of database resources, customer characteristics show a trend of diversification. The traditional method of customer segmentation is relatively simple and rough and can't well cater to the market business model (Yao et al. 2014). Therefore, choosing the appropriate data mining algorithm can effectively extract the effective characteristics of consumer behaviors for enterprises and institutions, use these characteristics to realize the division of different categories of consumer groups, finally evaluate the contribution of different customers to enterprises, and also facilitate the company to specify differentiated marketing programs.

Cluster analysis is a kind of algorithm frequently used in data mining technology, which is mainly used in the analysis of enterprise data information to observe the distribution characteristics existing in data sets, so as to achieve differentiated interpretation of different clusters (Oña et al. 2016). Among them, K-means algorithm is an optimization algorithm, which minimizes an objective function as an optimization criterion and chooses an optimal combination as the optimal clustering scheme. Although the K-means algorithm has a wide range of applications in helping telecom operators implement customer segmentation and accurately locate customers' market needs, there are still deviations in the analysis results (Luo et al. 2013). In the operation process of the traditional K-means algorithm, the number of clusters and the selection of the initial center point have a great influence on the clustering effect, and it may also result in a local minimum until convergence or a predefined number is reached. The stages of reducing the value of the objective function include the optimal degree of membership, or when the degree of membership is fixed. The continuous updating and improvement of the selection of cluster centers will inevitably increase the complexity of clustering algorithms. At the same time, due to the sensitivity of clustering results to the selection of cluster centers, there is no guarantee that the final clustering results will achieve satisfactory results. If the selection of cluster centers deviates from the set, large errors will occur (Zhang and Ma 2017).

Therefore, the current k-means algorithm is improved to explore the diversity of demands within the overall customer, so that companies can obtain more

sufficient customer information resources, obtain potential knowledge of customers, achieve favorable position in marketing and improve customer relationship management level, and finally provide favorable prerequisites for maintaining a leading position in the commercial battlefield.

2 Methodology

2.1 Theory related to customer segmentation

Since Wendell r. Smith, an American marketer, first proposed the concept of "segmentation", numerous experts and scholars have been attracted to study it. Segmentation, also known as customer segmentation, refers to the process of dividing a market into different buyers with different behaviors, characteristics, or needs (Hoegele et al. 2016). According to different times, markets and industries, the researchers proposed different market segmentation concepts: product-oriented segmentation and customer-oriented segmentation. Customer segmentation refers to a way of dividing according to different characteristics of consumer groups. It is an important part of customer relationship management and has gradually become an important prerequisite for enterprise to apply customer relationship management. The theoretical basis is that, in a clear strategic business model and a specific market, different types of market groups will, based on their actual conditions, decide whether to purchase a certain type of product in the market. Because of the demand of appearance, the attention of service level, or the inconsistency of the way the product realizes value, different consumer groups will give different consumption intentions. This theory proposes to study and predict the future consumption trend of customers in the way of segmentation of customer information and consumption behavior, as well as the profit market planning of enterprises, so as to achieve the goal of reasonable allocation of service resources and the most profitable design of customer marketing programs (Huang et al. 2014). The continuous development and improvement of Internet consumption methods can't be separated from the support of customer segmentation technology. The enterprise's cognition and trend analysis of customers' consumption habits play an important role in the competition between market positioning, marketing plans and competitors. The basic flow chart of customer segmentation is shown in Fig. 1.

2.2 K-means algorithm

K-means clustering algorithm is one of the clustering algorithms based on division. It adopts a heuristic iterative process to re-divide data objects and re-update cluster centers. The basic idea of the algorithm is: suppose a set with element objects and the number of clusters to be generated. In the first round, a sample element is randomly selected as the initial cluster center, and the distance between other sample elements and the center point is analyzed, and the clusters are respectively divided according to the distance. In each of the following

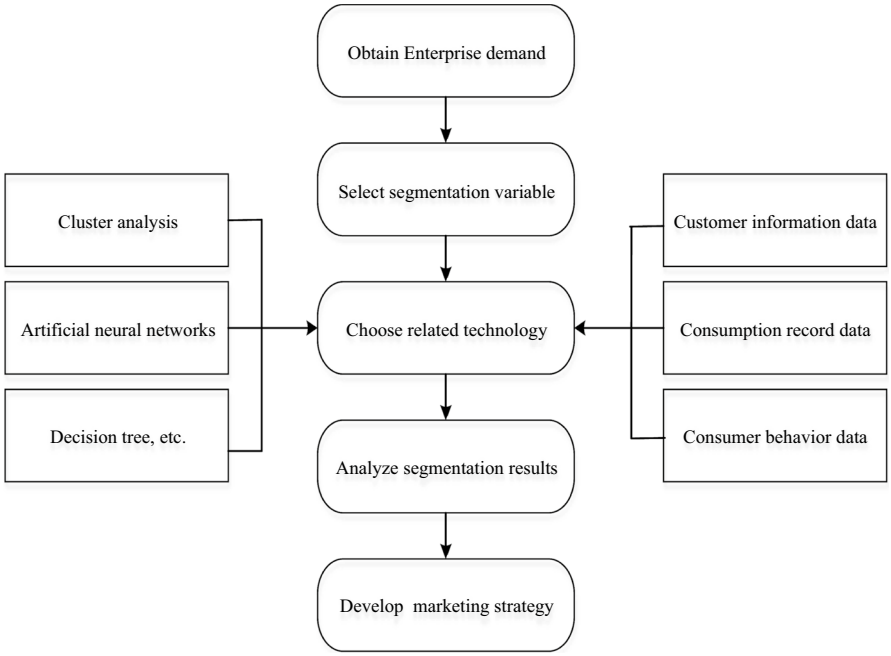

Fig. 1 Basic process of customer segmentation

rounds, the iterative operation of the above steps is continuously performed, and the average value of the element objects obtained this time is taken as the center point of the next round of clustering until the condition that the clustering center point no longer changes in the iteration process is met (Arora et al. 2016). The specific processing steps are as follows:

Given a data set D containing n data; given k clusters and k initial cluster centers $Z_f(I), j = 1, 2, \ldots k$;

Calculate the distance of each data object to the cluster center, $D(x_i, Z_j(I))$, $i = 1, 2, \ldots n$; $j = 2, \ldots, k$, if satisfied:

$$D\left(x_i, Z_k(I)\right) = \min\left\{D\left(x_i, Z_j(I)\right), j = 1, 2, 3 \ldots n\right\}, x_i \in w_k. \tag{1}$$

Determine a new cluster center; calculate the error square sum criterion function J;

$$J_c(I) = \sum_{j=1}^{k} \sum_{k=1}^{n_j} \left\| x_k^{(j)} - Z_j(I) \right\|^2. \tag{2}$$

Judge, loop 2 and 3, until each cluster no longer changes, i.e. the minimum, stops the algorithm;

Output k cluster sets.

In the iterative process of step 2 and step 3, there are two methods to modify the initial clustering center: batch modification method and individual modification method. The method of batch modification is to modify the clustering center after all data objects are classified; individual modification method means that each time an object changes its classification, it starts calculating the mean of the two classes involved and modifies the clustering center. According to the above description, it is obvious that the method of batch modification is small in calculation amount and the clustering speed is fast, but the clustering result has a large dependence on the initial cluster center; the clustering results of the individual modification method are related to the order in which the data objects are classified, so it is necessary to select representative objects as the clustering center. When using the method of individual modification, the number of clusters, the minimum distance between classes and the maximum distance within classes should be changed frequently to achieve multiple clustering and improve the clustering effect.

2.3 AP algorithm

AP algorithm is a clustering algorithm based on the similarity between N data samples. Its ultimate goal is to find the set of largest class centers with the sum of similarity by calculating the sum of the similarities of all data samples to the class center, that is, the optimal clustering result (Serdah and Ashour 2016). The AP algorithm doesn't need to give the initial cluster center or the number of clusters first, but treats all samples as potential cluster centers, called exemplar; it also establishes attractiveness information (that is, the similarity between any two data samples) for each data sample with other data samples and stores it in the similarity matrix S. In the AP algorithm, s (i, k) is used to indicate the probability of x_k as the clustering center of x_i. The value s (k, k) on the diagonal of the S matrix is taken as the evaluation criterion of whether x_k can become the clustering center, which indicates that the larger the value of s (k, k) is, the greater the possibility of this point becoming the clustering center is. This value is called the p (preference) in AP algorithm. The size of p affects the number of clusters: if each data sample can be used as a cluster center, then p takes the same value; if the mean value of the similarity of the data samples is taken as the value of p, a medium number of clusters can be obtained; if the minimum value of the similarity of the data samples is taken, fewer clusters can be obtained. At the same time, two important information are conveyed in the AP algorithm: r (responsibility) and a (availability). Where, r (i, k) represents the numerical message sent from point x_i to the candidate cluster center x_k, reflecting the suitability of point x_k as the clustering center of point x_i. a (i, k) represents the numerical message sent from the candidate clustering center point x_k to point x_i, reflecting the suitability of point x_i to select point x_k as its clustering center. The greater the sum of r (i, k) and a (i, k), the greater the possibility that point x_k can become clustering centers, and the greater the possibility that point x_i can belong to classes with point x_k as clustering centers. The iterative process of AP algorithm is the process of continuous update of each point's attraction and attribution value until

m high-quality cluster centers (exemplar) are produced, and then the remaining data samples are distributed to corresponding classes to complete the iterative process.

$$S(i,j) = -\left\|(x_i - x_j)^2 + (y_i - y_j)^2\right\|. \tag{3}$$

$$r(i,k) = s(i,k) - \max\{a(i,j) + a(i,j)\}, i \in \{1,2,\ldots N, j \neq k\}. \tag{4}$$

$$a(i,k) = \min\left\{0, r(k,k) + \sum_j \{\max(0, r(j,k))\}\right\}, j \in \{1,2,\ldots N, j \neq k \quad \text{and} \quad j \neq i\}. \tag{5}$$

The AP algorithm doesn't need to consider the problem of clustering center and clustering number, because all data samples can be regarded as potential clustering center, and the similarity between any two data samples is stored in the similarity matrix S.

2.4 SAPK + K-means algorithm

Both AP algorithm and k-means algorithm are algorithms implemented on the basis of k-center clustering. In the operation process of the classic k-means algorithm, the number of clusters and the selection of the initial center point have a great impact on the clustering effect, which may also lead to the local minimum value. Therefore, in order to achieve a more ideal clustering result, several different initial values should be given to implement the algorithm. However, if the data set needs to generate a large number of clusters, it will have little effect. Therefore, a new algorithm, SAPK + K-means, is proposed by combining the AP algorithm based on semi-supervised learning with the K-means algorithm. In the process of implementation, the improved algorithm constantly searches for the best center value of the cluster, and realizes the maximization of the objective function value, so as to obtain the best clustering effect; the SAPK + K-means algorithm realizes the initialization of the K-means algorithm, so its square error is smaller than AP algorithm. The algorithm of SAPK + K-means only needs one iteration to finish the operation, which can achieve better clustering effect for the number of elements and obtain satisfactory results. The specific operation steps of SAPK + K-means algorithm are as follows:

Enter the data set and initialize the parameters.

Data set is $\{x_1, x_2, \ldots x_N\}$; Bias parameter $p = p_m$; Falling step $step = p_{min}/10$; the technical mark H and the supervision mark HS were both zero; the convergence condition is that there is no change in the center circulation of the cluster for 30 times; the termination condition is that there is no change in the cluster center circulation for 300 times.

Run an iterative process.

Step 1 run the iterative process and obtain K clustering results. If HS = 1, go to step 3, otherwise proceed to the next step.

Step 2 check if the clustering result converges. If it converges, calculate *Sil*(*K*) and mark HS = 1, go to step 4; otherwise, go back to the previous step.

Step 3 check whether the clustering center meets the convergence condition. If it converges, obtain K clusters and calculate $Sil_{\max}$. If $Sil(K) < Sil(\mathrm{K}-1)$, then $\mathrm{H} = \mathrm{H}+1$; and if $Sil(K) > Sil_{\max}$, $\mathrm{H} = 0$.

Step 4 check whether $\mathrm{H} > \mathrm{Kl}/2$ and whether K is 2; and check whether the cycle number meets the termination condition. If it meets, go to step 5, otherwise go back to step 1.

Step 5 check the number of optimal clustering corresponding to $Sil_{\max}$, if it is 2, calculate *Hartigan* index and compare it.

Step 6 output the number of clustering and clustering center.

The the K-means algorithm is initialized by output cluster number K and the cluster center.

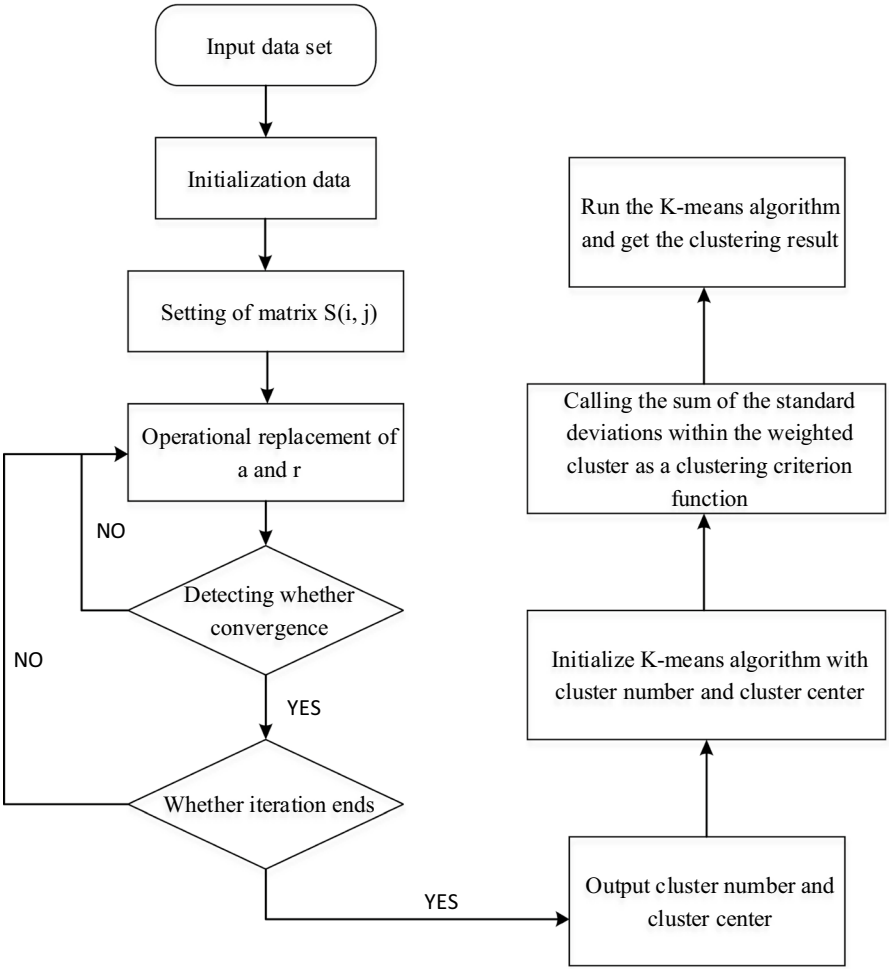

Fig. 2 Flow chart of the improved SAPK + K-means algorithm

Run the k-means algorithm to get the final clustering results. Draw the flowchart of SAPK + K-means algorithm, as shown in Fig. 2:

3 Results

3.1 Verification of algorithm

The improved SAPK + k-means algorithm is calculated by taking data set of iris and ionosphere as standard data, and the exact number of clusters obtained should be 4 and 3. As shown in Fig. 3, the time and error rate of a pair of standard data sets (iris, ionosphere) which adopt K-means algorithm, AP + K-means algorithm, and the improved algorithm SAPK + k-means algorithm are described. It can be seen from the analysis that, similar to the simulated data set, the SAPK + K-means algorithm improved in this study has a lower error rate in the clusters obtained from the two data sets, followed by the simple AP + K-means algorithm and the original K-means algorithm has the worst effect. However, in terms of time consumption, the improved SAPK + K-means algorithm in this study takes longer than the simple AP + K-means algorithm and the traditional K-means algorithm, which may be related to the clustering structure of the data set. Since the correct rate of the SAPK + K-means algorithm is high, the longer time consumed is reasonable.

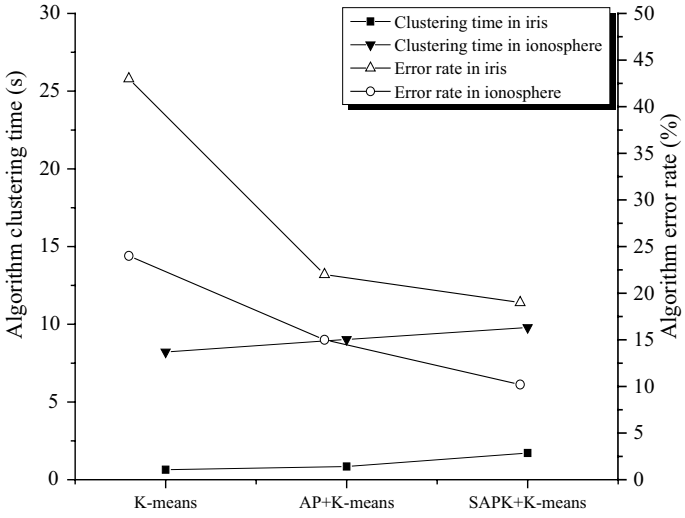

Fig. 3 Clustering comparison analysis of K-means algorithm, AP + K-means algorithm and SAPK + K-means algorithm

3.2 Steps of applying SAPK + K-means algorithm to e-commerce customer segmentation

Generally, according to the relevant information in the visitor log in the e-commerce website, the data is pre-processed first, then the relevant model is established, the customer is segmented by the clustering method, and the basis is provided for the enterprise to make decisions. The process of applying SAPK + K-means algorithm to customer segmentation is shown in Fig. 4. The specific steps are as follows:

Obtain the required data table from the related data table of the e-commerce website.

Determine whether the data table obtained has a clustering trend or not. If there is a clustering trend, conduct clustering; otherwise, cancel the following clustering steps.

Apply the SAPK + K-means algorithm to the acquired customer data set D, and divide the data set D into cl, c2, c3,… with clustering algorithm.

According to the data object characteristics in the class, summary each class into one or several rules corresponding to the characteristics of each class.

Evaluate the clustering results. If the clustering result is highly reliable, it is confirmed to be applied to the actual application; otherwise, the clustering analysis is performed again by other clustering algorithms.

3.3 Data acquisition of the SAPK + K-means algorithm applied to e-commerce customers

In this study, data information of a cosmetics e-commerce website is obtained, including customer information table, product information table and customer consumption information table. The relevant data of customers are recorded in Table 1, including customer number, gender, age, identity information, home address, etc.

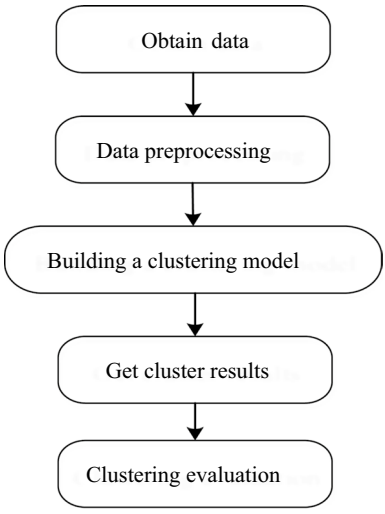

Fig. 4 The steps of the improved SAPK + K-means algorithm applied to e-commerce customer segmentation

Table 1 Customer information table

Attribute name	Data type	Length
Customer number	int	6
Gender	char	6
Age	int	2
Education background	char	6
Identity information	char	14
Home address	char	28

Table 2 Product information table

Attribute name	Data type	Length
Product number	int	4
Product name	char	20
Product type	char	8
Price	int	4

Table 3 Customer consumption information table

Attribute name	Data type	Length
Customer number	int	10
Product number	int	8
Product price	int	4
Consumption quantity	int	8
Total consumption	money	3

Relevant data of products are recorded in Table 2, including product number, product name, product type and price. The consumption data of consumers are recorded in Table 3, including customer number, product number, product price, consumption quantity and total consumption.

3.4 Data preprocessing of the SAPK + K-means algorithm applied to e-commerce customers

Processing of vacancy values. The information on the degree of education of customers has been largely missing, and it needs to be handled by manual processing. Since there is a lot of missing information, it is decided to discard this field. As for the field information of the total consumption of this time, the appropriate calculation method is selected to realize filling, and the information of the total consumption of this time is estimated based on the product price and consumption quantity.

Processing of noise data. As a part of data preprocessing, it is very important to confirm and process noise data. Noise data is different from normal and useful data information, which is based on randomness and deviation. It is mainly reflected in

the incomplete information record. For example, when the customer fills in the education level field, it is not always possible to fill in all the education experiences completely, or even not fill in at all, thus forming the potential noise data information; for the erroneous data, if the customer doesn't correctly fill in the age field, it exceeds the positive distribution threshold selected in the statistics = mean ± 2 × standard deviation, that is, all fields except the interval [12, 92] are regarded as wrong data; When repeated data occurs, simple fuzzy matching algorithm can be adopted, and angle similarity matching algorithm can be used to process.

Processing of inconsistent data. In the information table, there are some data information that customers fill out inadvertently or intentionally and don't match the format. For example, the customer fills in the gender information in the age field and adds the age information in the gender information field, but since this data is still rare, it can be solved by manual processing.

Convert variable names and their values in different databases to implement sample element conversion, as well as normalization processing, format conversion, and so on. To understand how long a customer has become a member of the site, it is calculated by subtracting the time of the last purchase from the time that the customer is registered as a member of the site, so that a derivative field that stores the most recent consumption time should be added. Finally, all the databases are integrated to obtain the final database set information, and new data information is finally run into the customer segmentation.

The data in Tables 1, 2 and 3 all describe the detailed information of consumers themselves and their consumption behaviour. For the goal of achieving the appropriate segmentation effect through the requirements of each division, it is required to extract the appropriate fields from the above table and generate the new table of required customer segmentation, as shown in Table 4.

3.5 E-commerce customer segmentation with SAPK + K-means algorithm

The improved new algorithm, namely SAPK + k-means algorithm, is applied in the customer segmentation process of the e-commerce website to obtain the data information of a cosmetics e-commerce website. 120 pieces of data are selected from them, in which the attribute variable selects the consumption number and average consumption amount of consumers during the specified period. After preprocessing the data, the algorithm is applied to the data information of the e-commerce website. The SAPK + K-means algorithm is adopted to segment the customers of e-commerce websites. After analysis, it is found that when k value is 4, a better clustering effect can be achieved. The segmentation results are shown in Fig. 5.

Table 4 New customer segmentation table

Attribute name	Data type	Length
Customer number	char	8
Evaluation of consumer price	char	16
Consumption quantity	int	6

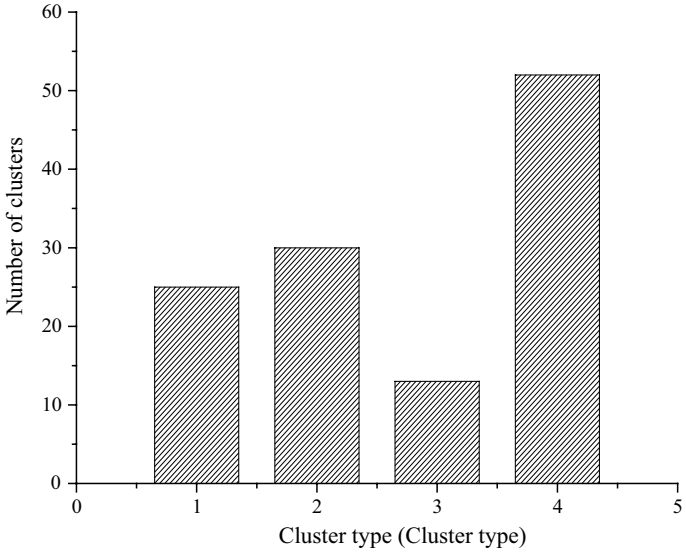

Fig. 5 The results of e-commerce customer segmentation with improved SAPK + K-means algorithm

Table 5 9 Customer consumption characteristics of different customer categories

Customer type	Consumption characteristics
Type 1	The average consumption times is 9.621 and the average consumption amount is 1028.3
Type 2	The average consumption times is 13.422 and the average consumption amount is 628.7
Type 3	The average consumption times is 4.124 and the average consumption amount is 1542.6
Type 4	The average consumption times is 2.876 and the average consumption amount is 212.8

According to the customer segmentation results obtained by SAPK-means algorithm and combined with the original data, the consumption characteristics of customers in the following customer categories can be obtained, as shown in Table 5. It can be concluded from Table 5 that different customer types have the following characteristics:

The number of the first type of customers is small, but the number of purchases is large, and the average consumption amount is also high. Most of these consumers live in first-tier cities and have the characteristics of higher education and higher income, and their age is often between 35 and 45 years old. According to the customer quantity ratio and value ratio, such customers have a higher frequency of consumption on the website, and the consumption amount is not low, which creates nearly half of the profits of the website. Therefore, such customers are called excellent customers, and enterprises should focus on maintaining such customers.

The number of the second type of customers is large, and the number of purchases is also the largest, but the average consumption amount is relatively small.

Such consumers mostly live in second and third-tier cities, and their education, age and income are of medium level. It can be concluded from the customer quantity ratio and value ratio that the number of such customers is general, but the revenue generated for the website is less, so such customers are called basic customers.

The third type of customers have the least number of customers and the least number of purchases, but the average consumption amount is the highest. Most of these consumers live in first-tier cities and have the characteristics of higher education and higher income, and they are mostly aged between 25 and 35. According to the customer quantity ratio and value ratio, such customers don't consume frequently on the website, but the consumption amount is not low, which creates a lot of profits for the enterprise, and there is still room for growth. Enterprises can take corresponding measures to encourage and promote such customers to increase the number of consumption, so such customers are called potential customers.

The fourth type of customers have the largest number of customers, but the number of purchases is less, and the average consumption amount is also the lowest. This kind of consumer lives more unevenly, have the characteristic of low education, low income, uneven age distribution. According to the customer quantity ratio and value ratio, such customers consume more frequently on the website, but consumption amount is low, which only creates a small amount of profit on the website. Therefore, such customers are called ordinary customers.

4 Conclusion

With the rapid development of computer network and e-commerce, there is a huge amount of business information stored in the database of e-commerce enterprises. In view of the defects of traditional enterprise customer segmentation, new features of customer segmentation in the e-commerce environment are proposed. By analyzing the deficiencies and existing defects of the traditional K-means algorithm and combining with AP algorithm, a new algorithm combining the semi-supervised AP algorithm and the classic K-means algorithm is proposed, namely SAPK + K-means algorithm. By applying it to e-commerce customers for segmentation research, the following conclusions are mainly drawn.

The improved SAPK + K-means algorithm has a low error rate in the clusters obtained from the two data sets, but the acquisition time is relatively long, which to some extent guarantees the high validity and accuracy of the SAPK + K-means algorithm for the clustering analysis of data information. The main steps of application of SAPK + K-means algorithm in customer segmentation include data acquisition, cluster analysis and evaluation of clustering results. And when k value is 4, SAPK + K-means algorithm has better clustering effect on customer segmentation of e-commerce websites. Analysis and evaluation are conducted from the perspective of customer value and customer behavior, and different marketing strategies are developed for different customers, further reflecting the practical value of SAPK + K-means algorithm for the segmentation of e-commerce customers.

References

Arora P, Deepali, Varshney S (2016) Analysis of K-means and K-medoids algorithm for big data. Procedia Comput Sci 78:507–512

De Roos AM, Galic N, Heesterbeek H (2009) How resource competition shapes individual life history for nonplastic growth: ungulates in seasonal food environments. Ecology 90(4):945–960

Hiziroglu A (2013) Soft computing applications in customer segmentation: state-of-art review and critique. Expert Syst Appl 40(16):6491–6507

Hoegele D, Schmidt SL, Torgler B (2016) The importance of key celebrity characteristics for customer segmentation by age and gender: does beauty matter in professional football? RMS 10(3):601–627

Huang S, Wang Q, School B (2014) Method for customer segmentation based on three-way decisions theory. J Comput Appl 34(1):244–248

Kuo RJ, Mei CH, Zulvia FE et al (2016) An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation. Neurocomputing 205(C):116–129

Luo Y, Cai Q, Xi H et al (2013) Customer segmentation for telecom with the k-means clustering method. Inf Technol J 12(3):409–413

Oña JD, Oña RD, López G (2016) Transit service quality analysis using cluster analysis and decision trees: a step forward to personalized marketing in public transportation. Transportation 43(5):725–747

Serdah AM, Ashour WM (2016) Clustering large-scale data based on modified affinity propagation algorithm. J Artif Intell Soft Comput Res 6(1):23–33

Yao Z, Sarlin P, Eklund T et al (2014) Combining visual customer segmentation and response modelling. Neural Comput Appl 25(1):123–134

Zhang T, Ma F (2017) Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function. Int J Comput Math 94(4):663–675

Reproduced with permission of copyright owner.
Further reproduction prohibited without permission.