Italian Journal of Food Science, 2024; 36 (4): 38–56

Multiple-perspective consumer segmentation using improved weighted Fuzzy k-prototypes clustering and swarm intelligence algorithm for fresh apricot market

Yan Shi[1], Siyuan Zhang[1], Siwen Wang[1], Hui Xie[2], Jianying Feng[1]*

[1]College of Information and Electrical Engineering, China Agricultural University, Beijing, China; [2]Research Institute of Horticulture, Xinjiang Academy of Agricultural Sciences, Xinjiang, China

***Corresponding Author:** Jianying Feng, College of Information and Electrical Engineering, China Agricultural University, No. 17 Qinghua East Road, Haidian District, Beijing 100083, China. Email: fjying@cau.edu.cn

Received: 5 March 2024; Accepted: 22 June 2024; Published: 26 September 2024

ORIGINAL ARTICLE

Abstract

Leveraging clustering technology for consumer segmentation is crucial for discerning the nuanced differences among fresh apricot consumer groups and subsequently executing precise marketing strategies. To achieve a more comprehensive and lucid consumer segmentation and identify typical characteristics of apricot consumers in different clusters, this research constructs a novel multiple-perspective segmentation indicator system for fresh apricot consumers. Given the diverse degrees of importance and types of consumer segmentation variables, and the inherent sensitivity of the original Fuzzy k-prototypes (FKP) algorithm to clustering centers, we proposed the weighted Fuzzy k-prototypes (WFKP) algorithms for mixed data (MD) optimized by the particle swarm optimization (PSO) algorithm (MDPSO-WFKP) and mixed data sparrow search algorithm (SSA) (MDSSA-WFKP), both incorporating information entropy weighting for mixed attributes. We test the proposed algorithms on four University of California Irvine machine learning repository (UCI) datasets and the consumer segmentation dataset, and the performance of all selected evaluation indexes shows significant improvement. These findings unequivocally validate the efficacy of the proposed methodologies. Since the MDSSA-WFKP algorithm has the best comprehensive effect on the evaluation indexes, we use it to conduct in-depth apricot consumer segmentation research and find that the apricot consumers can be subdivided into three groups with differentiation: 'Buddhist-like youths', 'Upscale attribute enthusiasts', and 'Quality-oriented consumers'. Finally, this paper gives the corresponding marketing suggestions based on the characteristics of the segmented groups.

Keywords: consumer segmentation; cluster analysis; MDPSO-WFKP algorithm; MDSSA-WFKP algorithm; precision marketing

Introduction

Promoting consumption is one of the key strategies for increasing economic development. China is undergoing a new stage of consumption upgrading and transformation, and consumers' demands for agricultural products show a trend of high-quality, diversified, and personalized requirements (Shi *et al.*, 2022). Many visionary enterprises nowadays commonly adopt consumer segmentation to provide personalized products and services to satisfy different types of consumers and enhance market competitiveness (Lee *et al.*, 2019). Consumer segmentation usually refers to the process of dividing consumers into several sub-groups with significant differences based on their demographic, geographical, behavioral, psychological characteristics, etc. (Bannor *et al.*, 2022, Wang and Scrimgeour, 2023). Common segmentation methods can be summarized as *a priori* segmentation and *post hoc* segmentation; *post hoc* segmentation has quite crucial research value and practical significance because of

 ISSN 1120-1770 online, DOI 10.15586/ijfs.v36i4.2552

its advantages of being more objective, comprehensive, and scientific (Kazbare *et al.*, 2010, Tohidi *et al.*, 2023). The fresh apricot enjoys widespread popularity among Chinese consumers. However, with the increasing supply and consumer demand, the market competition has become fierce and the difficulty of sales has developed. Therefore, stakeholders urgently need to understand consumers and their demands in a better manner. Fortunately, consumer segmentation research based on clustering algorithms can provide marketers with marketing decision-making advice.

From the perspective of data mining, *post hoc* segmentation is essentially a kind of data clustering analysis. As an effective data analysis tool, data mining technology based on cluster analysis is applied to the field of consumer segmentation. It has obvious advantages over the traditional segmentation method with standards for a single perspective (Abbasimehr and Bahrini, 2022; Sun *et al.*, 2021). In general, common clustering algorithms are classified into partition-based (Li *et al.*, 2021, Mollaei *et al.*, 2023), hierarchy-based (Bejaei *et al.*, 2020, Kuesten *et al.*, 2022), density-based (Bhattacharjee and Mitra, 2020, Hegazi *et al.*, 2021), grid-based (Du and Wu, 2022), model-based (Ghadiri and Mazlumi, 2020, Weber *et al.*, 2022), and modern clustering (Ezugwu *et al.*, 2022) algorithms. In the field of consumer segmentation of agricultural products, the clustering methods presently used are mostly simple and efficient partition-based (Chen *et al.*, 2021, Guo *et al.*, 2019) and hierarchy-based (Bejaei *et al.*, 2020) methods of traditional clustering.

Based on the preference characteristics of fresh corn consumption by urban residents in Beijing, Guo *et al.* (2019) adopted K-means to subdivide consumers after using factor analysis to extract three common factors: material demand, functional demand, and spiritual demand. Chen *et al.* (2021) obtained wine online review data based on text mining and applied K-means clustering to divide consumers into four segment markets. In addition, Kuesten *et al.* (2022) conducted agglomerative hierarchical clustering based on psychographic data, and segmented consumers into two classes with higher versus lower General self-efficacy. Although existing research provided methodological references and practical experience for consumer segmentation, they did not meet the demand for market segmentation under specific conditions. On the one hand, most of the consumer's characteristic attributes are mixed-type data, while the existing research mostly focuses on single-type data clustering, and the traditional clustering algorithm is easy to fail in the clustering analysis of mixed data. On the other hand, the traditional hard partition method makes it difficult to divide objects with fuzzy membership substantially, but in fact, consumers have uncertainty about their correspondence to specific subclasses. Therefore, it is particularly important to choose a suitable clustering algorithm for consumer segmentation, especially for consumer segmentation in the specific market condition from multiple-perspective characteristic attributes.

Fuzzy K-prototypes is a soft partitioning clustering algorithm based on K-prototypes, which can solve the mixed data clustering task by introducing the concepts of fuzzy parameter and membership, and it has shown excellent results in clustering mixed attribute data (Chen *et al.*, 2001). However, it still has some problems, such as sensitivity to initial clustering centers, a large influence of K value on clustering results, and the neglect of considering the unequal importance of variables. The existing studies tend to improve the algorithm only from a single perspective and the improvement effect is limited (Ouyang *et al.*, 2015, Ye and Liang, 2010). Consequently, we tried to optimize the Fuzzy k-prototypes from two aspects: variable weighted, and the initial clustering center determination, and proposed the mixed data (MD) optimized by the particle swarm optimization (PSO)–weighted Fuzzy k-prototypes (MDPSO-WFKP) and mixed data sparrow search algorithm (SSA)–weighted Fuzzy k-prototypes (MDSSA-WFKP) algorithms. It can not only meet the consumer segmentation in the mixed data context but also improve the segmentation effect based on the soft partitioning clustering algorithm, the Fuzzy k-prototypes algorithm, to improve the limitations of the above-mentioned algorithms.

The purpose of this research was to put forward a new and effective clustering algorithm to realize the market segmentation of fresh apricot consumers in China. To improve the accuracy of segmentation, this paper first constructed a multiple-perspective consumer segmentation indicator system. Then we proposed two improved Fuzzy k-prototypes algorithms for mixed data (MDPSO-WFKP and MDSSA-WFKP) that emphasized differences in the importance of segmentation variables and optimized the selection of clustering centers. After verifying the superiority of the proposed algorithms in four public datasets, the proposed methods were applied in apricot consumer clustering. Theoretically, we forwarded an optimized clustering algorithm to improve the clustering effect for mixed data. Practically, reasonable consumer segmentation could help stakeholders in the industry to better understand consumers' needs and preferences and finally promote fine management and precise marketing.

Materials and Methods

Establishment of empirical dataset

Data acquisition

There is no public dataset on the characteristics and purchasing preferences of fresh apricot consumers in

China. Therefore, we obtained data through a consumer questionnaire survey and finally established a segmentation dataset of fresh apricot consumers. We conducted a nationwide survey of Chinese fresh apricot consumers. According to the results of the seventh national census in China, a target sample size was designed for each of the 31 provinces, municipalities, and autonomous regions, based on a ratio of approximately three parts per million of the resident population in each region, and a random sampling survey was conducted. To expand the sample size and consider the requirements of epidemic prevention and control, 3,782 questionnaires were distributed through the online questionnaire platform, and 3,666 valid questionnaires were obtained after eliminating invalid and duplicate questionnaires, resulting in a sample return efficiency of 96.63%. Table 1 shows the distribution of characteristics of research sample. Finally, a fresh apricot consumer dataset containing 3,666 cases was established.

Data preprocessing

Different characteristic variables have different magnitudes. To reduce the relative relationship between magnitudes and eliminate the influence of magnitudes between characteristic variables, data normalization is needed (Singh and Singh, 2020). One of the commonly used and effective methods is the minimum–maximum normalization (MMN) method, also known as linear normalization or outlier normalization. It is a linear transformation of the original data and maps the result to the interval [0,1] (Kiran and Vasumathi, 2020). Its equation is as follows:

$$x'_{il} = \frac{x_{il} - x_{lmin}}{x_{lmax} - x_{lmin}} \quad (1)$$

where x_{il} represents the lth characteristic variable value of the ith sample, is the minimum value of the lth characteristic variable, and x_{lmax} is the maximum value of the lth characteristic variable.

Since both ordinal categorical variables and numerical variables have 'quantitative' differences, this study normalize ordinal categorical variables and numerical variables uniformly and treat them in a similar manner as numerical variables in the subdivision process, which are not pointed out separately later.

Construction of multiple-perspective consumer segmentation indicator system

Based on the theory of consumer behavior and related studies (Bannor et al., 2022, Park *et al.*, 2020), this paper establishes a fresh apricot consumer segmentation indicator system from five dimensions, such as demographic characteristics, geographic characteristics, behavioral characteristics, psychological characteristics, and cognition degree of product knowledge. Specifically, the

Table 1. Distribution of characteristic variables of research sample of survey questionnaire.

Characteristics	Percentage (%) (n = 3,666)	Characteristics	Percentage (%) (n = 3,666)
Gender		**Education level**	
Males	41.76	Bachelor's degree or above	60.09
Females	58.24	Junior college	19.04
Age		High school (including vocational middle school, vocational high school)	10.56
<20 years	4.01	Junior high school	7.31
20–28 years	48.31	Primary school and below	3.00
29–35 years	22.04	**Family situation**	
36–55 years	24.36	Single	49.15
>55 years	1.28	Married, no children ≤14 years old	19.18
Per capita monthly income of the family (yuan)		Married, with children ≤14 years old	31.67
<2,000	16.07	**Number of family members**	
2,000–3,000	14.98	<3	11.51
3,001–5,000	20.02	3–4	63.34
5,001–7,000	17.65	≥5	25.15
7,001–10,000	14.62	**Type of permanent residence**	
10,001–15,000	9.90	Town	70.79
>15,000	6.76	Village	29.21

demographic characteristics contain six original variables, such as gender, age, etc. The geographic characteristics contain two variables: type of permanent residence and area of permanent residence. The behavioral characteristics include two variables: purchase frequency and purchasing power. The psychological characteristics include seven original variables, which are sensitivities to price, freshness, and other attributes of fresh apricots. The cognition degree of product knowledge includes one variable.

In clustering analysis, too many clustering variables increase the complexity of the algorithm, because more running overhead, and easily lead to poor interpretability of results. Besides, there may be a certain degree of correlation between the relevant indicators but not a complete correlation. Therefore, directly deleting some of the indicators to achieve the purpose of reducing variables lead to a complete loss of information on some indicators (Hasan and Abdulazeez, 2021). Thus, the application of an appropriate feature extraction method to compress and transform clustering variables is helpful to reduce the dimension of clustering variables, improve the efficiency of clustering algorithms as well as further discover the potential structure between variables. In this research, the demographic characteristics of consumers are not replaced by each other, the geographic characteristics, behavioral characteristics, and cognition degree of product knowledge only have one or two variables; hence, it is not necessary to reduce the dimension.

However, there are seven original variables in psychological characteristics, the number is large, and there may be correlations among these variables. Therefore, factor analysis was used to reduce the dimension and transform original variables, then a few comprehensive indexes were extracted from original variables, and specific variable names were given according to the connotation of indicators. Such comprehensive variable indicators are independent of each other and do not overlap with each other (Taherdoost *et al.*, 2022).

After factor analysis, the cumulative variance contribution rate of four common factors was 84%, indicating that the seven original variables could be well explained by four common factors. The results of naming common factors according to the loading coefficient are shown in Figure 1.

So far, a multiple-perspective consumer segmentation indicator system was constructed; the clustering indicators and their attributes for consumer segmentation are shown in Table 2.

Proposed methods

In terms of multiple clustering variables, the entropy-weighted method was used to assign weights to variables to measure the unequal importance of variables and optimize the distance metric formula. As for the clustering

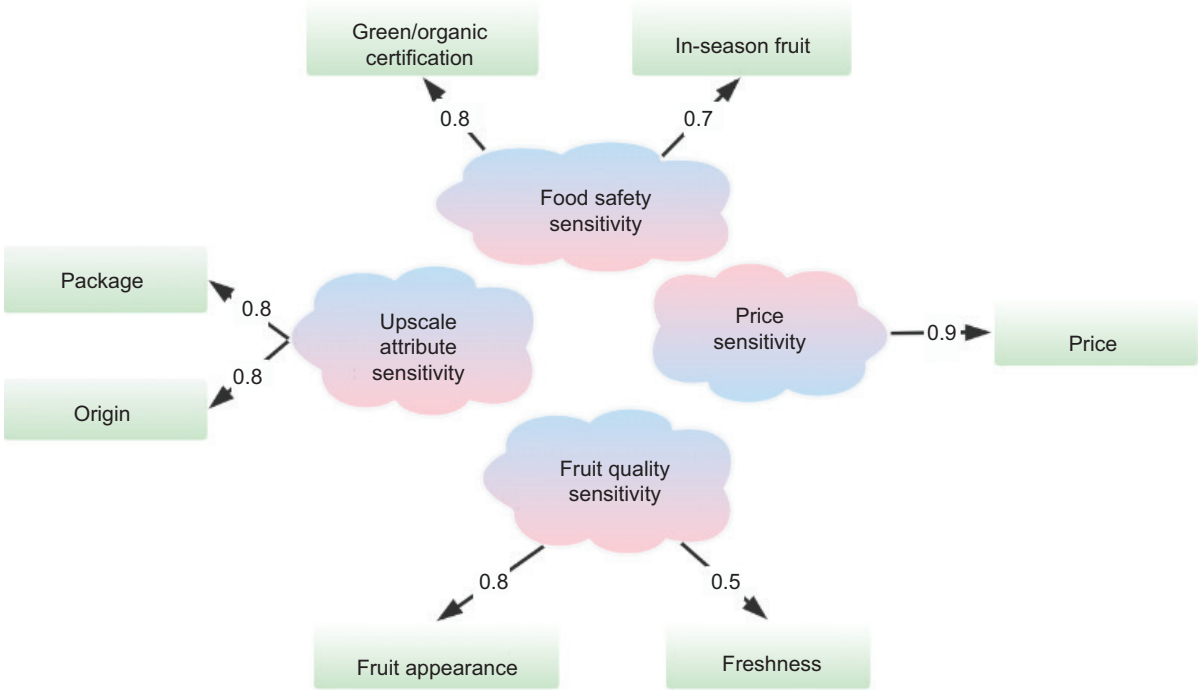

Figure 1. Factor loading diagram after factor analysis.

Table 2. Dataset variables and their attributes for consumer segmentation.

Variable	Types of variables
Age (A_1)	Numerical
Number of family members (A_2)	Numerical
Food safety sensitivity (A_3)	Numerical
Price sensitivity (A_4)	Numerical
Fruit quality sensitivity (A_5)	Numerical
Upscale attribute sensitivity (A_6)	Numerical
Cognition degree of product knowledge (A_7)	Numerical
Purchase frequency (A_8)	Ordinal categorical
Purchasing power (A_9)	Ordinal categorical
Educational level (A_{10})	Ordinal categorical
Per capita monthly income of the family (A_{11})	Ordinal categorical
Gender (A_{12})	Nominal categorical
Family situation (A_{13})	Nominal categorical
Type of permanent residence (A_{14})	Nominal categorical
Area of permanent residence (A_{15})	Nominal categorical

center selection, the traditional particle swarm optimization (PSO) algorithm and sparrow search algorithm (SSA) were not applicable to the optimization of mixed data. Therefore, based on the actual situation of the optimized target problem, initial clustering center determination for mixed data, we purposely proposed discrete attribute and boundary processing methods to apply to the processing of mixed data and perturb the population to a certain extent to avoid falling into the local optimum. Consequently, MDPSO as well as MDSSA were used for optimizing the selection of initial clustering centers for the Fuzzy k-prototypes algorithm.

Theoretical basis of Fuzzy k-prototypes

In this paper, we define the set of attributes as $A = \{A_1, A_2,\ldots,A_m\}$, and assume that $X = \{x_1,x_2,\ldots,x_N\}$ is a set of sample objects with mixed attributes, where $x_i = \{x_{i1},\ldots, x_{is},x_{i(s+1)},\ldots,x_{im}\}$ denotes a sample object with m attributes, and the first s attributes are numerical, while the s+1th to mth attributes are categorical attributes. The distance metric between the sample objects x_i and x_j is defined as follows:

$$d(x_i,x_j)=\sum_{i-1}^{s}(x_{il}-x_{jl})^2+\lambda\sum_{l=s+1}^{m}\delta(x_{il},x_{jl}) \quad (2)$$

where the first half denotes the distance measure of continuous attributes, while the second half denotes the distance measure of categorical attributes; λ is the weight that regulates the distance ratio of two types of attributes, called weight parameter; and the distance measure of categorical attributes is defined as shown in Eq. (2):

$$\delta(x_{il},x_{jl})=\begin{cases}1, x_{il}\neq x_{jl}\\ 0, x_{il}=x_{jl}\end{cases} \quad (3)$$

Assuming k is a positive integer, the objective of clustering X is to divide N samples into k different clusters, the objective function is as follows:

$$F(U,V)=\sum_{c=1}^{k}\sum_{i=1}^{N}(u_{ic})d(x_i,v_c) \quad (4)$$

where $u_{ic}\in[0,1]$, $\sum_{c=1}^{k}u_{ic}=1, 0<\sum_{i=1}^{N}u_{ic}<N$; U is the membership matrix; N is the sample number of objects; k is the number of clusters; u_{ic} is the membership of the ith sample object to the cth cluster; V is the set of clustering centers, $V = \{v_1,v_2,\ldots,v_k\}$. Parameter $\partial\in[1,\infty)$ represents the fuzzy parameter.

In the iterative process of the Fuzzy k-prototypes algorithm, the method of calculating membership u_{ic} is as follows:

$$\underset{\substack{1\le c\le k\\ 1\le i\le N}}{\forall} u_{ic}(x_i,v_c)=\begin{cases}1, x_i=v_c\\ 0, x_i=v_j, j\neq c\\ \dfrac{1}{\sum_{l=1}^{k}\left[\dfrac{d(x_i,v_c)}{d(x_i,v_l)}\right]^{\frac{1}{(\partial-1)}}}, otherwise\end{cases} \quad (5)$$

When iterating, the update equation for the lth($1 \le l \le s$) numerical attribute V_{cl} of the clustering center V_c is as follows:

$$v_{cl}=\frac{\sum_{i=1}^{N}(u_{ic})^{\partial}x_{il}}{\sum_{i=1}^{N}(u_{ic})^{\partial}} \quad (6)$$

For its $lth(s+1\le l\le m)$ attribute (categorical), $v_{cl}=a\in \mathrm{DOM}(A_l)$, a needs to satisfy the following conditions:

$$\sum_{i=1}^{N}((u_{ic})^{\partial}\mid x_{il}=a)\ge\sum_{i=1}^{N}((u_{ic})^{\partial}\mid x_{il}=t), a,t\in\mathrm{DOM}(A_l) \quad (7)$$

Optimized distance metric based on entropy weight method

The importance of a variable can be measured by the degree of heterogeneity of dataset relative to that variable. For a dataset, the more information a variable contains,

the more important that variable is. The concept of information entropy was introduced by Shannon (1948) for measuring uncertainty in the structure of information systems. The entropy-weighted method based on information entropy is an objective weighting method (Zhu *et al.*, 2020). According to the degree of variation of each variable, the entropy weight of each variable can be calculated by information entropy, and then the weight of each variable is corrected by entropy weight to obtain a more objective variable weight.

There are N samples and m feature variables, and the original data matrix $R = (x_{il})_{(m\times N)}$ is formed.

(1) Weight calculation of numerical variables
For the lth ($(1 \le l \le s)$ numerical variable A_l, there is information entropy:

$$e_l = -\sum_{i=1}^{N} p_{il} \ln p_{il} \tag{8}$$

where p_{il} is calculated as follows:

$$p_{il}^{num} = \frac{x_{il}}{\sum_{i=1}^{N} x_{il}} \tag{9}$$

Then the weight of the lth numerical variable is:

$$\omega_l = \frac{1-e_l}{\sum_{j=1}^{s}(1-e_l)} \tag{10}$$

(2) Weight calculation of nominal categorical variables
For nominal categorical variables, we define the range of variable A_l as $\mathrm{DOM}(x_l) = \{a_{l1}, a_{l2}, \ldots, a_{lr_l}\}(s + 1 \le l \le m)$, r_l is the number of categories of the first variable. Then the type p_{il} is calculated as follows:

$$p_{il}^{cat} = \frac{\sigma_{A_l=a_{lj}}}{N} \tag{11}$$

where $i,j \in \{1,2,\ldots,r_l\}$, and $\sigma_{A_l=a_{lj}}$ denotes the frequency of the variable value a_{lj} in the lth variable, N samples.

Equation (10) cannot be directly used to calculate the importance of categorical variables. We use the average value to adjust, that is, the importance of each variable to calculate its average entropy. For the lth ($s + 1 \le l \le m$) variable (categorical) A_l, there is information entropy:

$$e_l = -\frac{1}{r_l}\sum_{i=1}^{r_l} p_{il} \ln p_{il} \tag{12}$$

The weight of the lth variable (categorical) is calculated as follows:

$$\omega_l = \frac{1-e_l}{\sum_{j=1}^{m-s}(1-e_l)} \tag{13}$$

Based on the above weighting method and to further balance the weights, we define the distance metric of WFKP algorithm as follows:

$$d(x_i, x_j) = s\sum_{l=1}^{s} \omega_l (x_{il} - x_{jl})^2 + \lambda(m-s)\sum_{l=s+1}^{m} \omega_l \delta(x_{il}, x_{jl}) \tag{14}$$

where x_i, x_j represent the ith and jth samples in the dataset, respectively.

MDPSO-WFKP algorithm

The basic PSO algorithm is inspired by the behavior of foraging birds, and the basic idea is to randomly initialize a group of particles, ignoring their volume and mass, and consider each particle as a feasible solution to the optimization problem. The swarm of particles moves in the space of feasible solutions, and specific velocity variables determine the direction and distance of particle movement (Kennedy and Eberhart, 1995, Mellal *et al.*, 2023). The particles follow their current optimal particle locations and the population's optimal particle location for searching until the end condition is satisfied. In fact, the standard PSO algorithm is an improved PSO algorithm with inertia weights ω based on the basic PSO algorithm, which coordinates global and local search capabilities through inertia weights, and thus can ensure better convergence (Shi and Eberhart, 1999). We propose to use the standard PSO algorithm to optimize clustering centers based on mixed data as follows.

Assuming that there is a population of N particles in an M-dimensional search space, and in this paper, each particle represents the solution of a set of cluster centers, then the location of the ith particle at the tth iteration is denoted as $X_i^t = (x_{i1}^t, x_{i2}^t, \ldots, x_{im\times k}^t)^T$, and the particle velocity is $V_i^t = (v_{i1}{}^t, v_{i2}{}^t, \ldots, v_{im\times k}{}^t)^T$. The optimal location searched by the ith particle so far is called the individual extremum, which is denoted as $X_{ibest}^t = (b_{i1}^t, b_{i2}^t, \ldots, b_{im\times k}^t)^T$. The optimal location searched in the history of the population is called the global extremum and is denoted as $X_{gbest}^t = (x_{g1}^t, x_{g2}^t, \ldots, x_{gm\times k}^t)^T$. In the t + 1th generation, the velocity and location of the particle are updated according to the following equation:

$$v_{il}{}^{t+1} = \begin{cases} \omega \cdot v_{il}{}^t + c_1 r_1{}^t (b_{il}^t - x_{il}^t) + c_2 r_2{}^t (x_{gl}^t - x_{il}^t), & if\ 1 \le (l\%m) \le s \\ int\left(\omega \cdot v_{il}{}^t + c_1 r_1{}^t (b_{il}^t - x_{il}^t) + c_2 r_2{}^t (x_{gl}^t - x_{il}^t)\right), & otherwise \end{cases} \tag{15}$$

$$x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1} \quad (16)$$

where i = 1,2,...,N; j = 1,2,...,M, $M = m \times k$; ω is the inertia weight; c_1, c_2 are learning factors, also called acceleration constants; r_1^t and r_2^t are uniformly distributed random numbers in the range of [0,1], increasing the randomness of particle flight; b_{ij}^t represents the individual extreme value of the particle; x_{gj}^t is the total extreme value; and int represents rounding in the direction of 0.

For the particle X_{il} that exceeds the boundary, we define the operation:

$$X_{il} = \begin{cases} lb_l, & if\ X_{il} < lb_l \\ ub_l, & if\ X_{il} > lb_l \end{cases} \quad (17)$$

where $lb = (lb_1, lb_2, ..., lb_M)$, $ub = (ub_1, ub_2, ..., ub_M)$ represent the lower and upper boundary of particle location, respectively.

In this research, the fitness function is defined as follows:

$$f(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{a(i)-b(i)}{\max\{a(i),b(i)\}} \quad (18)$$

where $a(i)$ represents the average distance between sample i and other samples in the same cluster, and $b(i)$ represents the average distance between sample i and the closest samples in different clusters. This fitness function takes into account both compactness (CP) and separation (SP), and the smaller the obtained fitness, the better the obtained clustering results.

Consequently, the process of selecting initial clustering centers by MDPSO algorithm is shown in Table 3.

The flow chart of MDPSO-WFKP algorithm is shown in Figure 2.

MDSSA-WFKP algorithm

The sparrow search algorithm was developed based on PSO algorithm and was inspired by the predatory and anti-predatory behavior of sparrows in zoology (Xue and Shen, 2020). To optimize the selection of mixed attribute cluster centers, we propose the MDSSA algorithm based on the sparrow search algorithm. The sparrow set matrix is defined as follows:

$$X = [x_1, x_2, ..., x_n]^T,\ x_i = [x_{i1}, x_{i2}, ..., x_{i(m\times k)}] \quad (19)$$

where n is the population size of sparrows, i = 1,2,..., $m \times k$ represents the dimension of each sparrow, and the

Table 3. The process of selecting initial clustering centers by MDPSO algorithm.

Input:

data: dataset
k: number of clusters
m: number of attribute variables
s: number of numeric attributes
pop: population size
ω: inertia weight
c_1 and c_2: learning factors
t_{max}: maximum number of iterations
M: dimensionality of particle

Output: X_{best},f_{min}

1: Initialize relevant parameters
2: For *i = 1:pop*
3: Randomly take *k* samples from the dataset as the initial location of the particle, and randomly initialize the particle velocity
4: End for
5: While $t < t_{max}$
6: For *i = 1:pop*
7: Calculate the fitness value $f(X_i^t)$ according to Eq. (18)
8: If $f(X_i^t)$ is better than $f(X_{ibest}^t)$
9: Update $f(X_{ibest}^t)$ to the value of $f(X_i^t)$
10: End if
11: End For
12: Find the optimal fitness value in the population and update X_{gbest}^t;
13: For *i* = 1:*pop*
14: For *l* = 1:*M*
15: Calculate the particle velocity v_{il}^{t+1} according to Eq. (15)
16: Update the particle location x_{il}^{t+1} according to Eqs. (16) and (17)
17: ***end for***
18: ***end for***
19: $t = t + 1$
20: ***end while***
21: **return** X_{best},f_{min}

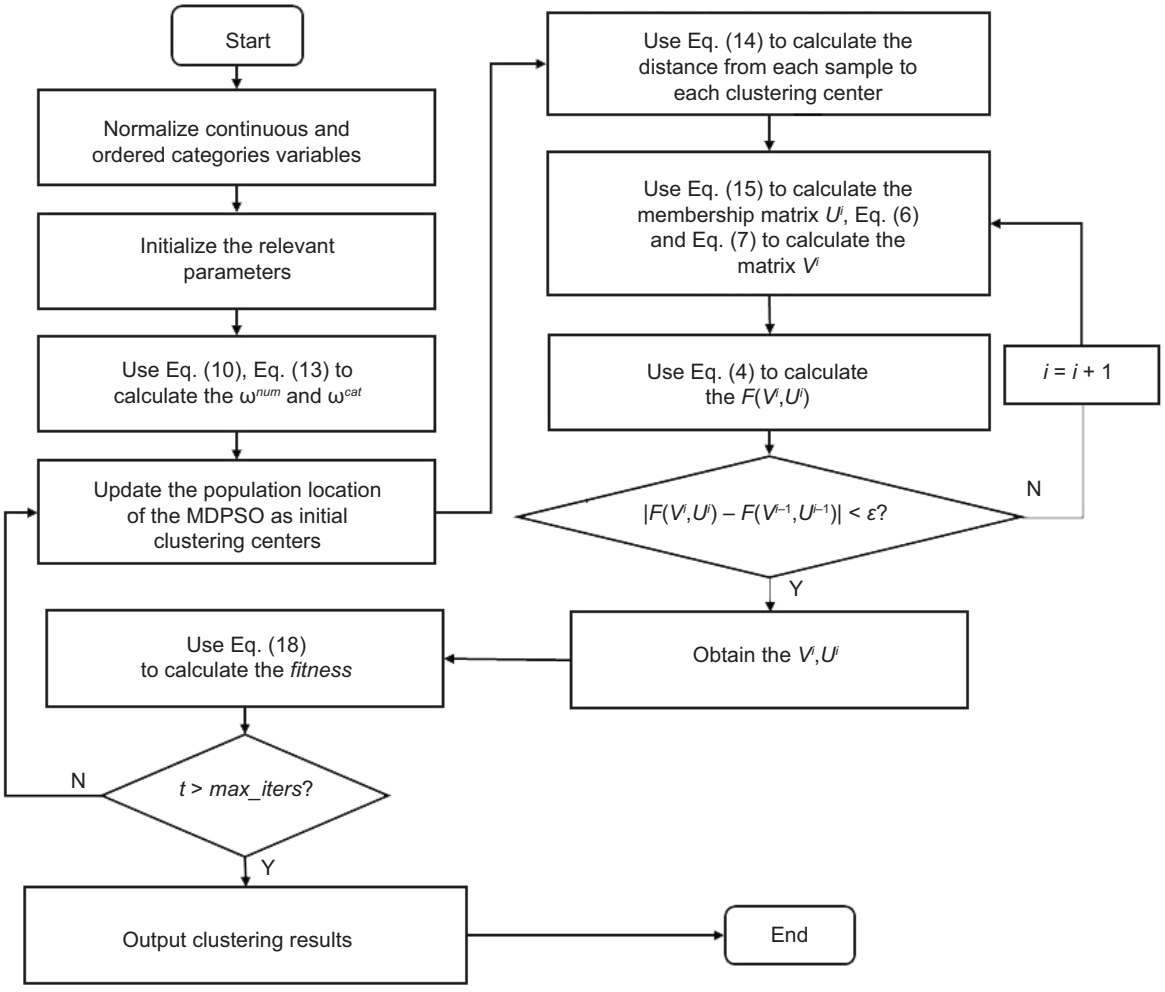

Figure 2. Flow chart of MDPSO-WFKP algorithm.

location of it in the population represents a set of clustering centers, that is, the location of sparrow is a clustering center in every m dimensions in order, totaling $m \times k$ dimensions. In the optimal selection of the initial clustering centers, the fitness matrix of sparrows is defined as follows:

$$F_x = [f(x_1), f(x_2), \ldots, f(x_N)]^T \quad (20)$$

$$f(x_i) = [f(x_{i1}), f(x_{i2}), \ldots, f(x_{im})] \quad (21)$$

where each value in F_x represents the fitness value of an individual sparrow, and the fitness function is calculated by Eq. (18). The sparrow with better fitness obtains food first and acts as a discoverer to lead the sparrow population closer to the food source. The strategy for updating the location of the discoverer can be written as Eq. (22).

$$X_{ij}^{t+1} = \begin{cases} X_{ij}^{t} \cdot exp\left(\dfrac{-i}{t_{max} \cdot \alpha}\right), r_2 < st \\ X_{ij}^{t} + Q \cdot L, r_2 \geq st \end{cases} \quad (22)$$

where X_{ij}^{t} denotes the location of the ith sparrow in the jth dimension, $j = 1,2,\ldots,m$; t refers to the current number of iterations; t_{max} denotes the total number of iterations of the algorithm; α is a random number ($\alpha \in (0,1)$); $r_2 (r_2 \in [0,1])$ represents the warning value; and st ($st \in [0.5,1]$) represents the safety value. Q is a random number obeying the normal distribution of [0,1]. L is the matrix with $1 \times m$ dimensions and all elements are 1. If $r_2 < st$, then there is no natural enemy nearby and the discoverer implements an extensive search pattern; if $r_2 \geq st$, then some sparrows in the population detect natural enemies, and the whole population is required to move to other safe areas as soon as possible.

The location of each follower is updated according to Eq. (23):

$$X_{ij}^{t+1} = \begin{cases} Q \cdot exp\left(\dfrac{X_{worst}^{t} - X_{ij}^{t}}{i^2}\right), i > \dfrac{n}{2} \\ X_{pbest}^{t+1} + \left|X_{ij}^{t} - X_{pbest}^{t+1}\right| A^{+} \cdot L, i \leq \dfrac{n}{2} \end{cases} \quad (23)$$

where X^t_{worst} represents the global worst location, and A is a $1 \times m$ matrix. Elements in this matrix are randomly assigned -1 or 1, $A^+ = A^T(AA^T)^{-1}$, if $i > \frac{N}{2}$, then the ith follower with poor fitness value does not get food, its energy value is low, and it needs to fly to other regions to supplement its own energy.

While the population is foraging, some of the sparrows in the population are responsible for vigilance and the location of each sparrow aware of the danger is updated by Eq. (24):

$$X_{ij}^{t+1} = \begin{cases} X_{gbest}^t + \beta \cdot \left| X_{ij}^t - X_{gbest}^t \right|, f_i > f_{gbest} \\ X_{ij}^t + r \cdot \dfrac{\left| X_{ij}^t - X_{worst}^t \right|}{f_i - f_{worst} + \varepsilon}, f_i = f_{gbest} \end{cases} \quad (24)$$

where X^t_{gbest} is the global optimal location of the current sparrow population search, β is the step control parameter, and $r(r \in [-1,1])$ is a uniformly distributed random number; f_i represents the fitness value of the ith sparrow of the current population, f_{gbest}, f_{worst} are the current global optimal value and worst fitness value, respectively; ε is the minimum constant to prevent the denominator from turning 0. If $f_i > f_{gbest}$, then the sparrow is at the edge of the population and is vulnerable to predator's attack; and $f_i = f_{gbest}$ means that the sparrows at the center of the population perceive the danger of predator attack and move closer to other sparrow individuals to avoid predator attack.

In the process of location update, for categorical discrete variables, we do the following processing based on Eq. (25):

$$X_{il}^{t+1} = round(X_{ij}^{t+1}) \quad (25)$$

where round(X_{ij}^{t+1}) represents rounding X_{ij}^{t+1} to the nearest integer.

For the sparrow locations exceeding upper and lower boundaries, Eq. (26) is used for processing:

$$s.t. X_{il} \left\langle lb_l \text{ or } X_{il} \right\rangle lb_l \begin{cases} X_{il} = rand\left(lb_l, ub_l\right), 1 \le \left(l\%m\right) \le s \\ X_{il} = rand\left(\mathrm{DOM}\left(x_l\right)\right), s+1 \le \left(l\%m\right) \le m \end{cases} \quad (26)$$

where rand(lb_l,ub_l) represents taking a random number between the upper and lower bounds of the lth dimension, and rand(DOM(x_l)) means taking a random value in the value domain of the lth dimension variable, that is, randomly selecting one of the categories of the variable.

The MDSSA algorithm is put forward to determine initial cluster centers, and the process is shown in Table 4. The flow chart of MDSSA-WFKP algorithm is shown in Figure 3.

Results and Discussion

Parameter setting experiment

The weight parameter λ is an important parameter in this paper, and the research (Ouyang *et al.*, 2015) indicated that the best clustering effect was achieved when λ was close to the value that classification attributes divided by the number of numerical attributes. We set the weight parameter λ as 0.36 according to this method.

Determination of fuzzy parameter ∂: Many studies set the fuzzy parameter as 2 (Ouyang *et al.*, 2015; Prasetyo, 2021), but literature (Wang and Zhu, 2005) reported that the optimal ∂ may be in (1, 1.5). Therefore, the experimental values were taken with a step of 0.05, and the values were taken from 1.05 to 2. For determining k, based on prior knowledge to determine the optimal range of k, the small- and medium-scale consumers are usually segmented into 3–10 groups with a good segmentation value (Chen *et al.*, 2022, Li *et al.*, 2021, Sun *et al.*, 2021).

To objectively determine the fuzzy parameter and the number of satisfactory clusters k in the experiment, as well as to verify the effectiveness of the proposed MDPSO-WFKP and MDSSA-WFKP algorithms, we chose the most common internal evaluation index in clustering, the silhouette coefficient (Pradana *et al.*, 2020), as the standard and experimented with the scenario of k = [3,10]. The Fuzzy k-prototypes algorithm was repeatedly run for 50 times. Eventually, the average results are shown in Figure 4.

It is observed that the value of silhouette coefficient is best when the fuzzy parameter is taken as 1.1 and k = 3. Therefore, in this research on the Chinese fresh apricot consumer segmentation dataset, we set the model parameters λ = 0.36, ∂ = 1.1, and k = 3. Furthermore, in both MDPSO-WFKP and MDSSA-WFKP algorithms, we set the population size to 50 and the maximum number of iterations to 30.

Performance evaluation of the improved clustering algorithm

Performance on University of California Irvine machine learning repository (UCI) datasets

In order to verify the effectiveness and feasibility of the algorithms proposed in this paper for clustering mixed datasets, four publicly mixed attribute datasets in the UCI Dataset (https://archive.ics.uci.edu/ml/index.php) with similar characteristics to the fresh apricot consumer segmentation dataset, namely Zoo, Heart Disease, Australian Credit Approval, and Hepatitis C Virus (HCV) for Egyptian patients, were selected. Three external evaluation

Table 4. The process of selecting initial clustering centers by MDSSA algorithm.

Input:

data: dataset
k: number of clusters
s: number of numeric attributes
lb: lower limit of sparrow location
ub: upper limit of sparrow location
pop: sparrow population size
Max_iters: maximum number of iterations
m: number of attribute variables
pd: proportion of discoverers in the population
sdNum: number of sparrows who perceive the danger
r_2: alarm value

Output: X_{best}, f_{min}

1: Initialize relevant parameters
2: ***for*** *i = 1:pop*
3: Randomly take *k* samples from data as the initial location of sparrow
4: **End for**
5: **While** *t < Max_iters*
6: Calculate the fitness values of sparrows and sort them to find the current optimal fitness value and the worst fitness value as well as corresponding locations
7: Select pd*pop sparrows with better fitness as discoverers
8: **for** *i = 1: pd*pop*
9: Update the location of discoverer according to Eq. (22)
10: The updated sparrow location is adjusted according to Eqs. (25) and (26);
11: **End for**
12: **For** *j* = (*1 – pd*) **pop:*
13: Update the location of follower according to Eq. (23)
14: Repeat operation 10 for follower's location
15: **End for**
16: Randomly select *sdNum* sparrows in the population to perceive danger;
17: **For** *I = 1:sdNum*
18: Update the vigilante's location according to Eq. (24)
19: Repeat operation 10 for vigilante's location
20: **End for**
21: Calculate current population's optimal location, according to the greedy rule; if the current population optimal location is better than the previous population optimal location, then optimal location is updated
22: *t = t + 1*
23: **End while**
24: Return

indexes, accuracy, normalized mutual information (NMI), adjusted Rand index (ARI), and two internal evaluation indexes, silhouette coefficient (SC) and Davies–Bouldin index (DBI), were used to compare and analyze the clustering results of Hard k-prototypes algorithm (Huang, 1997) as well as Fuzzy k-prototypes, MDPSO-WFKP, and MDSSA-WFKP algorithms. Among them, the larger the values of evaluation standards of the accuracy, NMI, ARI, and SC, and the smaller the value of DBI, the better the clustering effect. Table 5 shows the index test results of the proposed algorithms on public datasets, and the data are the average values of 50 repeated experiments.

The results show that both MDPSO-WFKP and MDSSA-WFKP algorithms are superior to the Hard k-prototypes and Fuzzy k-prototypes algorithms in terms of accuracy, NMI, ARI, SC, and DBI. Specifically, for the accuracy index, the MDPSO-WFKP algorithm improved by 0.1733, 0.0376, 0.0494, and 0.555, compared to the original Fuzzy k-prototypes algorithm on public datasets, namely Zoo, Heart Disease, Australian Credit Approval, and HCV, respectively, while the MDSSA-WFKP algorithm improved 0.2048, 0.0363, 0.0874, and 0.5571, respectively, showing a significant improvement. In addition, the MDSSA-WFKP algorithm had the best performance on five indexes on the Zoo dataset. On the Heart Disease dataset, it ranked second on four indexes only after the MDPSO-WFKP algorithm and exhibited the best performance in terms of SC. On the Australian Credit Approval dataset, it had the best performance in terms of four indexes. On the HCV dataset, it had the best performance in terms of all indexes. According to the above analysis, it was proved that both MDPSO-WFKP and MDSSA-WFKP algorithms, proposed in this paper, improved the limitations of original algorithm and clustering effectiveness.

Performance on the self-built dataset

All algorithms—Hard k-prototypes, Fuzzy k-prototypes, MDPSO-WFKP, and MDSSA-WFKP—were applied to segment apricot consumers. The performance results on the common three internal evaluation indexes, SC, CP, and SP, are shown in Figure 5, where the smaller CP

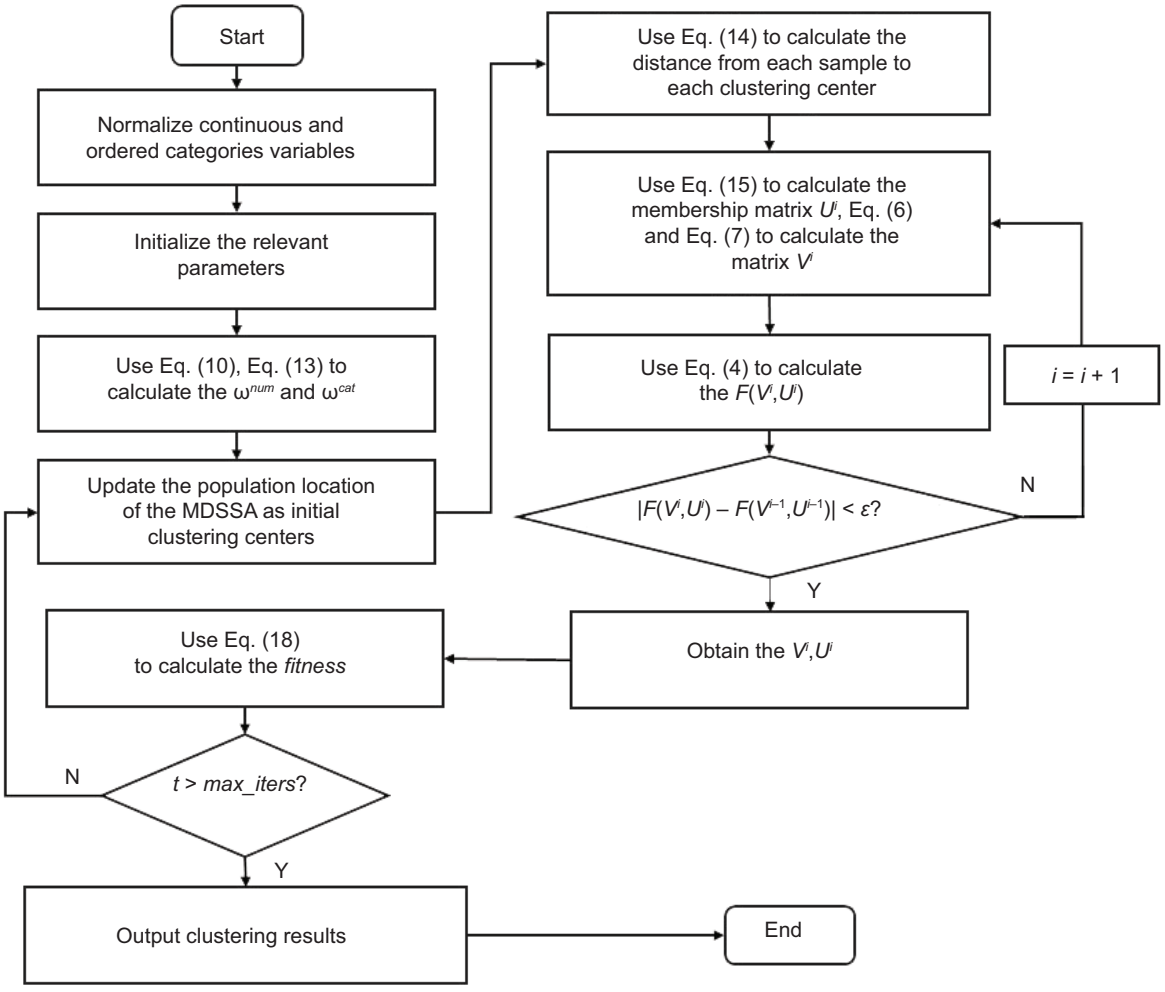

Figure 3. Flow chart of MDSSA-WFKP algorithm.

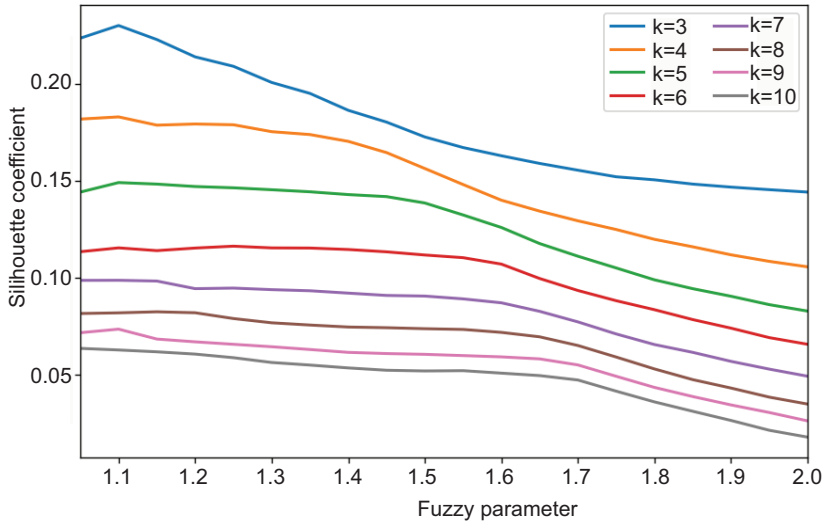

Figure 4. Reference results of parameter setting experiment.

Table 5. Clustering performance of all algorithms on public dataset.

		Hard k-prototypes	Fuzzy k-prototypes	MDPSO-WFKP	MDSSA-WFKP
Zoo	Accuracy	0.6293	0.7017	0.8750	**0.9065**
	NMI	0.7316	0.7666	0.8693	**0.8853**
	ARI	0.5516	0.6178	0.9138	**0.9301**
	SC	0.4035	0.4212	0.5580	**0.5740**
	DBI	3.3504	3.2274	2.7857	**2.5672**
Heart disease	Accuracy	0.7539	0.7718	**0.8094**	0.8081
	NMI	0.2235	0.2429	**0.2969**	0.2950
	ARI	0.2882	0.3089	**0.3836**	0.3776
	SC	0.2632	0.2653	0.2960	**0.2975**
	DBI	4.1188	4.1150	**3.9615**	4.0196
Australian credit approval	Accuracy	0.7479	0.7662	0.8156	**0.8536**
	NMI	0.2178	0.2476	0.3230	**0.4228**
	ARI	0.3161	0.3436	0.4008	**0.4995**
	SC	0.2218	0.2257	0.2526	**0.2612**
	DBI	5.1149	5.2839	**5.0822**	5.2424
HCV	Accuracy	0.3733	0.3657	0.9207	**0.9228**
	NMI	0.1513	0.1554	0.4275	**0.4287**
	ARI	0.0530	0.0543	0.6279	**0.6289**
	SC	0.2976	0.2999	0.6175	**0.6253**
	DBI	1.4746	1.4673	1.3806	**1.3508**

Values in bold are optimal values. HCV: Hepatitis C Virus (HCV) for Egyptian patients; NMI: normalized mutual information; ARI: adjusted Rand index; DBI: Davies–Bouldin index; SC: silhouette coefficient.

indicates the better clustering result, but SC and SP were opposite to it. The results are the average values taken from 50 repeated experiments.

The results show that both MDPSO-WFKP and MDSSA-WFKP algorithms proposed in this paper are superior to the original Fuzzy k-prototypes algorithm and Hard k-prototypes algorithm in terms of SC, CP, and SP on the segmentation dataset. Furthermore, MDSSA-WFKP is optimal on all three internal evaluation indexes. Based on the evaluation index results of algorithms on four public datasets and the segmentation dataset, we chose MDSSA-WFKP algorithm with the best overall performance to be applied to further segmentation research.

Analysis and discussion of consumer segmentation results

Based on the multiple-perspective consumer segmentation indicator system constructed in the previous text, the MDSSA-WFKP algorithm was used to achieve clustering segmentation of Chinese fresh apricot consumers. The results showed that the consumers could be divided into three groups, and the number of samples in each group accounted for 76.90%, 10.39%, and 12.71% of the total sample size.

Differences in the characteristics of three sub-categorized consumers are shown in Figures 6–10. Among them, Figure 6 shows the segmentation results of some categorical variables from the perspective of demographic characteristics and behavioral characteristics. Figures 7 and 8 show the segmentation results of the variables of age and number of family members, respectively, in the form of violin plots. Figure 9 shows the regional distribution of segmented groups; and Figures 10 and 11 show the clustering results from the perspective of psychological characteristics and cognition degree of product knowledge, respectively. Specific analysis is to be carried out later.

Description and analysis of consumer groups after segmentation

Segmented groups are often named after consumer preferences for attributes, behavioral differences, and other characteristics, resulting in clearer target markets and more targeted decisions. Through cluster analysis of

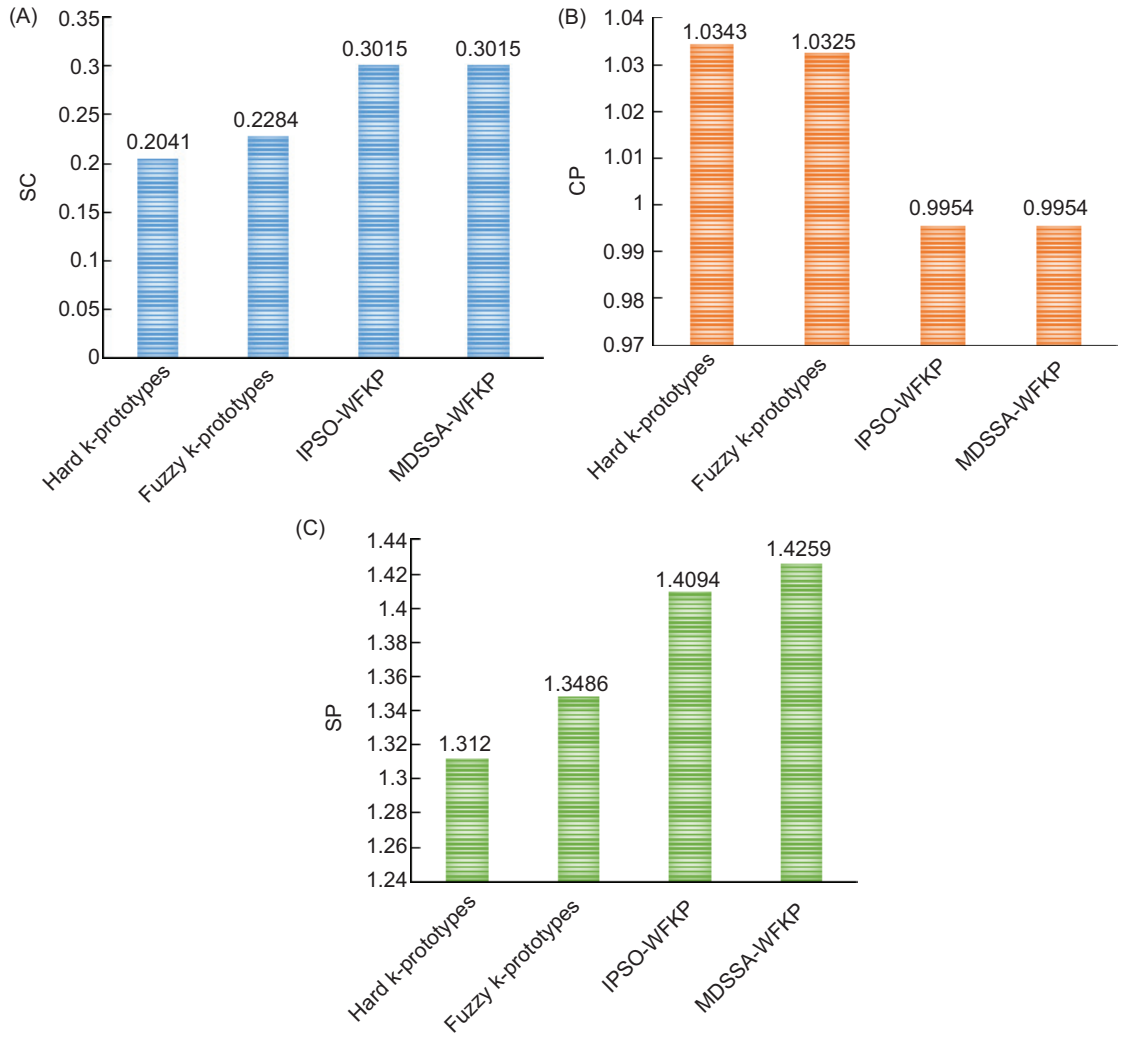

Figure 5. Performance of algorithms on the segmentation dataset: (A) SC; (B) CP; and (C) SP.

consumers, Chen *et al.* (2021) obtained the following four consumer groups: the pursuit-of-quality group, the cost-effective group, the service-preference group, and the appearance-and-logistics group. Based on the characteristics of purchase motivation, Bejaei *et al.* (2020) segmented out the following five target markets: better-eating quality seekers, better-eating quality seekers of familiar or good-looking apples, taste lover buyers, perfect product seekers, and cultivar-loyal buyers. The names of customer groups were developed by considering the frequencies of selection of the top four purchase reasons (i.e., visual appearance, texture, taste/aroma, and previous experience). The difference between the two representative researches above is: the first is based on the clustering of internal and external characteristics of products from the following five aspects: product quality, value and price, reputation and service, packaging and logistics, and preference recognition, taking into account a richer range of influencing factors; the second explored four main factors that affect consumers' purchase of apple varieties through a survey questionnaire, and then subdivided and named groups based on them, with different research processes.

In this paper, the segmented consumer characteristics are described and analyzed from the perspectives of demographic characteristics, geographic characteristics, behavioral characteristics, psychological characteristics, and cognition degree of product knowledge as shown in Table 6. Meanwhile, according to the previous relevant research and the most prominent characteristics of each group, the first cluster is named as 'Buddhist-like youths', the second cluster as 'Upscale attribute enthusiasts', and the third cluster as 'Quality-oriented consumers'.

Marketing suggestions

Based on the differentiated characteristics of segmented consumer groups, the marketing suggestions for different consumer groups are concluded.

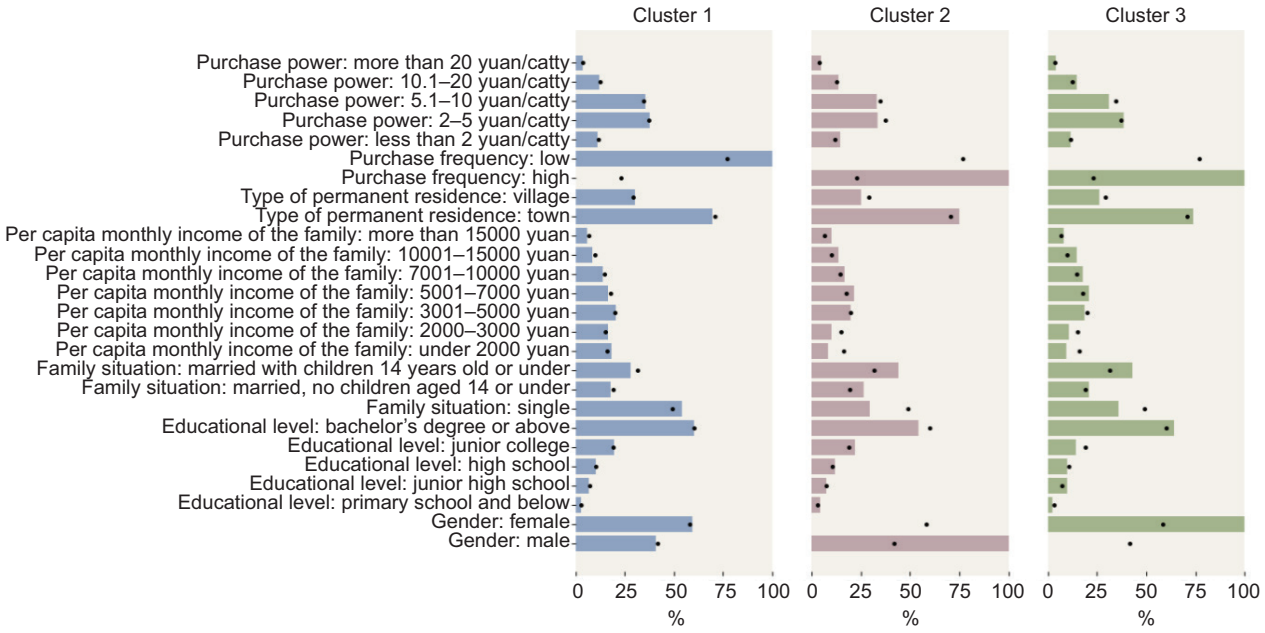

Figure 6. Segmentation results of categorical variables. Dots in the figure represent the distribution of each category in the overall sample.

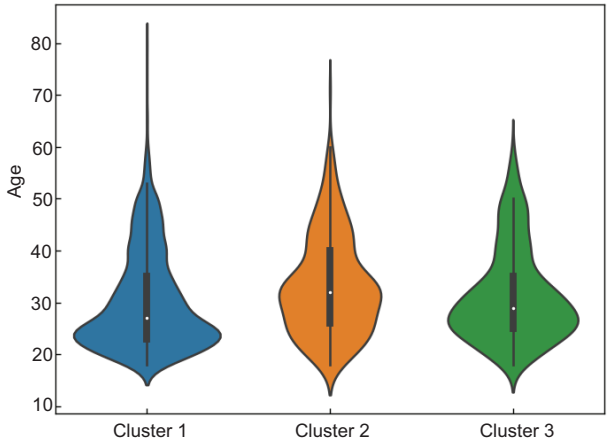

Figure 7. Age distribution of consumers in clusters.

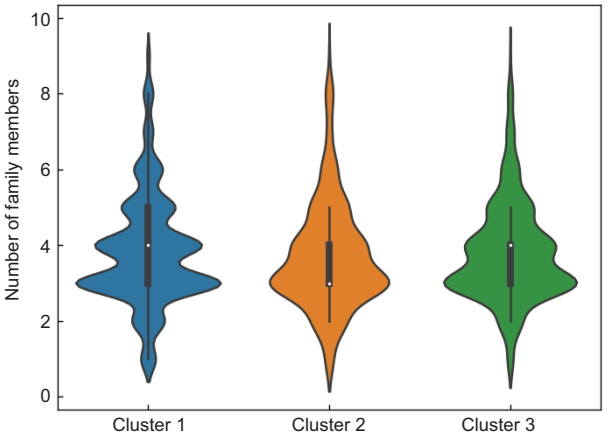

Figure 8. Number of family members in clusters.

For Group 1, 'Buddhist-like youths', considering their characteristics of lower income, purchasing power, and purchase frequency, but higher price sensitivity, the first strategy is to recommend low- and medium-price fresh apricot under the concept of small profits and quick returns. Second, based on the fact that the cognition degree of product knowledge of this group is not high, it is suggested to adopt precise information push strategy, enrich the forms of publicity, and carry out scientific popularization of apricot nutrition value to the audience in a targeted manner to prevent consumers from being affected by false folk rumors and improve their cognition toward apricots.

For the 'Upscale attribute enthusiasts', the first step is to integrate a marketing communication strategy to improve the targeting of communication and consolidate the brand loyalty of this group of consumers. Second, most consumers in this cluster have high purchase frequency and can accept medium and high prices, so the high-quality development and price differentiation strategy are recommended, specifically decision-makers can determine prices according to the quality of the fruit and implement the business strategy of high quality, high price. Third, producers and enterprises should increase brand publicity, appropriately focusing on the elegance of packaging and the specificity

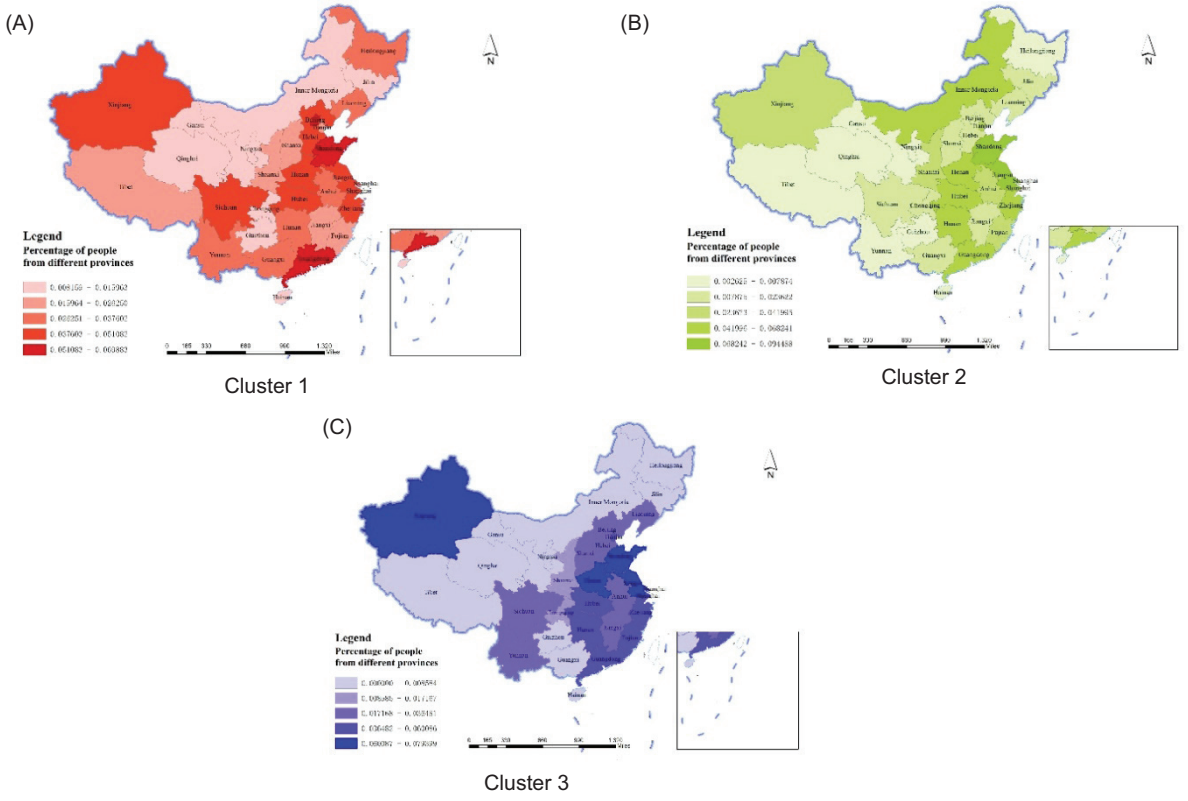

Figure 9. Regional distribution of clusters.

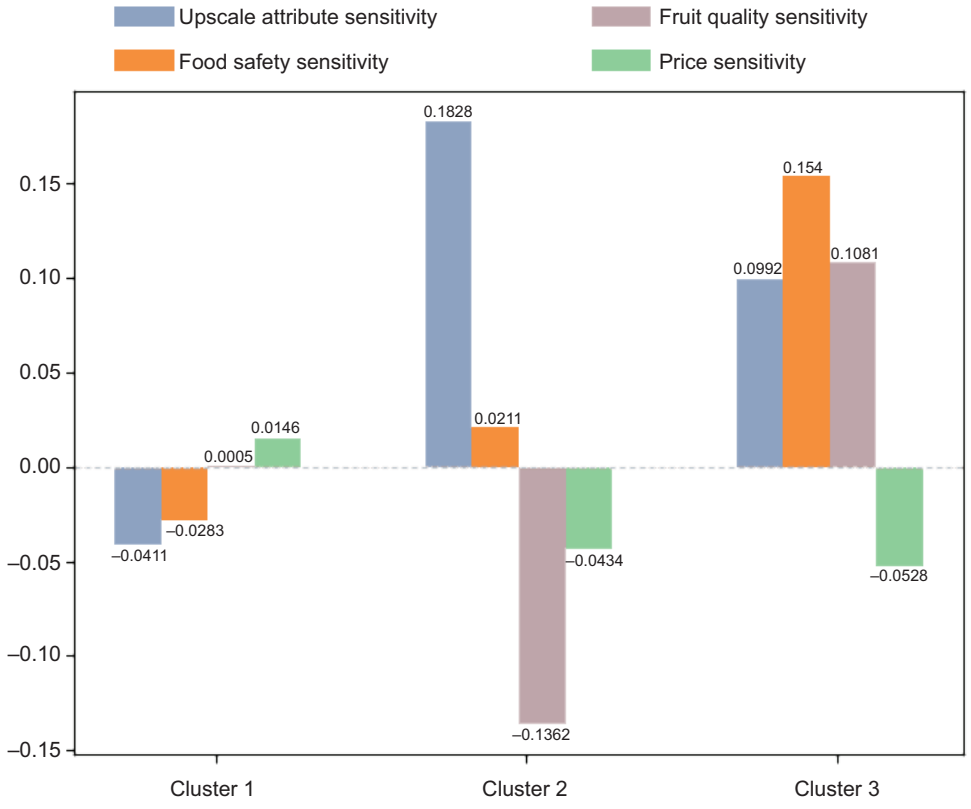

Figure 10. Indicator results from the perspective of psychological characteristics.

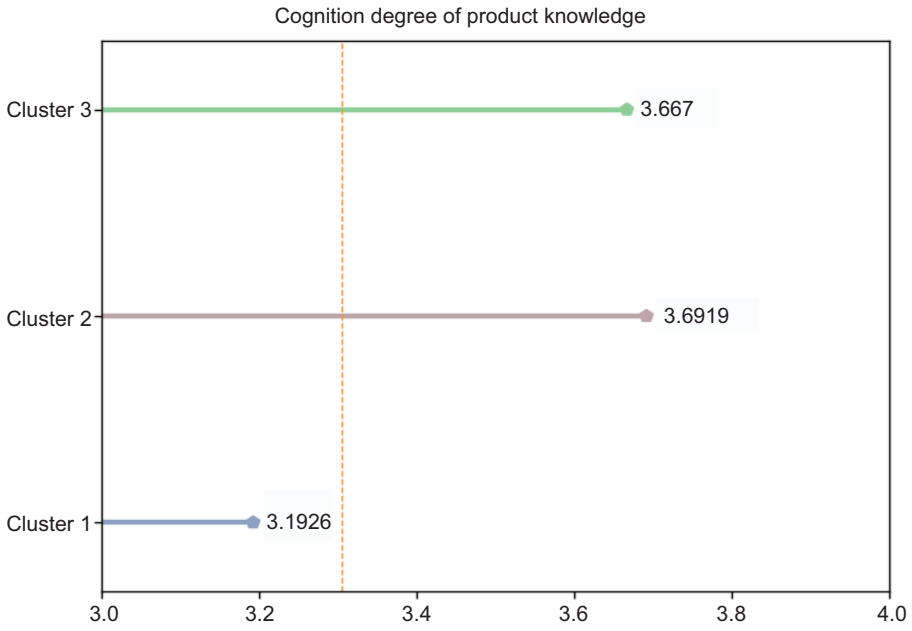

Figure 11. Cognition degree of product knowledge. Note: Dotted line represents the overall average.

Table 6. Description of segmentation group characteristics in multiple perspectives.

	Buddhist-like youths	Upscale attribute enthusiasts	Quality-oriented consumers
Age (in general [years])	<30	>30	20–30
Number of family members (average)	4	3	4
Gender	Male & female	Male	Female
Education level	Middle to higher	Middle to higher	High
Per capita monthly income of the family	Low	Middle to higher	Middle to higher
Family situation (in general)	Single	Married, with children aged <14 years	Married, with children aged <14 years
Type of permanent residence (in general)	Urban	Urban	Urban
Area of permanent residence	Beijing, Guangdong Province, etc.	Jiangsu, Shandong, Guangdong, Hunan, etc.	Shandong Province, Henan Province, Jiangsu Province, Xinjiang Uygur Autonomous Region, etc.
Purchase frequency	Low	High	High
Purchasing power	Low	High	High
Upscale attribute sensitivity	Low	High	Middle
Food safety sensitivity	Low	Middle	High
Fruit quality sensitivity	Middle	Low	High
Price sensitivity	High (relatively)	Low	Low
Cognition degree of product knowledge	Low	High	High

of origin, as a powerful promotional point, and take the upscale sales route.

For the 'Quality-oriented consumers', one piece of advice is to implement differentiated pricing strategies and recommend medium- and low-priced apricot to this group, for they have high purchase frequency but prefer lower prices. As consumers in this group are less sensitive to high-grade attributes of fruits, we can simplify fruit packaging and reduce cost of sales. The cost of sales is based on the characteristics of this group, which attaches more importance to fruit quality and food safety. Marketers should expand the external extension of apricots, pay attention to the preservation of this fruit, improve the

information identification of apricot sales, and accurately convey the advantages of sold apricots, such as freshness, beauty, and greenness, in appropriate forms.

Conclusion and Further Research

The research constructed a novel multiple-perspective segmentation indicator system for fresh apricot consumers based on five aspects: demographic characteristics, geographical characteristics, behavioral characteristics, psychological characteristics, and cognition degree of product knowledge in consumer segmentation. This clustering indicator system provides a characteristic framework for customer segmentation research of related fresh agricultural products.

Furthermore, to address the problems of unequal importance of attributes and sensitive clustering centers of the standard Fuzzy k-prototypes algorithm, the consumer segmentation methods were optimized from the perspective of cluster analysis for mixed dataset. Based on this concept, two algorithms, MDPSO-WFKP and MDSSA-WFKP, were presented. The proposed methods had two innovations: First, the information entropy weighted method was introduced into the Fuzzy k-prototypes algorithm, and the practical problem of unequal importance of variables was solved by improving distance measurement. Second, both MDPSO and MDSSA algorithms were proposed as applicable to the optimization of the initial clustering center selection of WFKP algorithm.

In this paper, both MDPSO-WFKP and MDSSA-WFKP algorithms were tested on four UCI public datasets and compared with both Hard k-prototypes algorithm and Fuzzy k-prototypes algorithm. The experimental results showed that the clustering algorithms proposed in this paper had significant improvement in three external evaluation indexes and two internal evaluation indexes. Among 20 indexes in four public datasets, MDSSA-WFKP performed best on 15 indexes and performed second only to the proposed MDPSO-WFKP algorithm on five indexes. On the segmentation dataset, MDPSO-WFKP improved 0.0731 and 0.0608 in the SC and SP, respectively, and reduced CP by 0.0371, compared to the original Fuzzy k-prototypes algorithm. Similarly, increased values of SC and SP and decreased value of CP for MDSSA-WFKP algorithm were 0.0731, 0.0773, and 0.0371, respectively. This verified the effectiveness of all algorithms proposed in this paper. Finally, MDSSA-WFKP with the best comprehensive effect was chosen to segment consumers, and the group characteristics in various forms were further visualized and analyzed. The three consumer groups were named 'Buddhist-like youths', 'Upscale attribute enthusiasts', and 'Quality-oriented consumers'. Moreover, the corresponding marketing suggestions in terms of different characteristics of segmented groups were put forward.

This paper synthesized five perspectives for cluster analysis of fresh apricot consumers, and the segmented groups showed significant differences. However, the study lacks consideration from the perspective of consumers' preference differences of fresh apricots in sensory attributes. Therefore, future research must further supplement to investigate the heterogeneity in this aspect.

Acknowledgments

This paper is our original work and has not been published or submitted elsewhere. It was supported by the earmarked fund of Xinjiang Apricot Industrial Technology System (XJCYTX-03). In addition, all authors have consented to the submission of this paper and declared no conflict of interest.

Author Contributions

Yan Shi and Jianying Feng prepared methodology and conceptualization, conducted the experimental work, and wrote original draft. Siyuan Zhang and Siwen Wang validated data. Hui Xie and Weisong Mu reviewed and edited original article.

References

Abbasimehr H. and Bahrini A. 2022. An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. Expert Syst Appl. 192: 116373. https://doi.org/10.1016/j.eswa.2021.116373

Bannor R.K., Abele S., Kuwornu J.K., Oppong-Kyeremeh H. and Yeboah E.D. 2022. Consumer segmentation and preference for indigenous chicken products. J Agribus Dev Emerg Econ. 12(1): 75–93. https://doi.org//10.1108/JADEE-08-2020-0162

Bejaei M., Cliff M.A. and Singh A. 2020. Multiple correspondence and hierarchical cluster analyses for the profiling of fresh apple customers using data from two marketplaces. Foods. 9(7): 873. https://doi.org/10.3390/foods9070873

Bhattacharjee P. and Mitra P. 2020. A survey of density based clustering algorithms. Front Comput Sci. 15(1): 151308. https://doi.org/10.1007/s11704-019-9059-3

Chen N., Chen A. and Zhou L. 2001. Fuzzy k-prototypes algorithm for clustering mixed numeric and categorical valued data. J Softw. 12(8): 1107–1119.

Chen T.-C., Ibrahim Alazzawi F.J., Mavaluru D., Mahmudiono T., Enina Y., Chupradit S., et al. 2022. Application of data mining methods in grouping agricultural product customers. Math Probl Eng. 2022: 3942374. https://doi.org/10.1155/2022/3942374

Chen H., Li S., Wang C. and Xu S. 2021. Influencing factors of consumers' purchase decision in wine online shopping and customer segmentation based on online review data. Liquor Making Sci Technol. (11): 127–132. https://doi.org/10.13746/j.njkj.2021070 (in Chinese)

Du M. and Wu F. 2022. Grid-based clustering using boundary detection. Entropy. 24(11): 1606. https://doi.org/10.3390/e24111606

Ezugwu A.E., Ikotun A.M., Oyelade O.O., Abualigah L., Agushaka J.O., Eke C.I. and Akinyelu A.A. 2022. A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng. Appl. Artif. Intell. 110: 104743. https://doi.org/10.1016/j.engappai.2022.104743

Ghadiri S.M.E. and Mazlumi K. 2020. Adaptive protection scheme for microgrids based on SOM clustering technique. Appl Soft Comput. 88: 1060. https://doi.org/10.1016/j.asoc.2020.106062

Guo Y., Liu R. and Jiang X. 2019. Fresh corn consumption by Beijing urban residents: market segmentation. Chin Agric Sci Bull. 35(32): 153–157. (in Chinese)

Hasan B.M.S. and Abdulazeez A.M. 2021. A review of principal component analysis algorithm for dimensionality reduction. J Soft Comput Data Mining. 2(1): 20–30. https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032

Hegazi A., Taha A. and Selim M.M. 2021. An improved copy-move forgery detection based on density-based clustering and guaranteed outlier removal. J King Saud University Comput Inform Sci. 33 (9): 1055-1063. https://doi.org/10.1016/j.jksuci.2019.07.007

Huang Z. 1997. Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining (PAKDD). CiteSeer, pp. 21–34.

Kazbare L., van Trijp H.C. and Eskildsen J.K. 2010. A priori and post hoc segmentation in the design of healthy eating campaigns. J Market Commun. 16(1–2): 21–45. https://doi.org/10.1080/13527260903342712

Kennedy J. and Eberhart R. 1995. Particle swarm optimization. Proceedings of ICNN'95 International Conference on Neural Networks. IEEE. 4: 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Kiran A. and Vasumathi D. 2020. Data mining: min–max normalization based data perturbation technique for privacy preservation. In: Raju, K., Govardhan, A., Rani, B., Sridevi, R., Murty, M. (eds.) Proceedings of the third international conference on computational intelligence and informatics. Adv Intell Syst Comput. 1090: 723–734. https://doi.org/10.1007/978-981-15-1480-7_66

Kuesten C., Dang J., Nakagawa M., Bi J. and Meiselman H.L. 2022. Japanese consumer segmentation based on general self-efficacy psychographics data collected in a phytonutrient supplement study: influence on health behaviors, well-being, product involvement and liking. Food Quality Pref. 99: 104545. https://doi.org/https://doi.org/10.1016/j.foodqual.2022.104545

Lee Y., Song S., Cho S. and Choi J. 2019. Document representation based on probabilistic word clustering in customer-voice classification. Pattern Anal Appl. 22: 221–232. https://doi.org/10.1007/s10044-018-00772-1

Li Y., Chu X., Tian D., Feng J. and Mu W. 2021. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. Appl Soft Comput. 113: 107924. https://doi.org/10.1016/j.asoc.2021.107924

Mellal M.A., Tamazirt I., Tiar M. and Williams E.J. 2023. Optimal conventional and nonconventional machining processes via particle swarm optimization and flower pollination algorithm. Soft Comput. 28: 3847–3858 https://doi.org/10.1007/s00500-023-09320-4

Mollaei S., Minaker L.M., Robinson D.T., Lynes J.K. and Dias G.M. 2023. Including sustainability factors in the derivation of eater profiles of young adults in Canada. Br Food J. 125(5): 1874–1894. https://doi.org/10.1108/BFJ-06-2022-0476

Ouyang H., Wang Z., Dai X. and Liu Z. 2015. A fuzzy K-prototypes clustering algorithm based on information gain. Comput Eng Sci (CES). 37(5): 1009–1014. (in Chinese)

Park H.-J., Ko J.-M., Lim J. and Hong J.-H. 2020. American consumers' perception and acceptance of an ethnic food with strong flavor: a case study of Kimchi with varying levels of red pepper and fish sauce. J Sci Food Agric. 100(6): 2348–2357. https://doi.org/10.1002/jsfa.10106

Pradana C., Kusumawardani S. and Permanasari A. 2020. Comparison clustering performance based on moodle log mining. In: 3rd International Conference on Engineering Technology for Sustainable Development (ICET4SD) 23–24 October 2019, Yogyakarta, Indonesia. IOP Conference Series: Materials Science and Engineering, vol. 722. IOP Publishing, Philadelphia, PA; p. 722: 012012. https://doi.org/10.1088/1757-899X/722/1/012012

Prasetyo H. 2021. Pengelompokan wilayah menurut potensi fasilitas kesehatan dan kejadian COVID-19 menggunakan algoritma fuzzy k-prototypes. Technol J Ilmiah. 12(4): 223–227. http://doi.org/10.31602/tji.v12i4.5631

Shannon C.E. 1948. A mathematical theory of communication. Bell Syst Tech J. 27(3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shi Y. and Eberhart R.C. 1999. Empirical study of particle swarm optimization. In: Proceedings of the 1999 Congress on Evolutionary Computation – CEC99 (Cat. No. 99TH8406). IEEE. 3: 1945–1950. https://doi.org/10.1109/CEC.1999.785511

Shi Y., He M., Xie H., Tian D. and Feng J. 2022. Chinese consumers' behavior and preference to fresh apricot. Chin Fruits. 2022(7): 84–90. https://doi.org/10.16626/j.cnki.issn1000-8047.2022.07.019

Singh D. and Singh B. 2020. Investigating the impact of data normalization on classification performance. Appl Soft Comput. 97: 105524. https://doi.org/10.1016/j.asoc.2019.105524

Sun Z., Zuo T., Liang D., Ming X., Chen Z. and Qiu S. 2021. GPHC: a heuristic clustering method to customer segmentation. Appl Soft Comput. 111: 107677. https://doi.org/10.1016/j.asoc.2021.107677

Taherdoost H., Sahibuddin S. and Jalaliyoon N. 2022. Exploratory factor analysis; concepts and theory. Adv Appl Pure Math. 27: 375–382.

Tohidi A., Mousavi S., Dourandish A. and Alizadeh P. 2023. Organic food market segmentation based on the neobehavioristic theory

of consumer behavior. Br Food J. 125(3): 810–831. https://doi.org/10.1108/BFJ-12-2021-1269

Wang O. and Scrimgeour F. 2023. Consumer segmentation and motives for choice of cultured meat in two Chinese cities: Shanghai and Chengdu. Br Food J. 125(2): 396–414. https://doi.org/10.1108/BFJ-09-2021-0987

Wang J. and Zhu Y. 2005. Research on the weighting exponent in fuzzy K-prototypes algorithm. Comput Appl. 25(2): 348–351. (in Chinese)

Weber C.M., Ray D., Valverde A.A., Clark J.A., and Sharma K.S. 2022. Gaussian mixture model clustering algorithms for the analysis of high-precision mass measurements. Nucl Instrum Methods Phys Res Sect A. 1027: 166299. https://doi.org/10.1016/j.nima.2021.166299

Xue J. and Shen B. 2020. A novel swarm intelligence optimization approach: sparrow search algorithm. Syst Sci Control Eng. 8(1): 22–34. https://doi.org/10.1080/21642583.2019.1708830

Ye Q. and Liang G. 2010. Fuzzy K-prototypes clustering based on quantum genetic algorithm. Comput Eng Appl. 46(1): 112–115. https://doi.org/10.3778/j.issn.1002-8331.2010.01.035. (in Chinese)

Zhu Y., Tian D. and Yan F. 2020. Effectiveness of entropy weight method in decision-making. Math Probl Eng. 2020: 3564835. https://doi.org/10.1155/2020/3564835