Applied Soft Computing 111 (2021) 107677

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

GPHC: A heuristic clustering method to customer segmentation

Zhao-Hui Sun [a,*], Tian-Yu Zuo [b], Di Liang [c], Xinguo Ming [a], Zhihua Chen [a], Siqi Qiu [d]

[a] Department of Industrial Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] School of Engineering, Westlake University, Hangzhou 310000, China
[d] SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 8 October 2020
Received in revised form 18 May 2021
Accepted 27 June 2021
Available online xxxx

Keywords:
Clustering
Customer segmentation
Customer requirement analysis
Heuristics information
Evolutionary algorithm

ABSTRACT

Customer segmentation refers to dividing customer groups into multiple different sub-communities according to customer characteristics. The accurate segmentation of customers is critical for decision-makers to fully understand the customer requirements (CRs) in the market and then design market activities to satisfy customers. In past studies, clustering algorithms have been widely used to solve customer segmentation. However, it is still difficult to divide customers clearly when facing real customer requirement data (CRD). To solve these difficulties, this paper develops a heuristic clustering method for customer segmentation, termed Gaussian Peak Heuristic Clustering (GPHC, for short). Specifically, this paper utilizes the entropy method and standardized Gaussian distribution to filter and model interval CRD. Then, the customer preference pattern hidden in CRD could be recognized by niching genetic algorithm and hierarchical clustering. Finally, the clustering result of CRD will be obtained by the *k*-means algorithm based on heuristics information from customer preference patterns. Furthermore, customer segmentation can be extracted from the clustering result. A practical case is used to illustrate the effectiveness of GPHC in solving the customer segmentation problem. Experiments show that the customer segmentation result output by our method is consistent with the customer segmentation result given by experts. Besides, the robustness of GPHC in the face of complex customer segmentation scenarios has been verified through numerical experiments.

1. Introduction

Nowadays, many far-sighted enterprises have generally provided the personalized products and services to satisfy different types of customers [1]. The strategy of enterprise development is essentially a customer-oriented market behavior [2,3]. It means that enterprises should be customer-centric [4,5] and focus on the requirements of target customers [6]. Therefore, as a key link in understanding the diversity of customer requirements (CRs), customer segmentation has become an unavoidable topic. Generally, business operators collect customer requirement data (CRD) by sending questionnaires to target customers or conducting interviews with core customers. Then, CR intention [7] can be extracted from CRD. With the emergence of intelligent interconnected devices, real-time CRD collection from mobile terminals is also helping enterprises better capture CR intention.

However, no matter how advanced the methods of capturing CRD, as a product-service enterprise still face the following two important issues.

(1) To better express CR intention, interval values are widely used in collecting CRD. It leads to CRD be full of ***ambiguity***. Such fuzzy data exacerbates the difficulty of capturing the actual CRs.
(2) In recent years, with the surge in CRs of large-scale personalized products [8], CRs unearthed from CRD are shown more ***diverse***. It makes enterprises impossible to respond to all CRs with limited resources.

From the point of view of data mining, customer segmentation for the supply chain is essentially data clustering analysis. ***Ambiguity*** causes clustering to be performed on CRD with fuzzy intervals. Besides, considering that these ***diverse*** CRs with different numbers of customers, the clustering of CRD becomes more difficult.

In addition to the difficulties of data clustering, the evaluation of the clustering performance is also a difficult problem. It is difficult to measure the quality of customer segmentation results through a classic cluster evaluation index. For the evaluation of the clustering performance, it should be measured whether the clustering result is helpful for enterprise decision-makers to understand CR intention. To explain more clearly, Fig. 1 gives an example of clustering. In detail, Fig. 1(a) shows a boxplot of a cluster generated by clustering CRD. It can be seen that it is difficult to obtain information about the CR intention. In contrast,

* Corresponding author.
E-mail address: zh.sun@sjtu.edu.cn (Z.-H. Sun).

https://doi.org/10.1016/j.asoc.2021.107677
1568-4946/

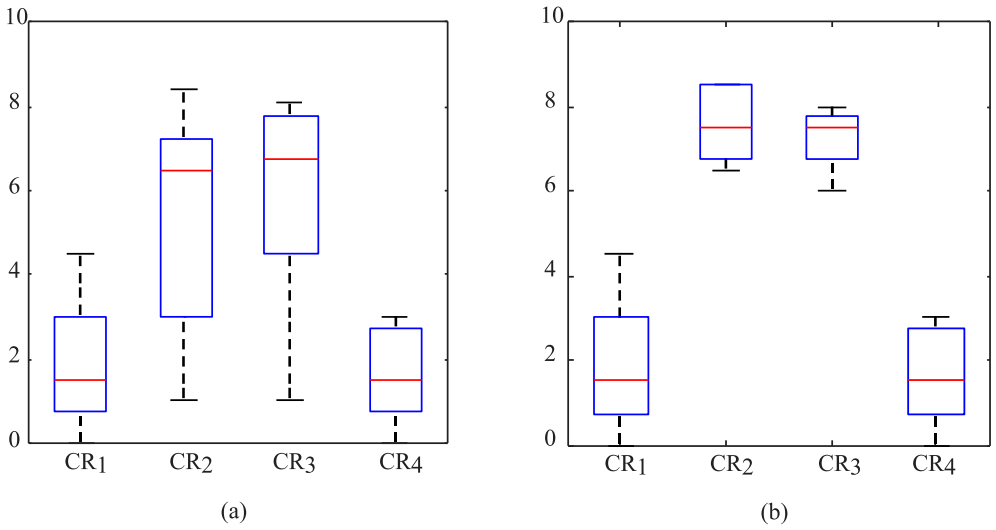

Fig. 1. An example of two boxplots from CRD clustering.

Fig. 1(b) expresses that CR_2 and CR_3 are more important than CR_1 and CR_4. The clustering result in Fig. 1(a) has little practical significance in customer segmentation. Since it cannot provide clear information for enterprises to understand customer preferences. The enterprise decision-maker hopes that each cluster generated by CRD clustering is as clear as possible (like Fig. 1(b)), which will help the enterprise accurately understand CR intention.

Based on the above considerations, a heuristic clustering method (termed Gaussian Peak Heuristic Clustering, GPHC) is proposed to analyze fuzzy CRD in this paper. In the first stage, the entropy method is used for data cleaning. In the second stage, the standardized Gaussian distribution function is used to model CRD. Next, the *customer preference pattern* is extracted from the Gaussian distribution function via the niching genetic algorithm (NGA) and agglomerative nesting (AGNES). In the last stage, the *customer preference pattern* obtained in the second stage serves as heuristics information for clustering, and then customer segmentation results can be finally obtained by match analysis.

Contribution: our contributions in this paper are illustrated as follows.

- The standardized Gaussian distribution function is used to model CRD, which better maintains the fuzziness of the raw CRD. This approach provides a new idea on the processing of interval data.
- As a typical method for multimodal optimization problems [9], NGA is first used to solve the problem of customer segmentation. In GPHC, the combination of NGA and AGNES provides heuristic information for the *k*-means clustering, which obtains a good clustering performance.
- An improved *k*-means that uses heuristic information is proposed. The algorithm structure of the proposed algorithm is similar to that of typical *k*-means, which makes it easier to understand by decision-makers than those customer segmentation methods or clustering analysis methods with complex structures. The interpretability of the algorithm would also facilitate decision-makers to intuitively understand CRs.
- In terms of solving practical problems of customer segmentation, our proposed GPHC successfully captures multiple intentions of requirement combination behind customer preference from a large amount of fuzzy CRD with noise. Practical case studies and numerical experiments have shown the superior performance and robustness of GPHC, especially in the face of complex customer segmentation scenarios.

The remainder of this paper is organized as follows. In Section 2, the literature review of related works is presented. Section 3 describes the proposed GPHC in detail. In Section 4, a case study is conducted to verify the feasibility and effectiveness of GPHC. In Section 5, numerical experiments are designed and conducted to evaluate the performance of GPHC in complex customer segmentation scenarios. The conclusion is summarized in Section 6.

2. Background and related work

2.1. Customer segmentation

Customer segmentation is aiming to divide customers into multiple subsets according to CR intentions (or preferences). After division, the customers in the same subset will have the same character while the customers in different subsets will have obvious differences. In the past research, many works have focused on customer segmentation for a specific user market based on traditional segmentation theories or proposed novel multiple segmentation attributes. Chen et al. [10] studied the smartphone market with rich functionality and multi-interaction characteristics where the customer's interaction preference can be fully released. They defined and proposed the usage pattern by analyzing the customer's usage frequency of APP, and then completed customer segmentation through the proposed algorithm. Zeybek et al. [11] identified six customer groups with different characteristics by combining with a multi-method approach for the customer segmentation of the rail freight traffic in the Turkish State Railways. Dzobo et al. [12] presented a multi-dimensional customer segmentation model for the reliability-worth analysis of power systems. Three customer attributes including economic size, economic activity, and energy consumption were considered in their proposed model and the hierarchical clustering technique was used to cluster electricity customers into customer segments of similar cost characteristics. Maria et al. [13] segmented the customers of cashback websites based on customers' commercial activity and role within the social network composed of the website. Nakano et al. [14] examined how customers choose multiple channels and media in modern retail environments. The customer segmentation in their work mainly considered the following attributes, including purchase channels of bricks-and-mortar and online stores, media touchpoints of PC, mobile, and social media, and psychographic and demographic characteristics.

Table 1
The summary of previous works.

Author(s)	Technique
Xu et al. [15]	Sparse fuzzy *k*-means
Murray, Agard and Barajas [16]	*k*-means
Mohammadzadeh, Zare Hoseini and Derafshi [17]	
Khalili-Damghani, Abdi and Abolmakarem [18]	*k*-means
Sano et al. [19]	*k*-medoids
Bose and Chen [20]	Extended fuzzy *C*-means
Llanos et al. [21]	SOM
Hong et al. [22]	Combination of SOM and *k*-means
Seret, Maldonado and Baesens [23]	
Dursun and Caber [24]	SOM, *k*-means, and RFM
Wei et al. [25]	Combination of SOM and LRFM
Wei et al. [26]	Combination of *k*-means and RFM
Cheng et al. [27]	Combination of *k*-means and LRFM

In the above papers, their researches are mainly focused on the design of customer segmentation attributes for different scenarios, but paid less attention to design the method to solve the customer segmentation problem. To solve the problem of customer segmentation, therefore, this paper proposed the *k*-means based framework for customer segmentation, termed GPHC. Different from studying the specific attributes of CRs, in our framework, a set of general modeling methods for CRD is explored, which could provide the methodological references for the research in different customer markets.

2.2. The methodology for customer segmentation

Effective customer segmentation could help the enterprise better understand CR intention. It provides the enterprise with a reference that guides the enterprise to make suitable market activity (such as conducting the market campaign, organizing production strategy, selecting suppliers) with limited resources. On the contrary, the inaccurate division of customers will lead to an unreasonable market activity design that could not respond to customer expectations of products and services. Consider the impact of customer segmentation on market activity design, many analysis methods have been used to solve the customer segmentation problem. Among these analysis methods, as a data-driven method, clustering analysis was widely studied. Clustering analysis methods can be roughly divided into two categories including non-heuristic clustering (such as *k*-means) and heuristic-based clustering.

2.2.1. Non-heuristic clustering

In previous works, many non-heuristic clustering methods were used to solve customer segmentation, such as *k*-means, self-organizing map (SOM), *k*-medoids, and so on. Some works combine different clustering algorithms or mix clustering algorithms with customer management models, such as RFM (Recency, Frequency, and Monetary). Table 1 summarizes recent works in the field of non-heuristic clustering and their applications in customer segmentation. It can be seen in Table 1, *k*-means is the most popular clustering algorithm.

Although many clustering algorithms have been used to solve customer segmentation problems, there is a lack of research on clustering performance on complex real data and unfavorable conditions, such as cluster overlap, outliers, and noise points. Given the important role of customer segmentation, it not only needs to focus on whether the clustering algorithm can be applied to customer segmentation but also needs to focus on the clustering performance.

2.2.2. Heuristic-based clustering methods

Given the recent advances in evolutionary computing algorithm [28,29], in the field of clustering, some researchers apply heuristic algorithms to improve the performance of clustering algorithms. Genetic algorithm (GA), artificial bee colony algorithm (ABC), and other evolutionary computing methods have been widely combined with clustering algorithms in previous work to solve the clustering problem. These previous works can be founded in [30,31]. Generally speaking, heuristic-based clustering methods express the initial clustering results in the form of solutions, and iteratively improve the quality of the clustering results through heuristic methods. Through the literature review, we found that the heuristic-based clustering method is rarely used to solve the problem of customer segmentation. In our study, NGA which is widely used in the search for the peak point of the multi-peak function is firstly applied in customer segmentation. As a significant process of GPHC, NGA provides effective heuristic information for clustering CRD to obtain exactly customer segmentation results. The introduction to NGA will be given in detail in Section 3.

3. Proposed method

In this section, the whole process of customer segmentation is introduced in detail. Firstly, data preprocessing is conducted to clean those CRD with no obvious CR intention. Then, cleaned CRD is transformed into a standardized Gaussian function. Finally, GPHC is implemented to clustering CRD based on the standardized Gaussian function for customer segmentation. Fig. 2 shows the overall technological process of our proposed method.

3.1. Basic concept

Before discussing the specific technical steps, the concept of CR item, CRD, and CR intention are introduced here.

Clustering is generally based on attributes, i.e., multiple samples will have multiple attributes. Therefore, for customer segmentation, it is also necessary to establish multiple CR attributes, which are termed CR items. A group of CR items is generally designed by experienced experts in combination with product and service characteristics. Around CR items, different customers' ratings for multiple CR items can be obtained through different information collection ways. The data collected from a large number of customers on CR items is termed CRD. The collection of CRD is used to understand customer behavior and analyze CR intention [32]. It should be noted that the CR intention mentioned in this paper is based on CR items. i.e., each CR intention is one combination of different CR items. Furthermore, the set of CR intentions existing in the actual market is a subset of the combinations of CR items.

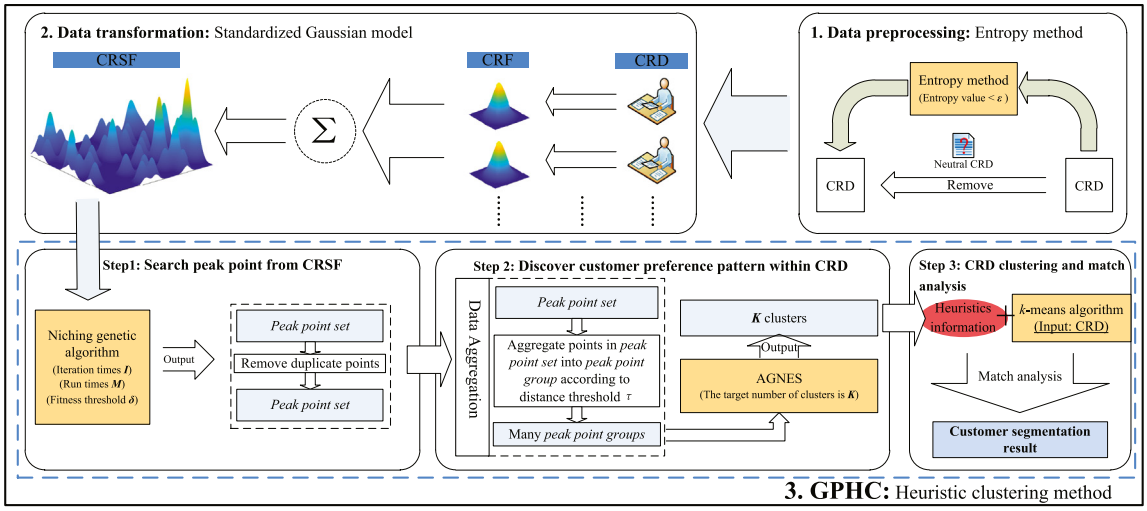

Fig. 2. The overall technological process.

Due to the ambiguity of CRD, CR items often use interval scores within a certain range for evaluation. To make the statement clearer, some explanations are made as follows.

(1) The score range of each CR item is often set from 0 to 10. A higher score indicates a higher customer expectation. For example, one customer scores "9" for "Quality", which means that the customer wants high-quality products. On the contrary, if one customer scores "3" for "Price", it means that the customer is less sensitive to the price of the products.

(2) The scores of different CR items from one customer can reflect the information of customer preference. For example, the scores of four CR items from one customer are [8,1,7,3]. It can be seen that the customer is particularly concerned about the first and third CR items. Therefore, to achieve high customer satisfaction, these CR items must be prioritized.

(3) Each score can be set as an interval value. Using a crisp number to indicate the expected degree of CR items will reduce the ability to reflect the subjectivity of CR intention. Therefore, interval scores will be more suitable for the evaluation of CR intention.

3.2. Data preprocessing: The entropy method

The CRD exported by an enterprise contains not only customer data with obvious CR intention [33] but also many neutral data without CR intention. Those neutral data are not helpful for enterprises to understand the CR intention, therefore should be deleted. For example, Table 2 shows the scores from three customers with three CR items in columns CR_1, CR_2, CR_3. It can be seen that the CR intention from $Customer_1$ is clear with an extremely low value in CR_1, CR_2, and extremely high value in CR_3. Different from $Customer_1$, the CR intention of $Customer_2$ and $Customer_3$ is neutral, and there is no obvious difference among different CR items. The neutral data will seriously affect the quality of clustering. Therefore, it is necessary to delete neutral data as much as possible before clustering. However, this is not easy work. Whether the CR intention is clear itself is a vague concept. For example, "9" is a very high score. It' s is clear for CR intention. But it is difficult to define whether the intention of score "6" is a high score or not. Therefore, it is impossible to establish an objective evaluation standard to distinguish whether CRD has obvious CR intention or not.

Considering above all, the entropy method is proposed to represent the uncertain degree of data [34], which could combine objective data evaluation and subjective experience judgment. The entropy method is described as follows.

Table 2
The scores from three customers with three CR items.

	CR_1	CR_2	CR_3	Entropy
$Customer_1$	[0,1]	[1,2]	[9,10]	0.808
$Customer_2$	[5,6]	[3,5]	[5,6]	1.570
$Customer_3$	[7,8]	[7,8]	[6,7]	1.582

First, a quantitative index based on entropy is used to describe the uncertainty in scoring CR items. The index is helpful to filter neutral data. Then, experience judgment based on experts' knowledge is used to determine the threshold of entropy filtering (termed ε). Assume a vector $x = [x_1, x_2, \ldots, x_i, \ldots, x_n]^T$ contains the scores of multiple CR items from one customer, where n is the number of CR items. And $x_i = [l_i, u_i]^T$, where l_i and u_i are the lower bound and upper bound of x_i, respectively. The mean value of each interval (i.e., $\overline{x_i} = (l_i + u_i)/2$) is taken to calculate the entropy. The entropy of each x (i.e., CRD) can be obtained according to Eq. (1).

$$H(x) = -\sum_{i=1}^{n} P(\overline{x_i}) \log_2 P(\overline{x_i}) \tag{1}$$

where $P(\overline{x_i}) = \overline{x_i} / \sum_{i=1}^{n} \overline{x_i}$ represents the score percentage of the i-th CR item to the total n CR items in one sample of CRD.

The larger the entropy value is, the more neutral the customer data is. Table 2 shows the entropy values corresponding to the three customers in the column of "Entropy". It can be seen that the entropy values of $Customer_2$ and $Customer_3$ are significantly higher than those of $Customer_1$. This is consistent with the obvious CR intention of $Customer_1$ among the three customers. We use entropy value to evaluate the degree of CR intention. When a threshold ε of entropy is set according to the experience of experts, the CRD with the entropy value greater than ε is removed but those with obvious CR intention could be retained.

3.3. Data transformation: Standardized Gaussian model

From the perspective of clustering, interval data is more challenging than a crisp number [35]. Lu et al. [36] proposed a fuzzy clustering algorithm for interval data based on Gaussian distribution, and D'Urso et al. [37] proposed a fuzzy c-ordered medoids clustering algorithm for interval data clustering. Inspired

Table 3
The scores from one customer with four CR items.

Quality	Reserve capacity	Delivery cycle	Price
[3,4]	[8,9]	[5,7]	[5,5]

by the above literature, the standardized Gaussian distribution is used to describe the interval CRD.

Firstly, the general form of the $(n-1)$-dimensional Gaussian distribution function is given in R^n space.

$$N(x, \mu, \sum) = \sqrt{\frac{1}{(2\pi)^{n-1}\det(\sum)}} \exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right) \tag{2}$$

where x is a vector with $n-1$ dimensions, $\Sigma = dig\ (\sigma_1^2, \sigma_2^2, \ldots, \sigma_{n-1}^2)$ is the covariance matrix, $\mu = (\mu_1, \mu_2, \ldots, \mu_{n-1})^T$ is the mean value vector.

For clarity, a specific example is used to illustrate the modeling of CRD. Here, the scores of four CR items from one customer are shown in Table 3.

According to Table 3, it can be seen that the CRs have four dimensions. Therefore, the Gaussian distribution function is established in the R^5 space.

Calculate the variance σ^2 and mean value μ in each dimension, respectively. For the calculation of σ^2, equidistant-discretization processing is conducted on interval data in each dimension. In this case, the discretization step is set as 0.1. The results of discretization for Table 3 are shown in Table 4. The special situations encountered in discretization processing are explained below.

For the case where the length of the interval is 0 (such as the interval [5, 5] of "Price" in Table 3), the interval length needs to be enlarged for discretization. The enlarged length is generally set to be the half value of the minimum length of the non-zero interval among all CR items. For example, the minimum length of the non-zero interval is 1 in Table 3, therefore the enlarged length is set to 0.5. Then, the interval [5, 5] can be enlarged to [4.75, 5.25] for subsequent discretization. Besides, if the data interval is an edge interval (such as [0, 0] or [10, 10]), follow the above steps, but remove the interval out of [0, 10].

By calculating, $\Sigma = dig(0.33, 0.33, 0.62, 0.62)$ and $\mu = (3.5, 8.5, 6.0, 5.0)^T$. Then, the scores of four CR items from one customer can be modeled by Eq. (2). By transforming all CRD into its corresponding Gaussian distribution, the ambiguity of customers' requirements can be well described.

Using Gaussian distribution to characterize CRs has the following properties.

Property: The $f(x)$ (i.e., function value) is considered as the 'preference degree' of mapping the corresponding interval CRD into a crisp vector x.

(1) The function takes its peak at (3.5, 8.5, 6.0, 5.0). Therefore, although each CR item is an interval value, CR intention is most likely to be expressed as μ. This is consistent with the habit that humans tend to use central values to describe fuzzy events.

(2) The function value decreases when x moves away from μ. It is indicated that the CR intention contained in interval CRD is not inclined to start point or endpoint (i.e., l_x or u_x).

The advantage of Gaussian distribution is that it can further use uncertainty to characterize CR intention while maintaining the ambiguity of customer rating, and reasonably enhance the representative of the scores of CR items.

However, there is still a serious problem in using the Gaussian distribution to describe CRD. The Gaussian functions established by different customers may have large differences in peak values, which leads to the "unequal status" between different customers. This is due to the different variance after discretization processing. The operation termed "standardization" is, therefore, to fix the term "$\sqrt{\frac{1}{(2\pi)^{n-1}\det(\sum)}}$" as 1. The standardized Gaussian distribution can be described below, termed customer requirement function (CRF).

$$N_S(x, \mu, \sum) = 1 \times \exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right) \tag{3}$$

To ensure that the weights of different customers are equal, the value range of the Gaussian function is stretched into [0, 1] via "standardization". Although the mathematical properties of the Gaussian function have changed with losing its coefficient after "standardization", covariance matrices and mean value vectors can still reflect the distribution information.

For visualization, a two-dimensional Gaussian distribution function in R^3 space is shown as an example to explain the "standardization" operation in Fig. 3. For different Gaussian functions corresponding to different customers, the large differences between their peaks can be solved by "standardization" operation, while maintaining the ability to reflect the distribution information. As shown in Fig. 3(a), assume $Customer_u$ and $Customer_v$ corresponds to two different CRD, where $\Sigma_u = dig(0.33, 0.62)$ and $\Sigma_v = dig(0.33, 0.33)$. Therefore, the coefficients before the exponential terms are 0.14 and 0.19 according to Eq. (2). It causes the peak value of CRF of $Customer_u$ to be lower than that of $Customer_v$. But after the "standardization" operation, the coefficients before the exponential terms of the two are both 1. It makes the CRFs corresponding to $Customer_u$ and $Customer_v$ have the same peak value. It eliminates the difference between the two as Fig. 3(b).

So far, each CRD corresponds to a CRF. To comprehensively analyze all CRD, the customer requirement sum function (CRSF) is obtained by the summation of m CRFs as follows.

$$CRSF = \sum_{i=1}^{m} N_S^i \tag{4}$$

m is the number of customers. Since CRSF contains all information of CRD, the analysis of CRD is transformed into the analysis of CRSF.

3.4. GPHC: A heuristic clustering method

To visualize the whole operation process of GPHC, an example of 50 CRD with two CR items is taken to analyze.

Perform equidistant-discrete processing on CRD, then use Eqs. (3) and (4) to obtain CRF and CRSF. The function image and heatmap of CRSF with 50 CRD are shown in Fig. 4(a) and Fig. 4(b), respectively.

3.4.1. Step 1: Search peak points from CRSF

(1) Why do search for peak points in CRSF

From the three-dimensional image of CRSF in Fig. 4(a), it can be found that CRSF has many peaks. This phenomenon is more clearly shown in the heatmap from Fig. 4(b). In the heatmap, the closer to yellow the color is, the greater the function value is, while the closer to blue the color is, the smaller the function value is. As the discussion in Section 3.3, CRSF reflects the comprehensive distribution of CRD. Therefore, the multi-modal phenomena showed in CRSF can be explained as there are many different CR intentions in CRD. The higher peak is due to the superposition of many CRFs in the same position, while the lower peak is the superposition of a few CRFs in the same position. The difference in height of peaks shows that there exists an imbalance in different

Table 4
The discretization result of Table 3.

Quality	Reserve capacity	Delivery cycle	Price
(3.0, 3.1, . . . , 3.9, 4.0)	(8.0, 8.1, . . . , 8.9, 9.0)	(5.0, 5.1, . . . , 6.9, 7.0)	(4.75, 4.85, . . . , 5.15, 5.25)

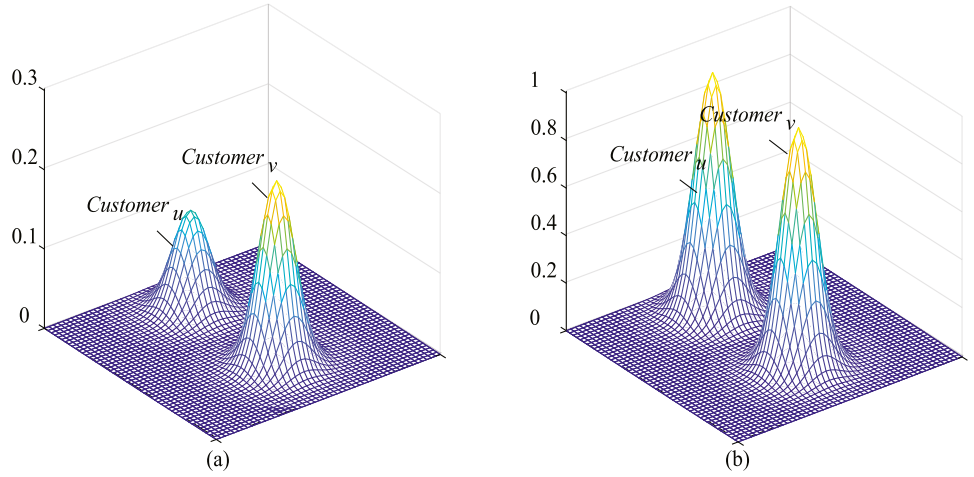

Fig. 3. The difference between the Gaussian distribution function and the standardized Gaussian distribution function. (a) shows the Gaussian distribution function corresponding to $Customer_u$ and $Customer_v$, and (b) shows the two Gaussian distribution functions after "standardization".

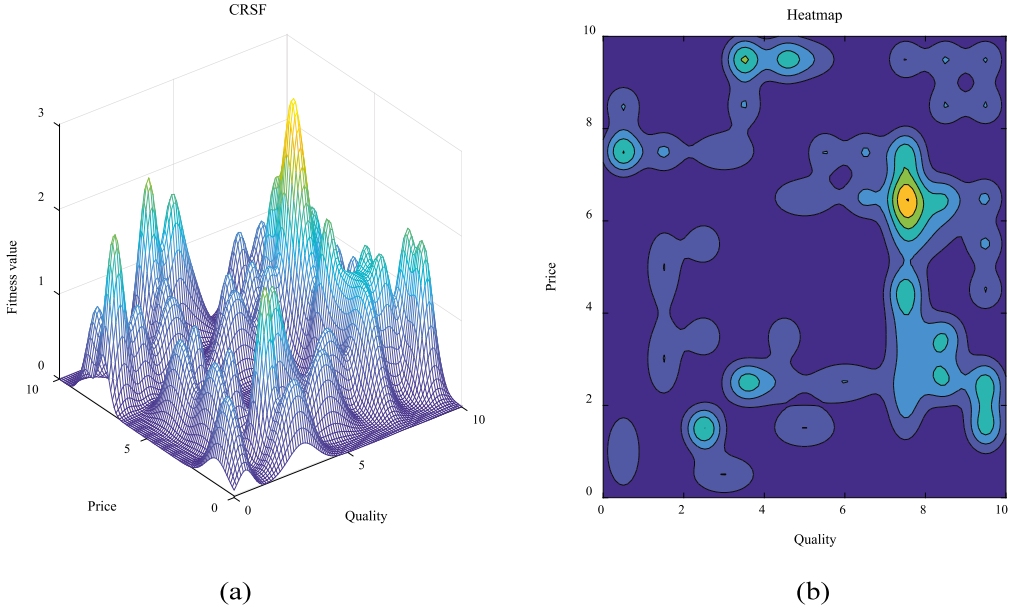

Fig. 4. The function image and heatmap of CRSF with 50 CRD.

CR intentions from CRD. In other words, the area with a higher peak is an area with dense customers, and an area with a lower peak is an area with sparse customers. By searching the peak points of CRSF, the clustering center of the customer group in the space can be found in advance before the start of clustering. It provides key heuristic information for subsequent clustering. Therefore, searching for peak points of CRSF is a key sub-problem of clustering CRD.

(2) How do search for peak points in CRSF: NGA

In our method, NGA is used to search peak points in CRSF, which could be considered a multimodal optimization problem [38]. The optimization objective can be described as:

$$\arg\max_{x}(CRSF) \tag{5}$$

As an extension of GA, NGA sets a certain mechanism based on GA to ensure the population can be scattered into the entire solution space [39,40]. Then, individuals can further evolve in different "niching" environments. NGA solves the problem that GA often converges to local peak points prematurely when searching for peak values of multi-modal function [41].

The implementation of NGA in this paper is based on the mechanism of eliminating a similar structure. Specifically, when one generation of populations is selected, crossed, and mutated, the distances between individuals are compared according to Eq. (6).

$$\|X_i - X_j\| = \sqrt{\sum_{k=1}^{l}(X_i^k - X_j^k)^2} \tag{6}$$

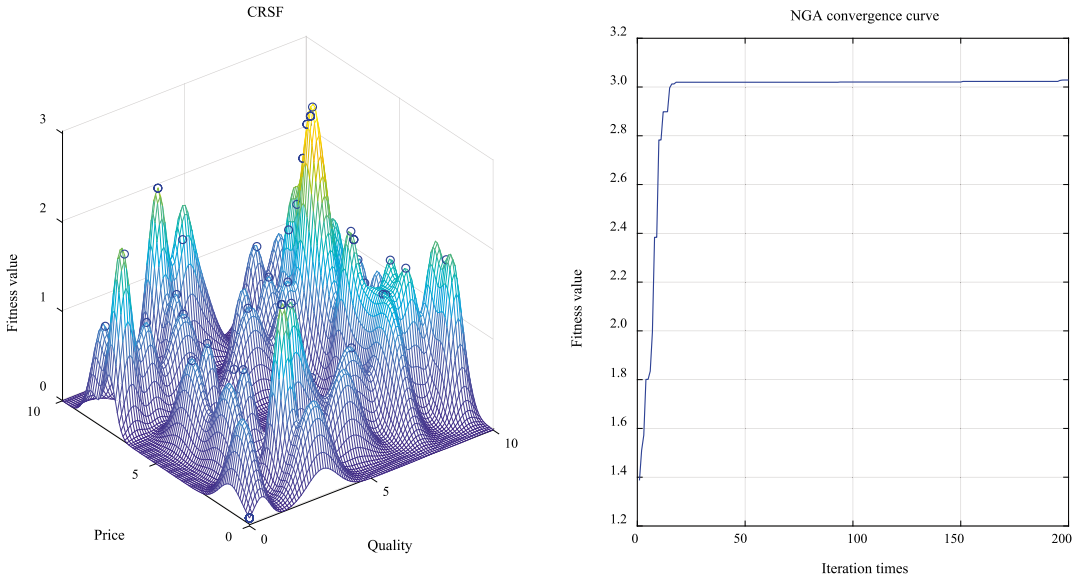

Fig. 5. The one search result of running NGA for CRSF. Each blue circle in (a) represents a local (or global) maximum point and (b) gives the convergence curve from one search of NGA.

where X_i, X_j represents individual i and individual j, l represents the length of the chromosome coding string.

When the distance is within the pre-specified niching radius L (i.e., $\|X_i - X_j\| < L$), the individual will be punished. Therefore, in the same niche environment, those individuals with lower fitness become smaller and then, more likely to be eliminated in the subsequent evolution process. It ensures that there is only one good individual in the same niche environment, and different individuals can be dispersed into the whole solution space. Besides, the elite strategy is added into NGA (i.e., the top N best individuals selected from each generation do not participate in the selection, crossover, and mutation operations of the next generation, but directly participate in the niching similar structure elimination operation with the new individuals). This strategy promotes NGA to converge to a globally optimal solution.

Fig. 5 shows the visualization results and algorithm convergence curve of searching peak points of CRSF by NGA. Each blue circle in Fig. 5(a) represents a local (or global) maximum point and the convergence curve from one search of NGA is shown in Fig. 5(b). It can be seen that the NGA not only converges to the global maximum value points but also finds some local maximum points scattered in the solution space.

(3) The necessity of Multi-running NGA

As shown in Fig. 5(a), one running of NGA cannot guarantee that all CRSF peak points be searched. To avoid the missing heuristic information, it is necessary to run NGA multiple times (M times) to find as many different peak points. In the search process, those points with very low fitness also will be searched by NGA. The fitness threshold δ is set to filter them. Finally, NGA will obtain a *peak point set*, which contains a large number of peak points. Besides, due to multiple times running of NGA, *peak point set* may contain many duplicate points. Those duplicate points should be deleted.

3.4.2. Step 2: Discover customer preference pattern within CRD

(1) Peak point aggregation: similarity measure

After deduplication, it contains a lot of similar data in the *peak point set*. These similar data will not helpful for the discovery of the clustering center of the customer group. Considering this problem, data points in the *peak point set* will be merged according to the similarity measure. The specific procedures are as follows. First, a distance threshold τ is set to define the similarity between data points. Then, traverse *peak point set* to merge those points with the Euclidean distance less than τ into a group (termed *peak point group*). Generally, τ is set far smaller than the niching radius L of NGA. Due to the small τ, the data points in the same group are highly similar.

(2) Customer preference pattern discovery: hierarchical clustering

Since τ is set to be a small value, there may still be similarities between *peak point groups*. AGNES [42], therefore, is used to further cluster similar *peak point groups*. Then, the *customer preference pattern* (i.e., the cluster center of the customer group) within CRD can be discovered. The application way of AGNES is as follows.

AGNES is a hierarchical clustering algorithm using a bottom-up aggregation strategy. First, treat each sample data as an initial cluster, and then find the two clusters with the closest similarity and merge them in each round of algorithm. This process is repeated until the preset number of clusters is reached. The central point of each *peak point group* is used as the representative point to input the AGNES, and the Euclidean distance is used as the similarity measurement between clusters to perform clustering work. Set the algorithm to stop when the number of clusters equals the preset K. After AGNES, all peak point groups are merged into K clusters. The centroid of each cluster corresponds to a *customer preference pattern*. The centroid of the i-th (i from 1 to K) cluster is calculated as follows.

$$c^i = (c_1^i, c_2^i, \ldots, c_n^i)^T = \frac{\sum_{j=1}^{Num_i}(x_1^{i,j}, x_2^{i,j}, \ldots, x_n^{i,j})^T}{Num_i} \tag{7}$$

where n *is* equal to the number of CR items, $(x_1^{i,j}, x_2^{i,j}, \ldots, x_n^{i,j})^T$ is a j-th data point in the i-th cluster, and Num_i is the number of the data point in the i-th cluster.

The obtained K centroids (*customer preference patterns*) provide high-quality heuristic information for subsequent CRD clustering.

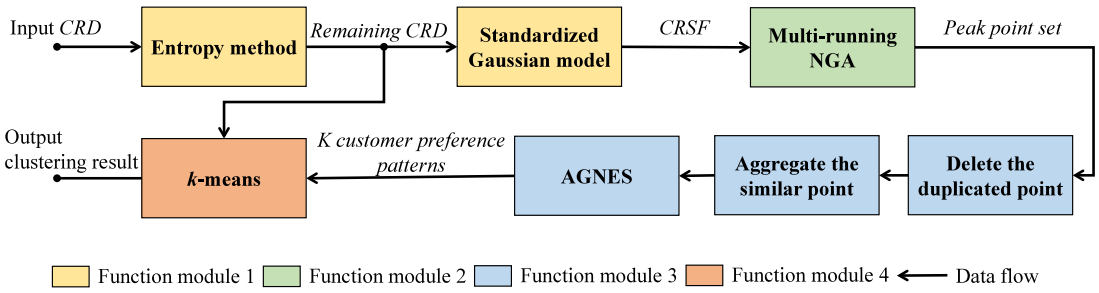

Fig. 6. The data flow between function modules of the proposed framework.

3.4.3. Step 3: CRD clustering and match analysis

As the last step of GPHC, the *k*-means algorithm is used to cluster CRD, where heuristic information, *K customer preference patterns*, are used as the initial cluster centroids of *k*-means. Considering that each data of CRD is an interval value, take the mean value of the interval as the input of *k*-means. Then, *k*-means will output *K* high-quality classes. In each class, CRD has similar data characteristics. Finally, introduce the expert experience to match *K* classes with different CR intentions. Compared with the method to label each data in CRD with CR intentions one by one, our proposed method only needs experts to label data *K* times. It greatly saves the expert manpower investment required by customer segmentation. Especially in the industry where the customer changes dramatically, it will be a great potential to use our method to analyze CRs and then quickly organize market activities that can accurately respond to CRs.

So far, the whole process of GPHC has been introduced. To express the above process clearly, the algorithm procedure of GPHC is given in Algorithm 1. The specific procedures of AGNES and heuristic *k*-means in GPHC are given in Algorithm 2 and Algorithm 3, respectively.

3.5. Data flow

To show the proposed framework more clearly, the diagram of data flow in our method is shown in Fig. 6. To obtain the customer segmentation result, the improved *k*-means method is used to cluster CRD. Different from the typical *k*-means, the initial centroids in *k*-means are obtained by our combination of multiple operations. Function module 1 is the modeling of CRD by CRSF. Then, the NGA is used to search peak points of CRSF. The output of function module 2 is the set of peak points. Function module 3 consists of three operations, which could output the *K customer preference patterns*. Finally, the *K customer preference patterns* are used as the *K* centroids of *k*-means for clustering.

3.6. Parameter setting recommendation

Table 5 summarizes all parameters used in our proposed methods, and put forward suggestions for the setting of key parameters. Setting recommendations for key parameters are given below.

For *M*, when the customer segmentation problem is complex, such as the high-dimensional CRD or serious imbalance in CRD, this parameter should be increased appropriately to ensure that NGA can find a variety of peak points. For *I*, it mainly needs to ensure that the single NGA can converge to the solution with a higher fitness value (i.e., global optimal or local optimal). For *L*, it is suggested to set to half of the number of CR items. For δ, it needs to be determined by the complexity of the actual problem. When the dimension of the problem (i.e., the number of CR items) is low, it is more likely that multiple CRFs gather in a certain region. The fitness values of peak points, which can be searched by NGA, are mostly greater than 1. δ is therefore suggested to set slightly less than 1. However, when the dimension of the problem increases, the dispersion of CRFs in space will increase. The fitness values of peak points, which can be searched by NGA, will be decreased. δ is therefore suggested to set a lower value. For τ, it is noted that a smaller τ will greatly increase the quality of the *peak point group*, but it is time-consuming and will seriously affect the efficiency of the algorithm. It is recommended to start with a value slightly lower than *L*. 10. For *K*, the choice of *K* is mainly based on the possible number of CR items combination and expert experience. The maximum number of CR items combination is calculated as follows.

$$T_{\max} = \sum_{i=1}^{n} \binom{n}{i} \tag{8}$$

where *n* is the number of CR items.

In most cases, only part of the CR items combination will appear in a batch of CRD, so the setting of *K* can be less than T_{max}. However, considering the possible clusters of outlier points, a better value for *K* is around T_{max}. Therefore, experts need to combine T_{max} and the possible CR intentions to set *K* comprehensively.

3.7. Computational complexity analysis

The computational complexity of our proposed method included data preprocessing (entropy method), data transformation (standardized Gaussian model), and the GPHC. The computational complexity of data preprocessing is $O(n)$, where *n* is the number of CRD. The computational complexity of data transformation is also $O(n)$.

The computational complexity of the GPHC as follows:

$$O(MI(N+E)^2) + O((\delta MI)\log(\delta MI)) + O({n_{pp}}^2) + O({n_{pps}}^2(n_{pps}-K)) + O(K\varepsilon nt).$$

$O(MI(N+E)^2)$ represents the computational complexity of running NGA multiple times. The notation of the above can be seen in Table 5. $O((\delta MI)\log(\delta MI))$ represents the computational complexity of deleting the duplicated points. *MI* is the number of peak points obtained by running NGA multiple times. Therefore, δMI is the number of peak points considered to be reserved. $O({n_{pp}}^2)$ represents the computational complexity of data aggregation, where n_{pp} is the number of the remaining peak points after data aggregation. $O({n_{pps}}^2(n_{pps}-K))$ represents the computational complexity of AGNES, where n_{pps} is the number of peak points in the *peak point set*, *K* is the predetermined target number of clusters. $(n_{pps}-K)$ is the number of iterations of the AGNES. $O(K\varepsilon nt)$ represents the computational complexity of *k*-means, where εn is the number of the remaining CRD. *t* the number of iterations of the *k*-means. According to the value range of each parameter in the actual business scenario, the computational complexity of the GPHC could be simplified as $O(MI(N+E)^2) + O({n_{pp2}}^2)$.

Algorithm 1: GPHC

Begin

Load CRD and CRSF;
//NGA
Set *peak point set* and δ;
for *NGA run times* ←1 **to** *M*
  for *Iteration number* ←1 **to** *I*
    NGA selection operation;
    NGA crossover operation;
    NGA mutation operation;
    Niching elimination operation;
    Save populations in which those fitness values of points are higher than δ into *peak point set*;
  end for
end for
Remove duplicate points from *peak point set*;
//Data aggregation
Set τ;
while *peak point set* $\neq \varnothing$ **do**
  Create a new *peak point group*;
  Randomly select data in the *peak point set* as a reference point and move it to the current *peak point group*;
  Calculate the Euclidean distance between each point in the *peak point set* with the reference point;
  Move those points which Euclidean distance is less than τ to the current *peak point group*;
  Save the current *peak point group*;
end while
//AGNES
Set *K*;
Implement AGNES (the whole process is shown in Algorithm 2);
Calculate the *customer preference patterns* according to Eq. (7);
//*k*-means
Heuristic *k*-means method (the whole process is shown in Algorithm 3);

End

Algorithm 2: AGNES in GPHC

Begin

Load *peak point groups* and *K*;
while the number of *peak point groups* > *K* **do**
  Calculate the average distance, d_{avg}, between every two groups;
  The d_{avg}^{ij} of $group_i$ and $group_j$ can be calculated by: $d_{avg}^{ij} = \frac{1}{|group_i||group_j|}\sum_{x\in group_i}\sum_{y\in group_j} dist(x,y)$
  Combine the two closest groups;
  Update the number of *peak point groups*;
end while
Output *K groups of peak point*;

End

Algorithm 3: Heuristic *k*-means in GPHC

Begin

Load *K* and *K customer preference patterns*;
Load the average value of each remaining CRD after entropy method (data sample);
Set the termination parameter *p*;
Set *K* as the target number of clusters;
Set *K customer preference patterns* as the initial cluster centroids;
repeat
  Calculate the distance between each data sample with *K* centroids of clusters, respectively;
  Back up data sample, and merge each data sample into the cluster that closest to it;
  Recalculate the *K* centroids of clusters according to Eq. (7);
until *K* centroids of clusters do not update *p* times;
Output *K* classes;

End

4. Case study

4.1. Data description

In this section, a practical case is used to verify the effectiveness of the proposed method. The CRD used in this section is derived from the statistical results of data obtained by a market survey. In the survey, customers are required to express their intentions towards one or more requirements when filling out the questionnaire. 300 feedback questionnaires with the scores of four CR items (Quality, Reserve capacity, Delivery cycle, and Price) have been fully recorded as CRD. The details of original data using in this section have been put on https://github.com/polysun/Customer-Segmentation.

4.2. Customer segmentation

4.2.1. Data preprocessing

Calculate the entropy value of each CRD according to Eq. (1) and sort CRD by entropy in ascending order. The sort results of

Table 5
The summary of parameters.

Parameter	Description	Belong to
ε	The threshold value in the entropy method	Data preprocessing
M	The times of NGA running	GPHC-NGA
I	Iteration times of NGA in a single running	GPHC-NGA
N	The number of populations	GPHC-NGA
E	The number of elite populations	GPHC-NGA
Pc	Crossover probability	GPHC-NGA
Pm	Mutation probability	GPHC-NGA
L	Niching radius	GPHC-NGA
$Penalty$	The penalty coefficient in the mechanism of eliminating similar structure	GPHC-NGA
δ	Fitness threshold	GPHC-NGA
τ	Distance threshold	GPHC-Data aggregation
K	The target number of clusters	GPHC-AGNES and k-means

Table 6
Entropy values of 300 CRD.

Order	Entropy value														
1–15	1.024	1.206	1.209	1.230	1.286	1.289	1.292	1.299	1.300	1.320	1.320	1.322	1.325	1.339	1.340
16–30	1.350	1.350	1.350	1.355	1.355	1.363	1.379	1.383	1.403	1.422	1.448	1.457	1.460	1.460	1.460
31–45	1.461	1.466	1.469	1.473	1.480	1.480	1.480	1.480	1.489	1.489	1.489	1.491	1.491	1.491	1.491
46–60	1.500	1.509	1.517	1.517	1.520	1.526	1.540	1.540	1.543	1.548	1.549	1.549	1.553	1.560	1.560
61–75	1.567	1.571	1.572	1.574	1.578	1.579	1.583	1.583	1.583	1.587	1.587	1.590	1.598	1.599	1.604
76–90	1.605	1.607	1.607	1.607	1.607	1.608	1.608	1.608	1.608	1.608	1.608	1.610	1.610	1.615	1.617
91–105	1.618	1.622	1.624	1.627	1.627	1.627	1.631	1.631	1.634	1.636	1.636	1.636	1.637	1.637	1.640
106–120	1.646	1.646	1.650	1.650	1.651	1.660	1.660	1.662	1.664	1.665	1.667	1.667	1.667	1.669	1.669
121–135	1.674	1.676	1.680	1.684	1.685	1.685	1.685	1.686	1.686	1.686	1.686	1.688	1.692	1.693	1.703
136–150	1.703	1.711	1.711	1.715	1.723	1.723	1.728	1.728	1.728	1.728	1.731	1.731	1.733	1.734	1.739
151–165	1.742	1.742	1.743	1.748	1.750	1.751	1.751	1.755	1.758	1.758	1.758	1.761	1.761	1.761	1.761
166–180	1.763	1.769	1.769	1.772	1.783	1.783	1.784	1.784	1.788	1.790	1.790	1.790	1.792	1.796	1.796
181–195	1.796	1.796	1.798	1.802	1.802	1.802	1.802	1.807	1.807	1.807	1.811	1.811	1.815	1.815	1.817
196–210	1.819	1.821	1.823	1.823	1.823	1.836	1.836	1.836	1.837	1.838	1.839	1.840	1.840	1.840	1.840
211–225	1.848	1.848	1.848	1.848	1.851	1.852	1.854	1.854	1.854	1.859	1.862	1.864	1.866	1.867	1.867
226–240	1.870	1.870	1.870	1.871	1.871	1.871	1.871	1.873	1.876	1.877	1.878	1.880	1.880	1.882	1.882
241–255	**1.882**	**1.883**	**1.883**	**1.887**	**1.896**	**1.896**	**1.902**	**1.905**	**1.908**	**1.908**	**1.908**	**1.908**	**1.910**	**1.913**	**1.917**
256–270	**1.918**	**1.918**	**1.923**	**1.929**	**1.930**	**1.930**	**1.934**	**1.940**	**1.943**	**1.945**	**1.945**	**1.954**	**1.954**	**1.959**	**1.959**
271–285	**1.964**	**1.973**	**1.978**	**1.978**	**1.978**	**1.980**	**1.980**	**1.981**	**1.984**	**1.984**	**1.986**	**1.986**	**1.986**	**1.987**	**1.987**
286–300	**1.989**	**1.990**	**1.992**	**1.992**	**1.993**	**1.994**	**1.994**	**1.995**	**1.995**	**1.997**	**1.997**	**1.998**	**1.998**	**1.998**	**1.998**

Table 7
Parameters setting in GPHC.

Parameter	I	N	E	Pc	Pm	M	L	δ	τ	K
Value	100	40	15	0.95	0.10	200	2	0.60	0.30	15

entropy are shown in Table 6. Then, 20% CRD with the highest entropy value is derived for experts to discuss to determine the entropy threshold ε. 20% CRD with the highest entropy values, in this case, has been shown in Table 6 in bold. After the analysis of CRD with high entropy by experts, the ε is finally determined to be 1.96. That is to say, experts believe that the 271st–300th CRD does not contain obvious CR intention, and should be deleted before GPHC.

4.2.2. Data transformation

After data preprocessing, 270 CRD are left. Through equidistant-discrete processing, these data can be converted into 270 CRFs by Eq. (3). Then, the CRSF can be obtained by Eq. (4).

4.2.3. GPHC

The setting of parameters in GPHC is shown in Table 7. The specific operations are divided into three steps.

(1) Implement NGA to search peak points of CRSF

Repeatedly run NGA 200 (M) times to search for the peak of CRSF to obtain the *peak point set*.

All parameters are set as follows. The iteration times I is set to 100 in each NGA running. The population N and the number of retained elites E are 40 and 15 in each generation, respectively. The crossover probability Pc is set to 0.95, and the mutation probability Pm is set to 0.10. Besides, the *Penalty* of the mechanism of eliminating a similar structure of NGA is set to a number close to 0, and the fitness threshold δ is set to 0.60.

(2) Data aggregation

Traverse all data in the *peak point set*, merging the data points with Euclidean distance below 0.30 (i.e., τ) into one group, thereby obtaining 613 peak points.

(3) Implement AGNES to obtain customer preference pattern

There are four CR items in CRD, so T_{max} (the maximum number of CR items combination) is 15, which is calculated by Eq. (8). For the convenience of later discussion, Table 8 shows all potential combinations. For the sake of simplicity, the 6th column in Table 8 defines the notation of each combination.

According to expert experience and the selection rule of K in Section 3.5, K is set to 15. Finally, the AGNES algorithm is used to aggregate 613 peak points into 15 clusters.

(4) Implement k-means guided by customer preference pattern to cluster CRD

Each cluster centroid (i.e., *customer preference pattern*) can be calculated by Eq. (7). The columns from the 2nd one to the 5th one in Table 9 give the cluster centroids of 15 clusters. For subsequent analysis, the experts are asked to label these *customer preference patterns* by using the notation in Table 8. The labels of each cluster are shown in the last column of Table 9. "=" is used to represent a category consisting of data that has no analytical value (including fuzzy CRD and some CRD with no obvious intention for these CR items). As shown in rows 8th and

Table 8
The 15 potential combinations of CR items.

Order	Quality	Reserve capacity	Delivery cycle	Price	Notation
1	√				**Q**
2		√			**Rc**
3			√		**Dc**
4				√	**P**
5	√	√			**Q & Rc**
6	√		√		**Q & Dc**
7	√			√	**Q & P**
8		√	√		**Rc & Dc**
9		√		√	**Rc & P**
10			√	√	**Dc & P**
11	√	√	√		**Q, Rc & Dc**
12	√	√		√	**Q, Rc & P**
13	√		√	√	**Q, Dc & P**
14		√	√	√	**Rc, Dc & P**
15	√	√	√	√	**Q, Rc, Dc & P**

Note: "√" represents customer cares about corresponding CR item.

Table 9
The customer preference pattern and corresponding CR intention.

Order	Quality	Reserve capacity	Delivery cycle	Price	CR intention
1	2.805	**7.600**	**7.825**	**8.670**	**Rc, Dc & P**
2	**9.123**	4.261	**8.093**	**8.076**	**Q, Dc & P**
3	**9.692**	0.686	2.575	**5.587**	**Q & P**
4	**6.725**	2.652	2.639	**6.451**	**Q & P**
5	0.681	**9.318**	**9.573**	3.703	**Rc & Dc**
6	3.637	2.779	**7.061**	**5.002**	**Dc & P**
7	**5.236**	**5.603**	2.841	2.949	**Q & Rc**
8	1.053	0.921	2.049	3.362	=
9	2.903	**8.536**	2.587	2.968	**Rc**
10	1.988	0.500	**7.709**	1.286	**Dc**
11	**8.900**	**7.157**	2.166	1.444	**Q & Rc**
12	**7.758**	2.315	**6.551**	2.807	**Q & Dc**
13	**8.050**	2.941	2.959	2.329	**Q**
14	2.845	3.168	3.693	1.861	=
15	2.729	**6.617**	**6.700**	2.587	**Rc & Dc**

Note: The value in bold on each row corresponds to its CR intention.

Table 10
The 15 centroids of classes output by *k*-means.

Order	Quality	Reserve capacity	Delivery cycle	Price	CR intention
1	2.750	**8.417**	**8.542**	**9.000**	**Rc, Dc & P**
2	**9.000**	3.300	**8.300**	**8.500**	**Q, Dc & P**
3	**8.643**	1.214	1.143	**6.429**	**Q & P**
4	**7.105**	2.289	2.368	**7.789**	**Q & P**
5	2.360	**7.760**	**8.620**	2.420	**Rc & Dc**
6	2.300	2.100	**7.425**	**7.025**	**Dc & P**
7	**5.667**	**6.000**	2.500	2.667	**Q & Rc**
8	1.000	1.583	2.667	3.833	=
9	3.500	**8.781**	2.531	2.000	**Rc**
10	2.667	2.167	**8.500**	1.625	**Dc**
11	**8.761**	**7.478**	1.913	1.500	**Q & Rc**
12	**7.471**	2.059	**7.382**	1.779	**Q & Dc**
13	**8.556**	2.500	2.583	1.972	**Q**
14	2.846	2.654	2.885	1.423	=
15	2.074	**6.868**	**6.279**	1.706	**Rc & Dc**

Note: The value in bold on each row corresponds to its CR intention.

14th of Table 9, their scores of four CR items are all low values. Therefore, experts mark with "=".

Then, the heuristics information is used to cluster 270 CRD. This means that *customer preference patterns* in Table 9 are used as the initial clustering centroids of *k*-means. The mean value of each dimension of each CRD is used to represent the original interval data as the data input of *k*-means. Finally, 270 CRD are merged into 15 CRD classes. Table 10 shows the centroids of 15 final classes output by *k*-means. Fig. 7 shows the number of CRD in each CRD class. Compared Table 9 with Table 10, it can be found that these centroids of the final classes are very close to the initial cluster centroids. Furthermore, *k*-means only slightly revised the initial cluster centroids in its iterative process. It shows that it is effective to use *customer preference patterns* as heuristic information to guide clustering CRD.

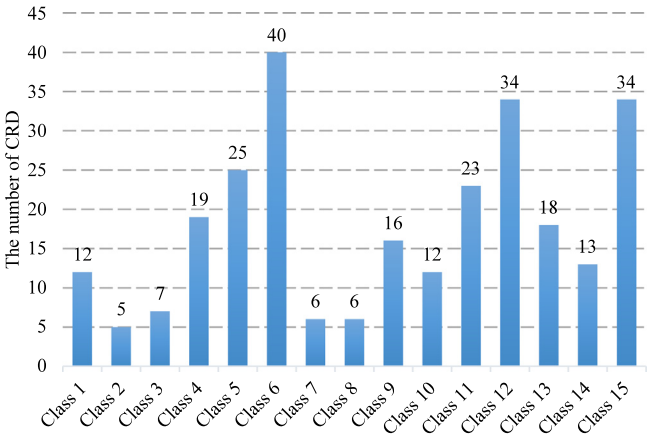

Fig. 7. The number of CRD corresponding to different CRD classes.

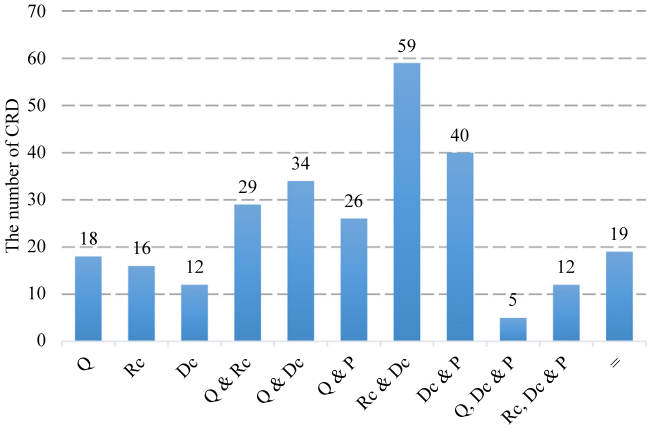

Fig. 8. The total number of CRD corresponding to different CR intentions.

(5) Matching analysis

For purpose of customer segmentation, the same CR intentions between different CRD classes should be combined. Table 10 also gives the expert label of each class. Then, those classes with the same CR intention (i.e., the same label) can be combined. Matching analysis is a relatively simple task. The expert only needs to label the given *K* CRD classes, and then merge the same labeled classes. According to matching analysis, the total CRD number of each CR intention can be further obtained, which is shown in Fig. 8.

4.3. Algorithm performance and practice comparison

GPHC consists of two targets. First, provide enterprise decision-makers with clear customer segmentation results to help accurately understand customer intention; (2) use computational intelligence to liberate human work on classification and clustering of customer data, and allow business operators to have more time to analyzing how to design market activities instead of data processing. Therefore, the performance of GPHC in practical customer segmentation deserves more attention. In short, two questions need to be answered: (1) Does GPHC provide clear customer segmentation results? (2) Do the customer segmentation results provided by GPHC help solve practical problems? The

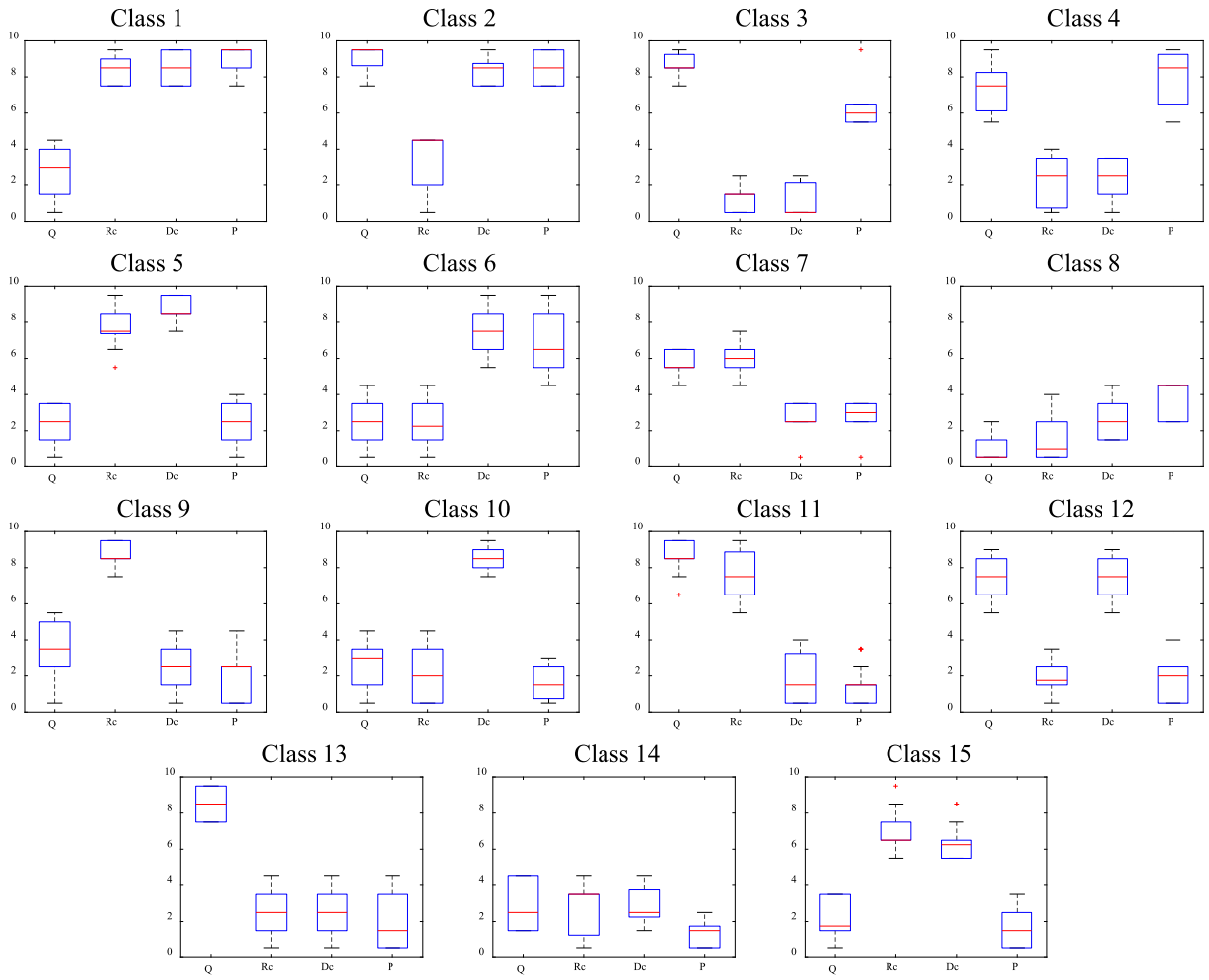

Fig. 9. The boxplots of CRD classes obtained by GPHC.

following gives an analysis both from algorithm performance and practice comparison.

4.3.1. Algorithm performance

To verify the performance of the proposed method in the customer segmentation problem, Fig. 9 illustrates the boxplot of 15 CRD classes obtained in Section 4.2.3. It can be seen that the relative positions of the four boxes are very clear in any boxplot. In other words, each class is made up of these CRD which belong to the same CR intention or these CRD with lower scores in all dimensions of CR items (e.g., Classes 8 and 14). GPHC provides clear customer segmentation results.

4.3.2. Practice comparison

To analyze the practice performance of GPHC, experts are asked to manually label the 300 CRD. The label can be selected from the notation of CR intention in Table 8.

Fig. 10 shows the results of customer segmentation of GPHC and the results provided by expert manual labeling. Due to the introduction of expert judgment, the customer segmentation results provided by manual labeling can be considered as the real label of CRD. It can be seen that the results from GPHC and expert experience are very close.

Similar results illustrate that the results given by GPHC are equivalent to those given by expert experience. It shows the effectiveness of GPHC in practical application.

The intuitive results of customer segmentation of GPHC in Fig. 10 can help enterprises quickly understand customer groups.

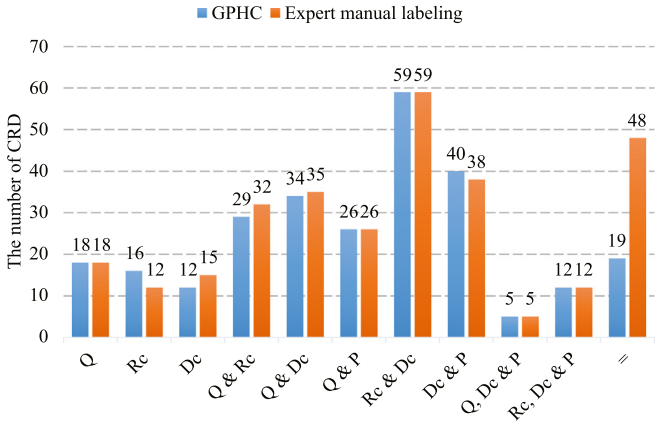

Fig. 10. Comparison between the customer segmentation results of GPHC and expert manual labeling. **Note**: The difference of 29 customers in "=" is since the enterprise does not use the entropy method to filter the neutral data, while GPHC previously deleted 30 neutral CRD.

For example, the following analysis is made based on Fig. 10. There are only 5 customers in the "Q, Dc & P". Therefore, this CR intention can be regarded as an outlier class that can be ignored in designing market activities with limited resources. But for "Rc", "Dc" and "Rc, Dc & P", they have 12, 15 and 12 customers

Table 11
The summary of the dataset using in Experiment 1.

CR	CR intention							
	1	2	3	4	5	6	7	8
CR1	√			√			√	
CR2	√	√			√			√
CR3		√	√			√	√	√
CR4		√		√		√	√	
Customers	510	460	420	20	560	540	480	20

Note: "√" represents the CRs that the customer cares about.

respectively. Considering that there are a certain number of customers, the enterprise may need to discuss whether customers corresponding to this CR intention should be met or not. Once having decided to respond to meet the requirement of such customers, to avoid affecting the final delivery cycle and stockout reducing customer loyalty, the enterprise must pay attention to the speed and yield of supply.

5. Numerical experiment

To illustrate the potential performance of GPHC, two numerical experiments are performed. Experiments 1 and 2 are designed to verify the performance of GPHC in high imbalanced CRD and high-dimensional CRD, respectively. It should be pointed out that the CRD used in this section is generated artificially by the authors. Therefore, CR intentions contained in the two datasets and the number of customers corresponding to each CR intention are known in advance. The detail of our method and other compared methods have been put on https://github.com/polysun/Customer-Segmentation.

5.1. Experiment 1: Discover customer preference pattern in a highly imbalanced scenario

GPHC is a heuristic clustering method, which heuristic information (i.e., *customer preference pattern*) is obtained through a combination of NGA and AGNES. Therefore, whether the *customer preference pattern* hidden in CRD can be completely discovered through NGA and AGNES plays a decisive role in the stability of GPHC.

In Experiment 1, 3010 CRD with 4-dimensional (i.e., 4 CR items) are generated. The collection of CRD exists highly imbalance in different CR intentions. Table 11 shows the detailed information of the 8 CR intentions contained in the dataset used in this experiment. It can be found that both the 4th and 8th CR intentions only have 20 customers, while the number of customers in other CR intentions is more than 20 times that of the 4th and 8th CR intentions.

Since the dimension of CR items of this experiment is consistent with the case study in Section 4, ε is set to 1.96 according to Section 4.2.1. Those CRD with entropy values higher than ε are deleted. Next, CRSF is constructed from 2894 retained CRD. Finally, GPHC is conducted. The setting of parameters is the same as Section 4.2.3.

Table 12 shows the 15 *customer preference patterns* discovered by GPHC. It can be found that GPHC has fully discovered 8 preset CR intentions including majority CR intentions and minority CR intentions. The experimental results illustrate the effectiveness of the combination of NGA and AGNES to search for *customer preference patterns* in the face of highly imbalanced CRD. It can provide a group of effective initial centroid points for the *k*-means.

Table 12
The 15 customer preference patterns discovered by GPHC in a highly imbalanced CRD scenario.

Order	Quality	Reserve capacity	Delivery cycle	Price	Belong to
1	**6.413**	3.566	2.469	**8.657**	CR intention 4
2	**8.980**	2.610	3.379	**7.603**	CR intention 4
3	**6.780**	3.722	**9.004**	**8.857**	CR intention 7
4	1.200	0.941	**5.597**	**6.633**	CR intention 6
5	0.437	**5.100**	**9.378**	**5.818**	CR intention 2
6	2.167	2.019	**8.794**	**6.851**	CR intention 6
7	**8.473**	2.356	**7.183**	**7.674**	CR intention 7
8	2.685	4.389	**7.114**	**7.526**	CR intention 6
9	2.852	**8.334**	**7.291**	**7.380**	CR intention 2
10	**5.066**	2.379	**7.294**	**6.863**	CR intention 7
11	2.750	**8.564**	**7.521**	3.364	CR intention 8
12	**8.588**	**5.825**	0.687	1.080	CR intention 1
13	1.304	1.614	**8.894**	1.814	CR intention 3
14	0.983	**6.570**	1.871	1.322	CR intention 5
15	4.167	2.523	**8.614**	1.154	CR intention 3

Note: The value in bold on each row corresponds to its CR intention.

5.2. Experiment 2: Obtain clear customer segmentation result in a highly dimensional scenario

In Experiment 2, 400 CRD with 20-dimensional CR items are generated, which include 17 CR intentions. Table 13 shows the detailed information of each CR intention. It can be found that the 4th CR intention in the dataset contains only 5 customers. Therefore, whether the 4th CR intention can be successfully discovered by GPHC is worth paying attention to. Experiment 2 gives a comparison between GPHC, *k*-means, and *k*-medoids. The difference between GPHC and *k*-means is whether heuristic information is used or not. The only parameter in *k*-means and *k*-medoids is *K*. Their values are consistent with the *K* setting in GPHC.

5.2.1. Boxplot comparison

To increase the difficulty of this problem, data preprocessing is not performed here. Table 14 shows the parameter settings of GPHC. The target cluster number of all methods is set to 20. Without affecting the analysis of the result, to save space, this paper only gives all boxplots output by *k*-means and a part of boxplots output by GPHC. Specifically, Fig. 11 shows boxplots of 20 classes output by *k*-means, and Fig. 12 shows a part of boxplots output by GPHC. It can find that the clustering result of *k*-means is confused. A part of boxplots output by *k*-means occupies the entire *y*-axis (such as classes 2, 7, and 15). 4th CR intention hidden in the CRD is also not discovered by *k*-means. In contrast, GPHC successfully discovered the 4th CR intention (Class 19) and other CR intentions with less than 20 customers.

5.2.2. Indicator comparison

Four common cluster validity indicators are used to compare the performance of GPHC with compared methods. These indicators are Dunn index [43] (DI, higher is better), Davies Bouldin index [44] (DBI, lower is better), Silhouette coefficient [45] (SI, higher is better), and Calinski–Harabasz index [46] (CH, higher is better), respectively. The experimental result takes the mean value of 20 runs. The experimental results are shown in Table 15. It can be seen that the scores of GPHC on DI, SI, and CH are higher than other comparison methods, and the score on DBI is lower than other comparison methods. It shows that the performance of GPHC outperforms other comparison methods.

5.3. Theoretical and practical implications

Through the above analysis, it illustrates that GPHC effectively performs the clustering task on scenarios of imbalance and highly

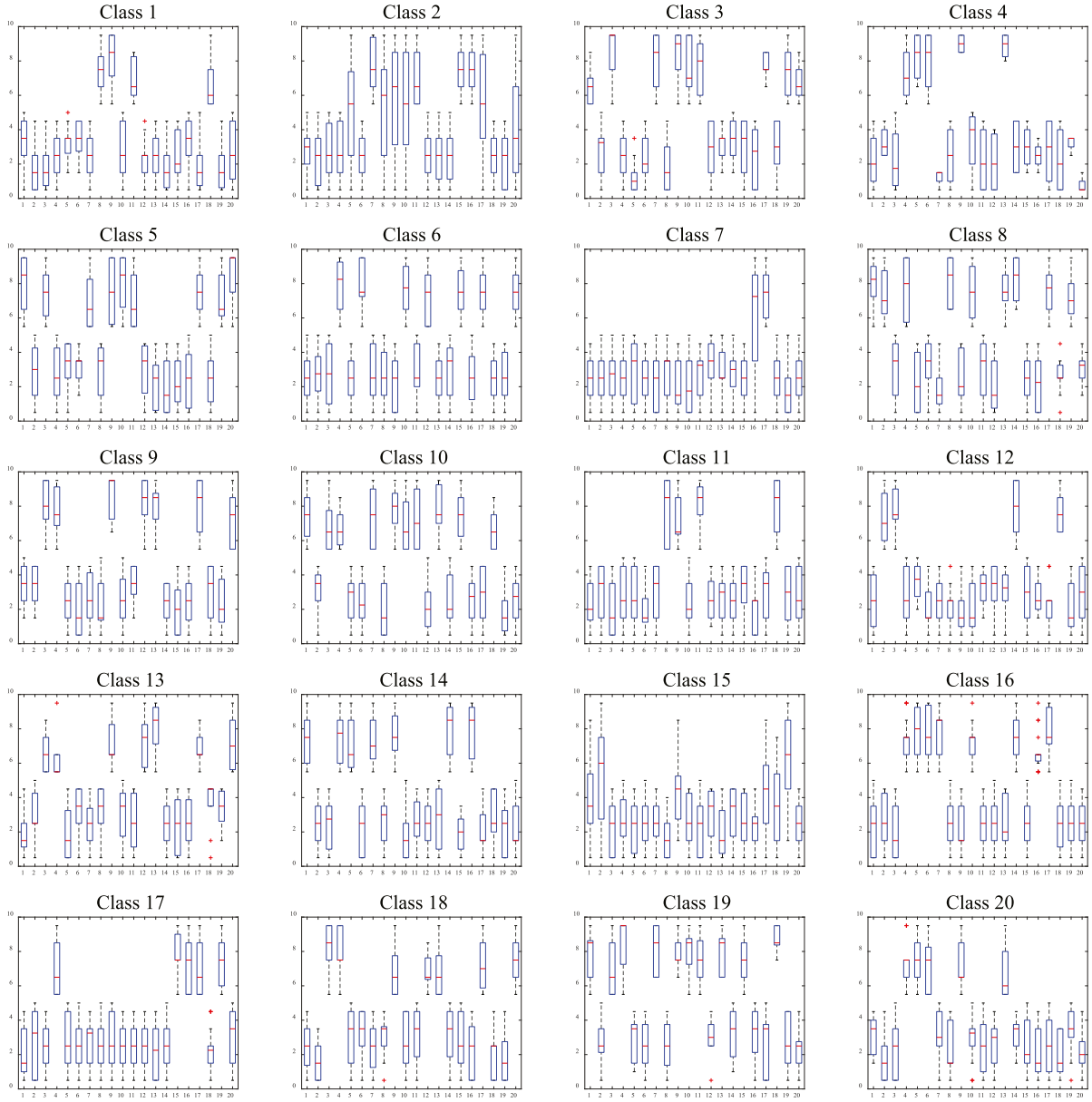

Fig. 11. All boxplots of CRD classes obtained by *k*-means in Experiment 2.

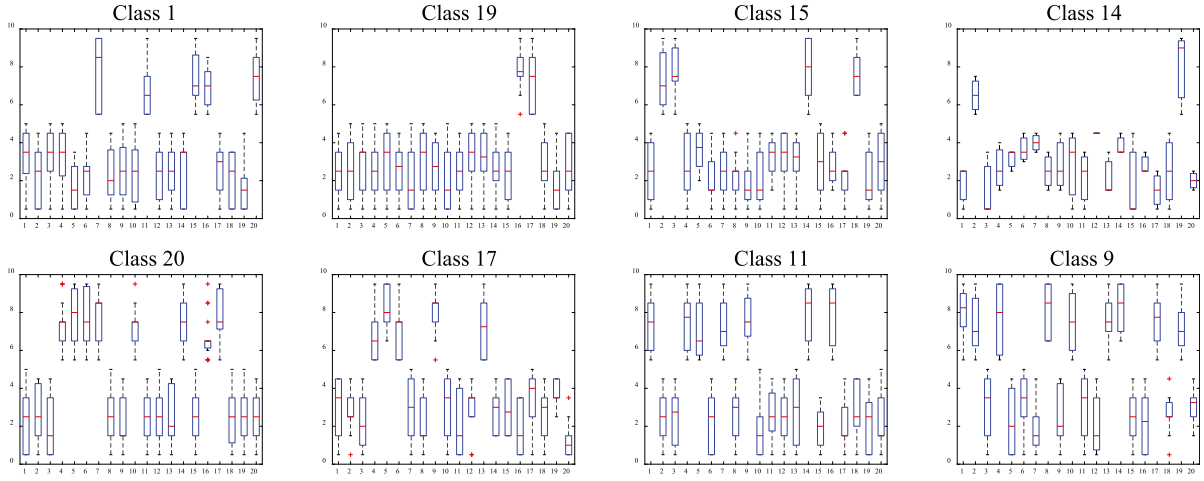

Fig. 12. A part of boxplots of CRD classes obtained by GPHC in Experiment 2.

Table 13
The summary of the dataset using in Experiment 2.

CR	CR intention																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
CR1			√						√	√					√	√	
CR2							√	√								√	
CR3					√		√		√	√							
CR4					√	√				√	√	√	√		√	√	
CR5	√											√	√		√		
CR6						√						√	√				
CR7	√	√							√	√		√			√		
CR8	√													√		√	
CR9	√		√		√				√	√			√	√	√		
CR10	√					√			√	√		√				√	
CR11	√	√							√	√				√			
CR12					√	√											
CR13					√					√			√			√	
CR14							√					√			√	√	
CR15	√	√				√				√	√						
CR16	√	√	√								√	√			√		
CR17	√		√	√	√	√			√		√	√				√	√
CR18				√			√			√				√			
CR19								√	√		√					√	
CR20		√			√	√			√								
Customers	26	17	32	5	41	32	12	11	33	29	34	19	20	40	16	12	21

Note: "√" represents the CR item that the customer cares about.

Table 14
Parameters setting in GPHC in Experiment 2.

Parameter	I	N	E	Pc	Pm	M	L	δ	τ	K
Value	1000	40	15	0.95	0.10	200	9	0.00001	0.50	20

Table 15
Indicator comparison between GPHC and other compared methods with DI, DBI, SI, and CH.

Methods	DI	DBI	SI	CH
k-means	0.30 ± 0.05	2.01 ± 0.14	0.22 ± 0.02	47.34 ± 2.80
k-medoids	0.37 ± 0.03	2.07 ± 0.07	0.24 ± 0.01	50.63 ± 0.20
GPHC	**0.44 ± 0.04**	**1.42 ± 0.11**	**0.29 ± 0.01**	**50.90 ± 2.51**

dimension, so that enterprise can accurately and easily perform customer segmentation under difficult data distribution scenarios. By GPHC, experts only need to simply judge the CR intentions expressed by each CRD class. After that, customer segmentation can be obtained, which results will be similar to manually labeling. The proposed GPHC in this paper greatly reduces the human resources and time cost of customer segmentation.

Through customer segmentation, enterprises could accurately divide customer groups into multiple communities with different requirement characteristics. It enables enterprises to better understand the overall situation of CRs for products and services in the market and further locks in those that can bring profits to enterprises. This is a necessary condition for long-term profitability and sustainable development of enterprise. The advent of the digital age means that more personalized requirements of customers are displayed in the form of data. Therefore, segmenting CRs based on CRD and providing more targeted products and services is also an emerging topic that a large number of companies are facing. This paper also hopes to provide a reference for enterprises to explore the supply–demand relationship in the digital age through the study of this topic.

6. Conclusion

This paper divides the customer segmentation problem into three stages based on interval CRD. The first stage is data preprocessing that uses entropy to filter neutral CRD. The second stage is data transformation. A data transformation scheme based on the standardized Gaussian distribution is proposed to model CRD, which retains the ambiguity of requirements expressed by the original interval CRD. The third stage is data clustering. A heuristic clustering method, termed GPHC, is proposed. This method uses heuristics information to accurately cluster CRD into different classes.

A practical study case is conducted to verify the feasibility and effectiveness of the proposed method. By analyzing the clustering results of GPHC, and comparing it with manual labeling by experts, our method shows potential advantages in customer segmentation. In numerical experiments, the performance of GPHC in face of complex data distribution has been verified.

Through the proposed solution (including data preprocessing, data transformation, and GPHC), the customer segmentation problem is solved systematically. Enterprises can directly use the proposed method to perform customer segmentation tasks, and further adjust market activity in a dynamic customer environment in time.

Although our method could effectively solve the customer segmentation problem, there is still much room for further research. For example, how to realize the parameter adaptation of GPHC to avoid manual setting. Also, as a new heuristic clustering method, the popularization and verification of this method in other fields need to be developed urgently.

CRediT authorship contribution statement

Zhao-Hui Sun: Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing, Supervision. **Tian-Yu Zuo:** Formal analysis, Methodology, Validation, Writing -

original draft. **Di Liang:** Writing - review & editing. **Xinguo Ming:** Funding acquisition. **Zhihua Chen:** Visualization. **Siqi Qiu:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the 2019 Industrial Internet Innovation Development Project from the Ministry of Industry and Information Technology of China under Grant TC190A3X8-10, the 2019 Shanghai Industrial Internet Innovation and Development Special Project from Shanghai Municipal Commission of Economy and Informatization, China under Grant 2019-GYhlW-01007, and the National Natural Science Foundation of China under Grant 71632008.

References

[1] Z. Zhang, H. Lin, K. Liu, D. Wu, G. Zhang, J. Lu, A hybrid fuzzy-based personalized recommender system for telecom products/services, Inform. Sci. 235 (2013) 117–129.
[2] P. Zheng, X. Xu, S.Q. Xie, Integrate product planning process of OKP companies in the cloud manufacturing environment, in: Proceedings of the International Federation for Information Processing 2015, IFIP, 2015, pp. 420–426.
[3] H. Li, Q. Jiao, X. Wen, G. Luo, C. Wu, Implementation solution planning methodology of enterprise product-service system oriented to customer demand, Comput. Integr. Manuf. 23 (8) (2017) 1750–1764.
[4] K. Khalili-Damghani, F. Abdi, 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, Appl. Soft Comput. 73 (2018) 816–828.
[5] A. Garai, T.K. Roy, Multi-objective optimization of cost-effective and customer-centric closed-loop supply chain management model in T-environment, Soft Comput. 24 (1) (2020) 155–175.
[6] V. Holý, O. Sokol, M. Černý, Clustering retail products based on customer behavior, Appl. Soft Comput. 60 (2017) 752–762.
[7] P. Zheng, X. Xu, S.Q. Xie, A weighted interval rough number based method to determine relative importance ratings of customer requirements in QFD product planning, J. Intell. Manuf. 30 (1) (2019) 3–16.
[8] E.G. Kyriakidis, T.D. Dimitrakos, C.C. Karamatsoukis, Optimal delivery of two similar products to N ordered customers with product preferences, Int. J. Prod. Econ. 209 (2019) 194–204.
[9] W. Luo, Y. Qiao, X. Lin, P. Xu, M. Preuss, Hybridizing niching, particle swarm optimization, and evolution strategy for multimodal optimization, IEEE Trans. Cybern. (2020) 1–14.
[10] H. Chen, L. Zhang, X. Chu, B. Yan, Smartphone customer segmentation based on the usage pattern, Adv. Eng. Inf. 42 (2019).
[11] H. Zeybek, Customer segmentation strategy for rail freight market: the case of turkish state railways, Res. Transp. Bus. Manag. 28 (2018) 45–53.
[12] O. Dzobo, K. Alvehag, C.T. Gaunt, R. Herman, Multi-dimensional customer segmentation model for power system reliability-worth analysis, Int. J. Electr. Power Energy Syst. 62 (2014) 532–539.
[13] T.B. Maria, G.C. Pilar, J. Sainz, Customer segmentation in e-commerce: applications to the cashback business model, J. Bus. Res. 88 (2018) 407–414.
[14] S. Nakano, F.N. Kondo, Customer segmentation with purchase channels and media touchpoints using single source panel data, J. Retail. Consum. Serv. 41 (2018) 142–152.
[15] J. Xu, J. Han, K. Xiong, F. Nie, Robust and sparse fuzzy K-means clustering, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16, 2016, pp. 2224–2230.
[16] P.W. Murray, B. Agard, M.A. Barajas, Forecasting supply chain demand by clustering customers, IFAC-PapersOnLine 48 (3) (2016) 1834–1839.
[17] M. Mohammadzadeh, Z. Zare Hoseini, H. Derafshi, A data mining approach for modeling churn behavior via RFM model in specialized clinics case study: A public sector hospital in tehran, Procedia Comput. Sci. 120 (2017) 23–30.
[18] K. Khalili-Damghani, F. Abdi, 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, Appl. Soft Comput. 73 (2018) 816–828.
[19] N. Sano, R. Tsutsui, K. Yada, T. Suzuki, Clustering of customer shopping paths in Japanese grocery stores, Procedia Comput. Sci. 96 (2016) 1314–1322.
[20] I. Bose, X. Chen, Detecting the migration of mobile service customers using fuzzy clustering, Inf. Manage. 52 (2015) 227–238.
[21] J. Llanos, R. Morales, A. Núñez, D. Sáez, M. Lacalle, L.G. Marín, R. Hernández, F. Lanas, Load estimation for microgrid planning based on a self-organizing map methodology, Appl. Soft Comput. 53 (2017) 323–335.
[22] T. Hong, E. Kim, Segmenting customers in online stores based on factors that affect the customer's intention to purchase, Expert Syst. Appl. 39 (2012) 2127–2131.
[23] A. Seret, S. Maldonado, B. Baesens, Identifying next relevant variables for segmentation by using feature selection approaches, Expert Syst. Appl. 42 (2015) 6255–6266.
[24] A. Dursun, M. Caber, Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis, Tourism Manage. Perspect. 18 (2016) 153–160.
[25] J.T. Wei, S.Y. Lin, C.C. Weng, H.H. Wu, A case study of applying LRFM model in market segmentation of a children's dental clinic, Expert Syst. Appl. 39 (5) (2012) 5529–5533.
[26] J.T. Wei, M.C. Lee, H.K. Chen, H.H. Wu, Customer relationship management in the hairdressing industry: An application of data mining techniques, Expert Syst. Appl. 40 (2013) 7513–7518.
[27] C.H. Cheng, Y.S. Chen, Classifying the segmentation of customer value via RFM model and RS theory, Expert Syst. Appl. 36 (2009) 4176–4184.
[28] G.T.S. Ho, W.H. Ip, C.K.M. Lee, W.L. Mou, Customer grouping for better resources allocation using GA based clustering technique, Expert Syst. Appl. 39 (2012) 1979–1987.
[29] C.-Y. Chiu, Y.-F. Chen, I.-T. Kuo, H.C. Ku, An intelligent market segmentation system using k-means and particle swarm optimization, Expert Syst. Appl. 36 (2009) 4558–4565.
[30] K. Krishna, M.N. Murty, Genetic K-means algorithm, IEEE Trans. Syst. Man Cybern. PartB: Cybern. 29 (1999) 433–439.
[31] C. Zhang, D. Ouyang, J. Ning, An artificial bee colony approach for clustering, Exp. Syst. Appl. 37 (2010) 4761–4767.
[32] P. Luo, S. Yan, Z. Liu, Z. Shen, S. Yang, Q. He, From online behaviors to offline retailing, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 175–184.
[33] O. Taylan, A hybrid methodology of fuzzy grey relation for determining multi attribute customer preferences of edible oil, Appl. Soft Comput. 13 (5) (2013) 2981–2989.
[34] X. Zhang, C. Mei, D. Chen, J. Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[35] R.M. Souza, F.D.A. Carvalho, Clustering of interval data based on city–block distances, Pattern Recognit. Lett. 25 (3) (2014) 353–365.
[36] Z. Lu, H. Jin, P. Yuan, D. Zou, A fuzzy clustering algorithm for interval-valued data based on Gauss distribution functions, Acta Electron. Sin. 38 (2) (2010) 295–300.
[37] P. D'Urso, J.M. Leski, Fuzzy c-ordered medoids clustering for interval-valued data, Pattern Recognit. 58 (2016) 49–67.
[38] X. Lin, W. Luo, P. Xu, Differential evolution for multimodal optimization with species by nearest-better clustering, IEEE Trans. Cybern. 51 (2) (2019) 970–983.
[39] G.T. Ho, W.H. Ip, C.K.M. Lee, W.L. Mou, Customer grouping for better resources allocation using GA based clustering technique, Expert Syst. Appl. 39 (2) (2012) 1979–1987.
[40] N.N. Glibovets, N.M. Gulayeva, A review of niching genetic algorithms for multimodal function optimization, Cybern. Syst. Anal. 49 (6) (2013) 815–820.
[41] Y. Liang, K.S. Leung, Genetic algorithm with adaptive elitist-population strategies for multimodal function optimization, Appl. Soft Comput. 11 (2) (2011) 2017–2034.
[42] A. Chitra, A. Rajkumar, Paraphrase extraction using fuzzy hierarchical clustering, Appl. Soft Comput. 34 (2015) 426–437.
[43] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (3) (1973) 32–57.
[44] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[45] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1) (1987) 53–65.
[46] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat.-Theory Methods 3 (1) (1974) 1–27.