Application of Improved K-means Algorithm in E-commerce Data Processing

Wenwei Chen*, Qindi Wang
School of Business Management, Hangzhou Polytechnic, Hangzhou 311402, China
E-mail: chenwenwei1979@163.com
*Corresponding author

Keywords: k-means; e-commerce goods; genetic algorithm; recommender system; singular value decomposition

Received: April 8, 2024

Accurate recommendation processing for a large number of e-commerce products can play a role in increasing e-commerce sales and improving the user's consumption experience. This study uses genetic algorithm, coefficient of variation method to design an improved k-means algorithm, and design an improved singular value decomposition ++ algorithm, so as to construct an e-commerce product data recommendation model. The model uses the improved singular value decomposition ++ algorithm to extract the hidden features of the data, and the improved k-means algorithm to realize the recommendation of the products. The performance test results revealed that when the number of recommendations was 15, the area under the recommended precision, recall, and receiver operating characteristic curves of the designed recommendation model were 85%, 87%, and 0.83, respectively, which were higher than all the ablation experiment comparison models and advanced recommendation models. The average computational time consumption of ISVD++_I_k-means, RLRA, TRRA, and CF models were 54.2s, 73.8s, 83.3s, and 58.7s, respectively. Among them, ISVD++_I_k-means consumed less time, but the computational memory consumption of the designed model was in the worse level among all the comparison models. The test results demonstrate that, although there are certain drawbacks in terms of resource consumption, the recommendation model developed in this study can successfully increase the efficiency and quality of recommendations. The research results are beneficial to provide reference for e-commerce platforms to design more efficient recommendation models.

Povzetek: Študija uporablja izboljšan algoritem razvrščanja z voditelji (k-means) za obdelavo podatkov v e-trgovini z uporabo genetskega algoritma in metode koeficienta variacije, kar izboljšuje točnost in učinkovitost priporočil za izdelke.

1 Introduction

In the current era, people can buy most of the items needed for life in e-commerce platform (ECP) [1]. However, too much product information can also make people uncertain about their purchasing choices, thus wasting users' shopping time and degrading the shopping experience [2]. Massive volumes of data about online shopping are produced as the e-commerce sector grows and matures [3]. Effective utilization of online user behavior data in e-commerce industry is the key to enhance user's consumption experience, so recommender system (RS) has gradually become the focus of academic research [4]. RS is an information filtering approach that aims to help users find more suitable content in a huge ocean of information [5-6]. The fields of content-based recommendation, collaborative filtering, and hybrid recommendation have all evolved since the 1990s, when RS was first proposed. During the course of their gradual evolution, the corresponding recommendation algorithm models have also grown more intelligent and effective, however issues like data sparsity and cold start persist [7]. Since RS can find the objects that users may be interested in from the huge amount of product data by means of precise data analysis, it is beneficial to improve the user's stickiness and shopping experience, and thus RS is a key element in the development of ECP. However, considering the shortcomings of the traditional RS model mentioned above, it is necessary to improve it.

K-means algorithm (KMA) is a division based clustering algorithm that finds several cluster centers (C-C) in an iterative manner and divides the points to the nearest C-Cs to cluster the data [8]. This algorithm has demonstrated excellent performance in areas such as market segmentation, social networking, image processing, etc. Since the k samples closest to a particular sample in the feature space can be obtained in KMA, it naturally has the potential to be applied to recommendation work, but KMA also still has some problems in clustering quality and stability [9]. Specifically, the traditional KMA obtains the initial cluster centers (ICC) by random extraction or weighted random extraction, which brings large uncertainty to the operation results of the algorithm and weakens the stability of the algorithm. If the ICCs are in some special

positions may cause the algorithm to converge very slowly or even fail to converge [10]. In addition, traditional KMA often uses equal weights or simple weighting when determining the nearest neighbor (N-N) list, which cannot reflect the difference in the importance of different features for different items, thus leading to lower quality of recommendation results. Many improvement strategies, based on sample weights according to the distribution density of the samples in the feature space, have been proposed in academia and industry to address the aforementioned shortcomings. These strategies help to mitigate the negative effects of the random initialization of the C-Cs, but they are not able to address the core issues with KMA. To create a recommendation model (RM) for the e-commerce goods (ECG) data recommendation task, the goal of this research is to enhance KMA. This should enable ECP users to receive recommendations for goods of a higher caliber.

The following sections comprise the primary material of this research. The first part provides background information on the evolution of the e-commerce business and the resulting rise in demand for product recommendations, which is what made this research necessary and intended. The second part designs the ECG RM based on improved singular value decomposition++ (SVD++), genetic algorithm (GA) to improve KMA. The third part of the study focuses on designing two experiments for testing the recommendation accuracy, recommendation efficiency, resource consumption and other metrics. The content of the last part is to summarize the research content and findings of the whole study, and to elaborate and analyze the limitations of this research and the future research directions.

2 Related works

Many computer experts and academics have studied RS because it has great application value in retail business scenarios with a lot of information. Gwadabe and Liu's research team found that if there is no user archived information, e-commerce websites can improve recommendation results by using interaction transformations across conversations, a method known as session-based recommendation. However, this advice was limited because of the scant data and the erratic nature of user activity. Consequently, the study created a RM using a recurrent neural network that was based on session-based recommendation. Then the study tested the model for recommendations and got the following outcomes. The model outperformed common RMs on Yoochoose and Diginetica datasets [11]. Roozbahani et al. proposed an integrated RM based on a multilayer network, which also introduces a semi-supervised module that allows it to achieve better recommendation results even with insufficient training data. Test results revealed that the designed model significantly outperformed the pre-improvement model in the social network data recommendation task [12]. Choudhary et al. designed a deep neural network-based RM using an integrated approach which is capable of analyzing both ratings and reviews data and sentiment analysis of reviews for subsequent processing data. Moreover, the model had a hidden layer structure, which could improve the overall learning ability. Test outcomes revealed that the Top-5 recommendation accuracy of the RM designed by the authors was 76.3% higher than the pre-improvement model and the collaborative filtering RM [13]. Rabiu et al. research team found that the historical rating data used for recommendation in collaborative filtering RS is generally sparse or unbalanced, and the combination of user comments and ratings can better capture user sentiment and thus help users make high-quality recommendations. In light of the aforementioned considerations, the team responsible for developing this RM opted to construct it on the basis of long- and short-term memory neural networks, with the objective of capturing the emotional variations between user ratings and ratings. The outcomes of the tests based on Amazon ECP revealed that the RM designed by the authors is able to output high-quality recommendations for cell phones, babies, and fine foods for users, which has certain application potential [14].

KMA is simple in principle, fast convergence, only one parameter k needs to be changed when tuning the parameters, good interpretability, this is one of the most important reasons why it is used most in the industry. For the liver and liver tumor segmentation challenge, V. N. Pattwakkar et al. suggested a segmentation model based on SegNet deep neural network and KMA. The experimental results showed that the Dice coefficient of $96.46 \pm 0.48\%$ and the Jaccard index of $93.16 \pm 0.89\%$ were superior to other models [15]. For the short-term traffic flow (TF) prediction problem, Sun et al. suggested a prediction model based on KMA and gated recursive unit prediction. The historical TF data were clustered using KMA by the model, and the N-N classification approach was utilized to identify the historical TF pattern that most closely resembled the TF trend on the forecast date. The model enhanced prediction accuracy and took into account the diversity of TF patterns, according to experimental data [16]. In reaction to the grey wolf optimization (GWO) method's propensity to enter a local optimum, Mohammed et al. suggested changing the algorithm by employing KMA. By segmenting the population into distinct groups, the K-means clustering algorithm (KMCA) in the modified GWO algorithm will improve the performance of the original GWO. According to experimental data, KM-GWO outperforms the other algorithms in terms of significant values. Furthermore, pressure vessel design issues were successfully resolved using KM-GWO [17]. For feature selection in supervised situations, Ziabari et al. presented an effective infinite feature selection technique based on K-means (KM). The process began with

clustering the feature space into a predefined number of subspaces. Next, the features in each subspace were ranked using the Inf-FS approach. Lastly, the resulting subclassifications were combined using a measure of information theory and cluster size. The accuracy, runtime, and memory consumption of this method are better than those of infinite feature selection, according to experimental data [18]. Natarajan and Rebekka proposed an optimization strategy for SC switching algorithm based on KMA and dynamic loading for the problem of optimizing the energy efficiency of topological network systems with small base stations and auxiliary macro base stations [19]. The specific research results of various scholars are shown in Table 1.

Table 1: Specific research achievements of scholars

Lead author	Method	Result	Disadvantage
Gwadabe and Liu [11]	Improving graph neural networks based on session recommendation systems through non sequential interaction	More than 10% higher performance than other state-of-the-art models on the Yoochoose and Diginetica datasets	Relatively dependent on session-based recommendation models
Roozbahani et al. [12]	Integrated model based on multi-layer networks	This model can prevent information loss in the network	Social network data is dynamically changing, and there may be some errors in the results
Choudhary et al. [13]	Recommendation model based on integrated deep learning methods	The model performs well in user preference recognition	The two outputs in the integrated network must be within the same range
Rabiu et al. [14]	recommendation system based on adaptive long short term memory network	This model is superior to existing static and dynamic models	There may be some errors in simulating user and project characteristics
Pattwakka et al. [15]	A liver tumor segmentation model based on K-means clustering and segNet deep neural network	The model performs well in liver tumor segmentation	False positives can affect the accuracy of liver tumor segmentation
Sun et al. [16]	Short term traffic flow prediction model combining K-means clustering and doorstep recursive units	This model considers the diversity of traffic flow patterns and improves prediction accuracy	This model can only predict short-term traffic flow and cannot achieve coverage of the entire road network
Mohammed et al. [17]	Engineering problem solving method combining grey wolf optimization algorithm and K-means clustering	This method effectively solves the problem of pressure vessel design and has better performance than other algorithms	May fall into local optima
Ziabari et al. [18]	Infinite feature selection method based on K-means clustering	On six benchmark datasets, this method outperforms the infinite feature selection method in terms of accuracy, runtime, and memory consumption	The algorithm needs to set the number of subspaces to be partitioned in advance, which may affect the feature selection results
Natarajan and Rebekka [19]	Small cell handover algorithm based on K-means clustering and dynamic load	Compared with traditional K-means clustering methods, this method improves energy efficiency by 20% and system throughput by 16%	The setting of dynamic load thresholds has a certain impact on switching decisions

In summary, although previous researchers have conducted a lot of studies to improve the recommendation accuracy and quality of RS, most of the studies still choose to use traditional recommendation algorithms. Due to the issues with these RMs—such as their excessive reliance on data and cold start—this study avoids using them and instead builds RMs using KMA, which is unaffected by the aforementioned issues.

3 E-Commerce data processing model based on improved SVD++ with improved K-Means

With the development of e-commerce, consumers of ECPs face a large number of choices when shopping. In addition, KMA is a method that can cluster the samples according to the similarity principle based on the data features [20-22]. Therefore, applying KMA to ECG RMs can achieve better interpretability than neural networks and machine learning algorithms [23]. However, the performance of this approach in RMs is impacted by the typical KMA's shortcomings, which include unpredictability in the selection of initial clustering centers, inappropriate feature weighting calculations, and inadequate attention to hidden features (HFs) in the data [24]. Therefore this study improves KMA so as to construct a RM for ECG data.

3.1 Hybrid improved SVD++ and K-Means recommendation model for E-Commerce Goods

The classic SVD++ algorithm's flow is depicted in Figure 1. The SVD++ algorithm is chosen to be fused with the KM model in order to enhance the latter's capacity for data processing because of its benefits in HF calculation. The SVD++ algorithm has the following problems. First, the SVD++ algorithm calculates the implicit feedback by training the implicit scores through multiple iterations [25]. Second, this SVD++ algorithm does not take enough consideration of realistic factors, such as the frequency of users' ratings versus the frequency of items being rated, which can affect the overall computational accuracy of the SVD++ algorithm [26]. A new and improved SVD++ algorithm is currently suggested to tackle the aforementioned issues. This work modifies the implicit feedback calculation to address the SVD++ algorithm's low computing efficiency drawback. The SVD++ technique is fused with KMA based on the principle of N-N in order to increase its computational accuracy. This improvement brings the recommendation results closer to the user's tastes.

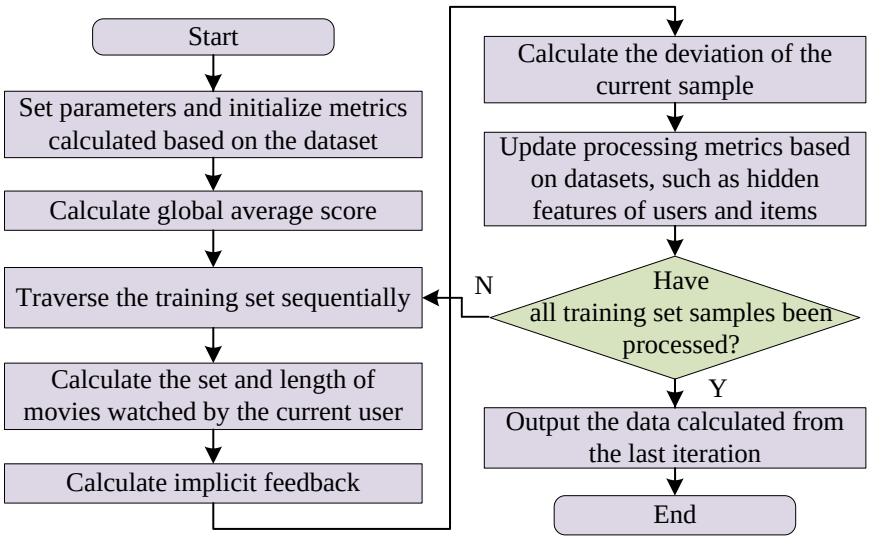


Figure 1: Running process of traditional SVD++algorithm

The computational inefficiency of the SVD++ algorithm is due to the fact that implicit feedback can only be obtained through implicit scoring training, so this study proposes to avoid this approach to obtain implicit feedback [27-28]. Implicit feedback also exists for items from the perspective of the recommended item (RI), but this feedback exists in a passive form due to the reciprocal nature of the interaction behavior between the item and the user [29]. The user's own implicit feedback behavior and the strength of the item's audience can be indirectly reflected when the user takes action on the item [30]. Therefore, in this study, the map is chose to the

user's implicit feature vector p_u to implicit feedback and superimpose it to the item's implicit feature q_i. After improving the implicit feedback calculation method, the SVD++ algorithm also needs to be improved to accommodate more elements that affect the quality of recommendation. When applied to the work with suggestions, the typical SVD++ algorithm frequently ignores the user's rating count and the number of times the item has been rated. In addition, these two points will bring obvious influence on the user's preference judgment. As a result, the SVD++ algorithm's cost function needs to be revised; Equation (1) displays the updated cost function.

$$\min_{p^*,q^*,b^*} \sum_{(u,i)\in trainset} \left(r_{ui} - \mu - b_u - b_i - \left(q_i - ppp[i]\right)^T \cdot p_u + \sqrt{\frac{L_u}{L_{all}}} + \sqrt{\frac{L_i}{L_{all}}} \right)^2 + \lambda\left(\|p_u\|^2 + \|q_i\|^2 + \|b_u\|^2 + \|b_i\|^2\right) \quad (1)$$

In Equation (1), $ppp[i]$ denotes the hotness of the corresponding item i on the HF, L_i denotes the total users who have rated item i. L_u is the total items rated by user u and L_{all} is the total samples in the training dataset. $\sqrt{\frac{L_u}{L_{all}}}$ and $\sqrt{\frac{L_i}{L_{all}}}$ represent the activity of the corresponding user and the total heat of the corresponding item, respectively. p_u is the user's HF and q_i is the item's HF. b_u and b_i denote the deviation of the user and the deviation of the item, respectively. r_{ui} is the rating of user u on item i calculated in Equation (2).

$$r_{ui} = q_i^T p_u \quad (2)$$

In Equation (1), both $ppp[i]$ and $\sqrt{\frac{L_i}{L_{all}}}$ can be interpreted as the hotness of the item, the former represents the hotness of the corresponding item on the HF, which can also be interpreted as the feature hotness. For example, a certain item has a high artistic component, and also many users go to buy and evaluate it, which leads to a higher heat of the item on the artistic component. That is to say, feature heat represents the length of the item in a certain feature dimension that makes the user favorite. $\sqrt{\frac{L_i}{L_{all}}}$ can be interpreted as the overall heat of the item, indicating that when an item is rated more times than the rest of the items, the item is able to appeal to more than one group of users on multiple feature dimensions. For example, in the field of commodity recommendation, a commodity contains artistic components, practicality, value preservation and so on at the same time, which can satisfy the needs of different consumer users. Furthermore, an item's rating count may be a good indicator of a user's general interest in it. If an item has no selling point or is difficult to arouse interest, its total number of ratings will be very small. Therefore, the total number of ratings of an item can be used to describe the overall hotness of the item.

So far the computational method of the HF generation module of ECG RM based on SVD++ algorithm with KMA, i.e. ISVD++ algorithm is designed. The computational flow of this algorithm is shown in Figure 2. As shown in Figure 2, the ISVD++ algorithm first needs to input the user's rating data for ECGs and construct rating matrices based on these rating data. The aforementioned rating matrices will then be divided in order to calculate the overall heat matrix of the item, the implicit feedback obtained from the user's implicit features, and the user's activity data. Furthermore, the high-end implicit feedback will be transformed into the feature specificity of the item. These data will be stored in a database for subsequent KMCA modules.

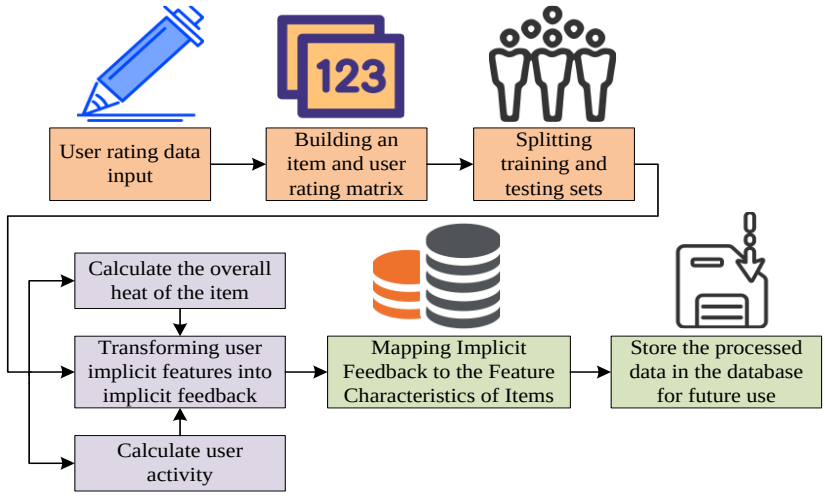


Figure 2: Calculation process of ISVD++algorithm in e-commerce product recommendation

Specifically, the first step is to construct a user-item rating matrix using the input data, i.e., the user's rating data, which will be used as input to the ISVD++ algorithm. But first, it's also required to separate the data into a training set and a test set before feeding the matrix into the ISVD++ algorithm. In the third step i.e. to start the model training, firstly L_{all}, L_i and L_u need to be calculated based on the training set. Then L_u is mapped as the user's activity metrics and BB is mapped as the item's overall hotness by using Equation (1). After the mapping computation is completed, iterative computation is started with the aim of transforming p_u into implicit feedback, and then finally obtaining the characteristic hotness of the item. Considering that the N-N users of the current user need to be obtained for ECG recommendation as a reference for predicting the interest of the current user, KMA constructed according to the N-N idea is chosen to design the e-commerce RM. Furthermore, KMA is better suited for the e-commerce recommendation task due to its straightforward implementation and high interpretability.

3.2 Improved design of KMA in recommendation modeling

The KMA in the ECG RM designed in this research has shortcomings such as unreasonable calculation of indicator weights, so the KMA is now improved. The coefficient of variation method is essentially an objective assignment method, which calculates the corresponding weight coefficients through the information in the indicators. This time, the coefficient of variation method is used to calculate the variable weights of the recommended commodities in the RM, which can play a role of reflecting the differences between items more objectively. In the case of the recommended commodities, the variable's significance for the present commodity increases with its difference; hence, the coefficient of variation approach will assign this variable a higher weight. Now the process of calculating the weights of item variables for the coefficient of variation method is specifically designed in a variable parameterized manner. Let a dataset contains n samples, each sample contains r variables, so the dataset N_{data} can be described according to Equation (3).

$$N_{data} = \begin{bmatrix} x_{11} & x_{22} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2r} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nr} \end{bmatrix} \tag{3}$$

In Equation (3), x_{np} represents the r th variable data for the n th sample. The mean $\overline{x}_j$ of indicator j can be described by Equation (4).

$$\overline{x}_j = \sum_{i=1}^{k} x_{ij} \Big/ n \tag{4}$$

Similarly, the standardized mean deviation σ_j of indicator j can be described by Equation (5).

$$\sigma_j = \left(\frac{1}{n-1} \sum_{i=1}^{k} \left(x_{ij} - \overline{x}_j \right) \right)^{1/2} \tag{5}$$

Therefore, the coefficient of variation V_j for indicator j can be calculated using Equation (6).

$$V_j = \sigma_j / \overline{x}_j \tag{6}$$

On the basis of Equation (6), the weight coefficients W_i for each indicator of sample i can be calculated by

Equation (7).

$$W_i = V_j \Big/ \sum_{j=1}^{r} V_j \tag{7}$$

KMA also has some pitfalls in the ICC selection. In this study, the ICC selection and updating method are first analyzed from the traditional KMA point of view, as shown in Figure 3. The quality of the current clustering results in Figure 3 is evaluated according to the index P calculated in Equation (8).

$$P = \sum_{i=1}^{n} \sum_{j=1}^{C} \left\| x_i - c_j \right\|^2 \tag{8}$$

In Equation (8), C represents the total cluster classes for the current data, x_i is the i th data within the j th cluster, and c_j is the j th cluster class centroid. In Figure 3, the number of clusters is first determined and then the cluster centroids are randomly initialized. Subsequently, the sample points are assigned by computing the distance of each sample from the C-C point. The average value of each cluster is determined and used as the new cluster clustering center once all the samples have been assigned to the closest C-C. The sample points are then redistributed using the previously described sample point allocation procedure. The ICCs are typically selected at random or weighted randomly based on the density of the data distribution; however, there is a degree of randomness in both approaches, which reduces the stability of the clustering outcomes.

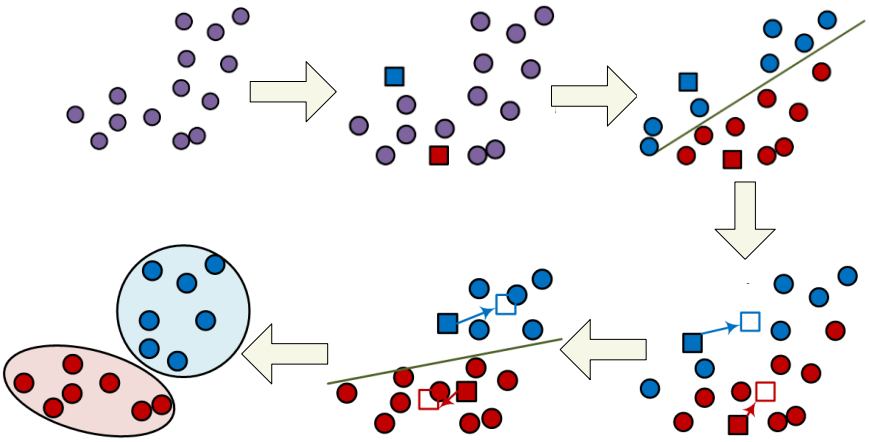


Figure 3: Traditional k-means algorithm C-C initialization and update method

GA as a type of heuristic intelligence algorithm, is extensively employed in the field of mathematical model and algorithm parameter optimization due to its superior capacity for automated computing and global optimization seeking. The system is based on the principle of natural selection and focuses on the encoding of decision variables as its primary operational objective. The algorithm does not require an understanding of the problem itself. Rather, it utilizes the fitness value of chromosomes as search information, retaining high-fitness chromosomes to eliminate low-fitness chromosomes. Ultimately, the optimal or quasi-optimal solution is obtained through population iteration. In Figure 4, the GA computational flow is displayed.

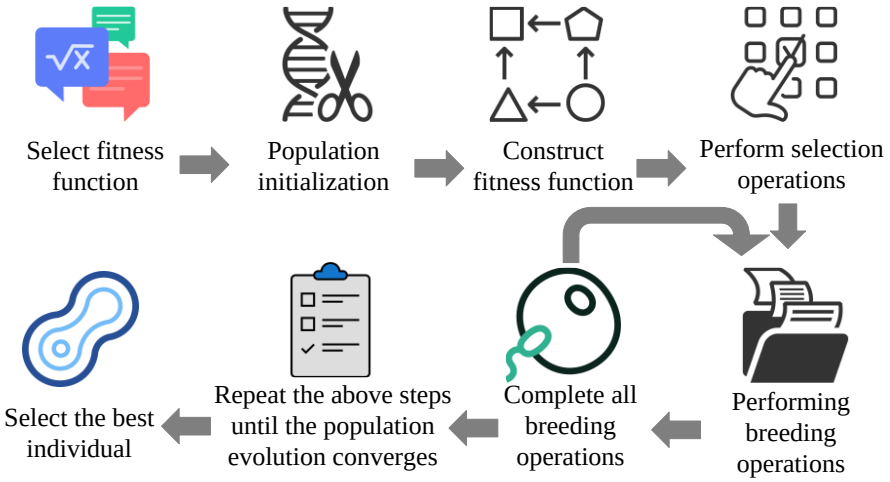


Figure 4: GA calculation process

Therefore, this study uses GA to optimize the ICC calculation of KMA. The optimal ICC can be identified through the use of the global search capability of the GA. This is followed by the implementation of the improved KMCA, which enhances the stability of the algorithm and yields satisfactory clustering results. The chromosome encoding method of GA has numerical encoding and binary encoding, but considering that the e-commerce data is more complex, it is more appropriate to choose binary encoding, as shown in Equation (8).

$$g\left(x_l^t,k\right)=u_k+\frac{u_k-v_k}{2l-1}\left(\sum_{j=1}^{t}x_t^{(kl+j)}\times 2^{j-1}\right) \quad (8)$$

In Equation (8), x_l^t represents the t th chromosome calculated to the l th generation. k is a real number parameter, u_k and v_k are the upper and lower limits of parameter k, respectively. $g\left(x_l^t,k\right)$ represents the gene position of x_l^t chromosome under the condition of parameter k. l represents the number of gene loci of the chromosome.

The first step of KM clustering using the GA is to binary encode the centroids and then select the appropriate ICCs according to the fitness function output by the GA. Since the research topic of this study is ECG recommendation, each set of similar e-commerce data can be regarded as a cluster. Suppose the dataset obtained by the RM from the database is $D=\left\{d_j\,|\,j=1,2,...,n\right\}$, and d_j is the j th data among them. Then let the number of C-Cs in D be $N_{k=}=\left\{C_i\,|\,i=1,2,...,k\right\}$ and C_i be the i th cluster. Cluster optimization is performed according to the standard deviation sum as a metric, which is shown in Equation (9).

$$f\left(P_k\right)=\sum_{i=1}^{k}\sum_{P\subset C_i}\left|P-m_i\right|^2 \quad (9)$$

In Equation (9), $f\left(P_k\right)$ represents the sum of standard deviations between clusters in the data set. m_i is the mean value of the C_i clusters and is calculated in Equation (10).

$$m_i=\frac{1}{t_i}\sum_{p\subset C_i}p \quad (10)$$

In addition, the similarity between the data is calculated according to Euler's formula as shown in Equation (11).

$$d\left(x_i,m_j\right)=\left(\sum_{i=1}^{d}\left(x_{it}-m_{jt}\right)^2\right)^{1/2} \quad (11)$$

In Equation (11), x_{it} and m_{jt} are the x_i data and cluster C_j means at the t th calculation, respectively.

So far the hybrid GA improved KMCA can be obtained and its computational flow is shown in Figure 5.

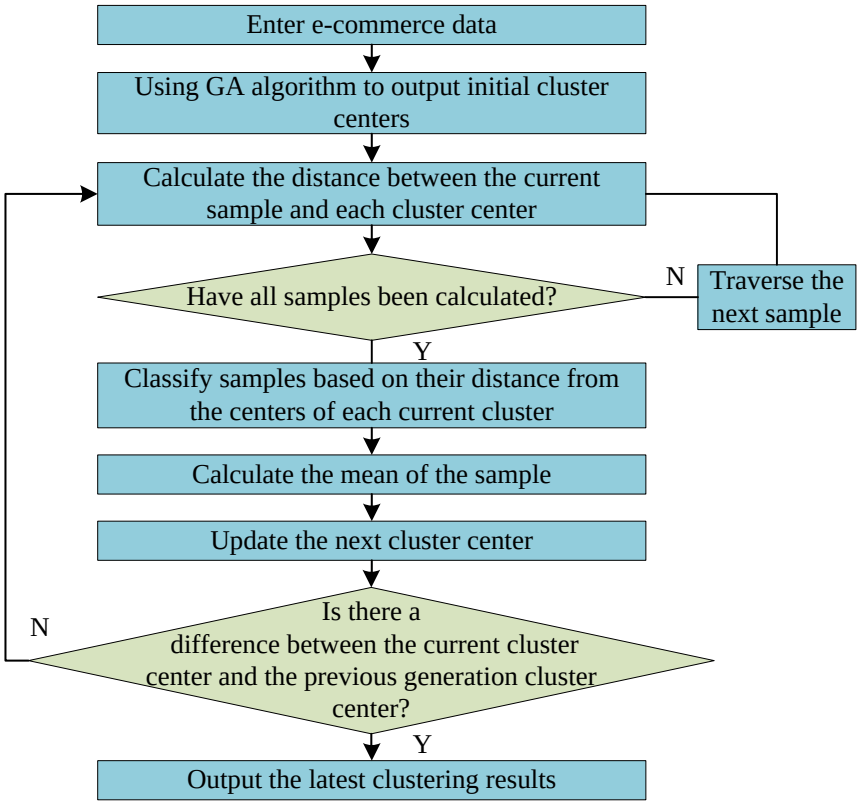


Figure 5: Calculation steps of improved k-means clustering algorithm for hybrid GA

In Figure 5, the clustering is excellent when Equation (9) is calculated to obtain the minimum value of the function, and this study follows this way of calculating the fitness, as shown in Equation (12).

$$f\left(R_i\right)=E_{\max}-E\left(R_i\right) \tag{12}$$

Equation (12) in which $E_{\max}$ represents the maximum value of chromosomal variance in population R and $E\left(R_i\right)$ represents chromosomal variance in population R_i of the i th population. $f\left(R_i\right)$ represents the calculated comfort level. In addition, after incorporating the GA, the selection probability is calculated in KM according to Equation (13).

$$P_i=f\left(R_i\right)\Big/\sum_{i=1}^{n}f\left(R_i\right) \tag{13}$$

In Equation (13), $n=1,2,...,n$, P_i represents the selection probability P_i for the i th data.

Now the GA is combined with the hybrid SVD++ designed above with the improved KM of the coefficient of variation method to construct the ECG-oriented RM. In addition, Figure 6 depicts the particular model structure. The first step of the model computation is to train the ISVD++ model, which aims to provide input data for the subsequent computation boards. In the first step, the cost function is optimized according to the stochastic gradient descent method, and the HF matrix of items and users can be output. Moreover, in the first step, the sparse rating matrix needs to be complemented and the descending process is performed according to the ratings to get the recommendation list of all users. The second step of the model is to cluster the users using improved KMA. The elbow technique is chosen in this case to identify the clusters or user groups K due to the sizeable user population in the dataset and the presence of variances depending on gender, position, age, interests, and other variables. Subsequently, the specific user u can be identified where the cluster, as determined by the KMA, can be located in the N-N of the user B. This serves as the basis for determining the user u 's preferences. Processing the user recommendation list and removing the items' current user ratings is the third step. Because these items are already known to the user, do not need to do a repeat of the recommendation, the rest of the list to intercept the first h items that is the user's recommendation of the product program. The fourth step is to measure whether an item needs to be recommended to the corresponding user according to the set strategy. The precise procedure is as follows: Determine the average score and the total frequency of identical items in

the suggested data. Additionally, the mean score and the frequency of comparable occurrences can be determined. Ultimately, the final list of recommendations is obtained by sorting the things in descending order of frequency.

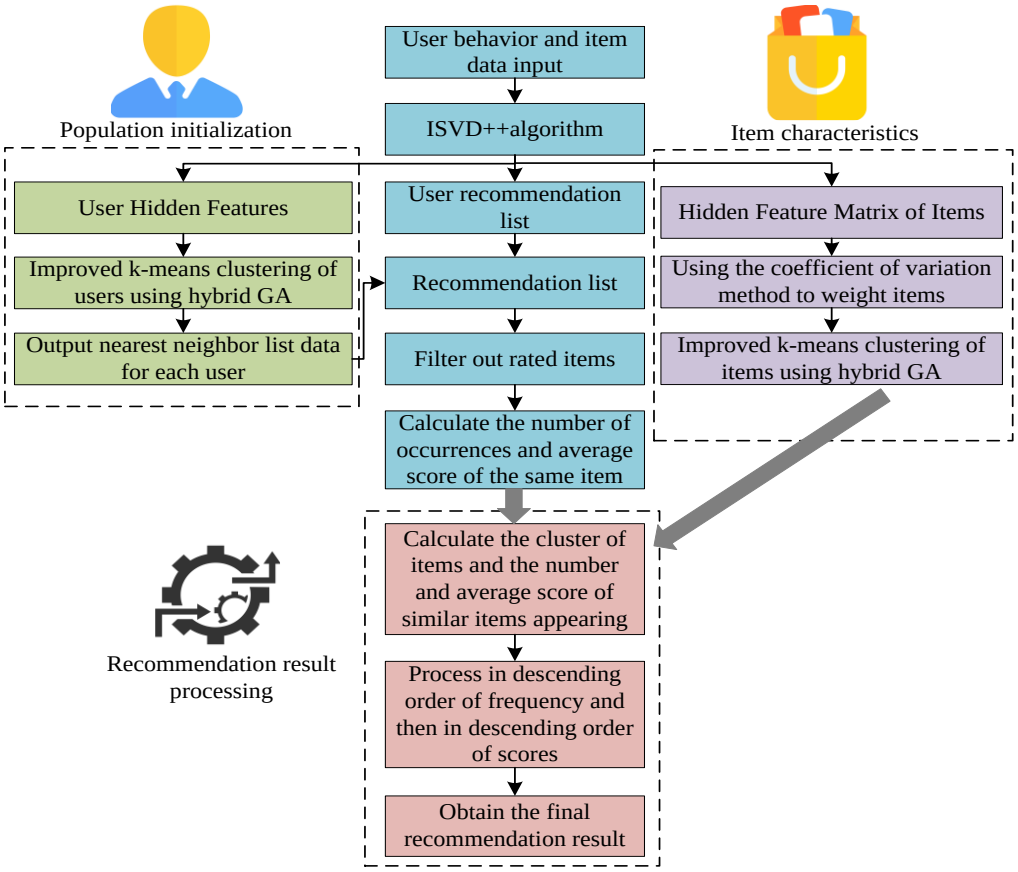


Figure 6: E-commerce product recommendation model structure of hybrid ISVD++and improved k-means algorithm

4 E-Commerce data processing model testing based on improved SVD++ with improved K-Means

To compare the effectiveness of the ECG data RM based on Improved SVD++ and Improved KM proposed in this research in the recommendation task, a test experiment is now being planned and carried out. The creation of the experimental protocol, the analysis of the ablation experiment's results, and the analysis of the control experiment's results comprise the test experiment.

4.1 Testing of experimental programs

The dataset chosen for the test experiments is obtained from Taobao, a Chinese online shopping platform, which contains a total of 5842 users and 6447 items. There are 827,384 user rating records. The experimental setup and parameters are as follows, taking into account the size of the dataset and the computational load of the recommendation task (Table 1). The parameters in Table 1 are specifically determined in a larger range of values according to the grid analysis method.

Table 1: Experimental environment and parameters

Type	Name	Number	Set results
Model parameter	Population size	#01	158
	Cross probability	#02	0.7
	Mutation probability	#03	0.05
	Selection method	#04	Tournament selection
	Chromosome length	#05	89
	GA iterations	#06	300

	K-mean algorithm convergence threshold	#11	0.001
	Distance measurement method	#12	Euclidean distance
Hardware	Central Processing Unit	#21	i7-7560U
	Hard disk	#22	Solid state drive, 512GB
	Memory	#23	16GB
	Graphics processing unit	#24	GTX1060
Software environment	Simulation software	#31	Jupyter Notebook
	Development environment	#32	Python3.7
	Operating system	#33	Windows 10 professional edition

The comparison algorithms SVD++, KM, and improved k-means with hybrid GA (GA-KM) are selected for the ablation experiments to construct the RM. Collaborative filtering (CF) is selected for the control experiments as well as the advanced temporal-aware recommendation algorithm (TRRA), reinforcement learning-based recommendation algorithm (RLRA) to construct the comparison model. The dataset requires result preprocessing with the aim of removing null values, illegal data, and useless features. The dataset is split in a 7:3 ratio between the training and test sets. Precision, recall, area under curve (AUC) of receiver operating characteristic curve (ROC), computation time consumed, and memory consumption of TOP-N are chosen as the evaluation indexes of model performance in the test.

4.2 Analysis of model ablation experiments

Initially, it examines the data from the ablation experiment. In Figure 7, the statistical findings for each comparison model's precision and recall throughout the training phase are displayed. When the number of suggestions is set to 30, Figure 7 displays the results of the computation. The model "ISVD++_I_KM" is created specifically for this study. As training times and quality increase, each model's Precision and Recall start to rise upward and then gradually stabilize. When more than 200 iterations are completed, each model's recommendation accuracy index fluctuates less. It can be assumed that the training of each model is completed when the number of iterations is 400, at which time the Precision and Recall of ISVD++_I_KM, GA-KM, SVD++, and KM models are 93.5%, 91.2%, 87.2%, and 82.9% versus 92.4%, 91.3%, and 86.4%, respectively, 81.7%.

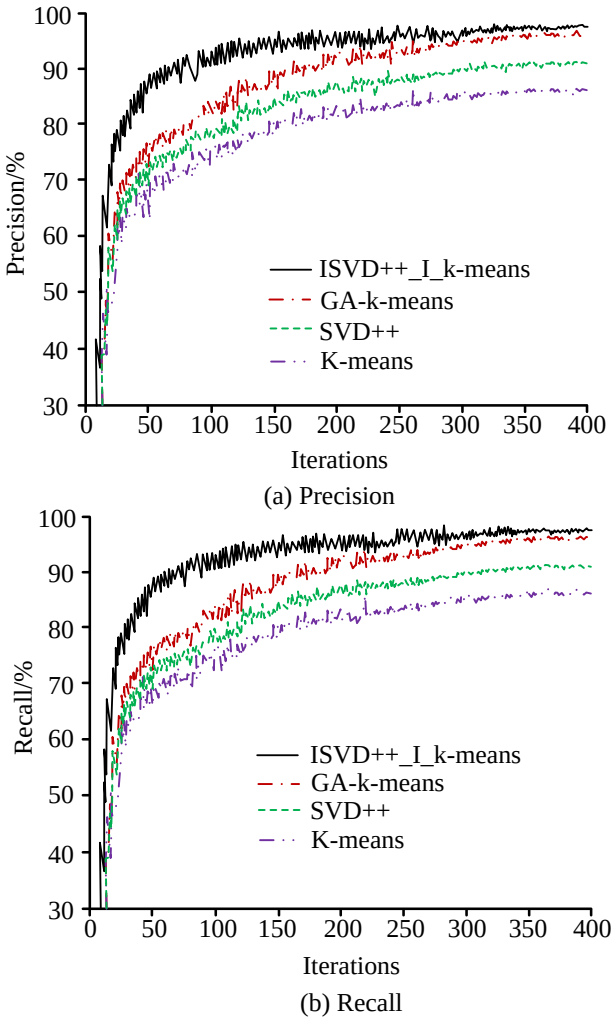


Figure 7: Precision and recall of each model during the training phase

Figure 8 displays the statistical findings of Precision and Recall for each model under various numbers of recommendations in the test set. The number of RIs in the output is shown by the values of each axis in Figures 8(a) and 8(b), respectively. The recommendation accuracy of each model exhibits an overall monotonically growing change trend when the number of RIs drops, regardless of Precision or Recall, while the growth rate is gradually decreasing. When the number of recommendations is 15, the Precision of ISVD++_I_KM, GA-KM, SVD++, and KM models are 87.9%, 85.3%, 82.9%, and 76.4%, respectively, and when the number of recommendations is 30, ISVD++_I_KM, GA-KM, SVD++, and KM models have Precision of 93.1%, 91.90%, 86.7%, and 83.0%, respectively. It can be noticed that after the number of RIs grows from 15 to 30, the recommendation accuracy does not show a large increase, and then considering that recommending too many items will increase the difficulty of the user's choice, so the recommendation number parameter of the subsequent ablation experiments is fixed at 15.

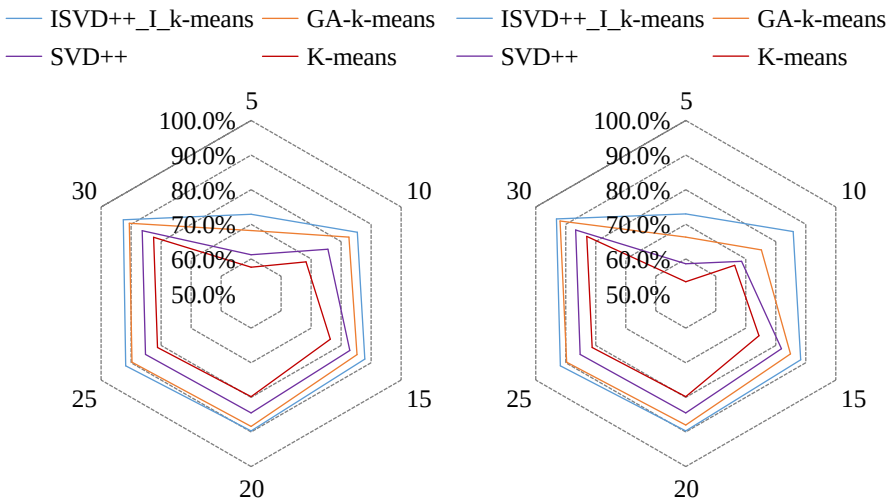


Figure 8: Precision and recall of ablation experimental models under different recommendation numbers

Finally, under the condition that the number of RIs is 15, all the metrics of the statistical ablation experimental model are shown in Table 2. The ISVD++_I_KM model has a greater accuracy and AUC than all of the comparative models combined. However, the computational time consumption with average memory consumption and maximum memory consumption of ISVD++_I_KM model is also higher than the comparison models.

Table 2: All performance indicators of the ablation experimental model under 15 recommended item conditions

Index	ISVD++_I_k-means	GA-k-means	SVD++	K-means
Top_N_Pre /%	85.4	85.3	82.9	76.4
Top_N_Rec /%	87.9	84.9	81.9	74.4
AUC	0.83	0.79	0.74	0.65
Average calculation time/s	54.2	47.3	36.9	45.1
Average memory consumption/MB	16.8	11.2	6.5	9.4
Maximum memory consumption/MB	17.1	11.8	8.9	13.5

4.3 Analysis of Model-Controlled experiments

The statistical results of the recommendation metrics of the ISVD++_I_KM model and the rest of the state-of-the-art models for each number of recommendations are shown in Figure 9. The recommendation accuracy of each model increases with the increase in the number of RIs. In addition, the Top_N_Pre and Top_N_Rec of ISVD++_I_KM model are always higher than the rest of the comparison models. When the number of RIs is 15, the Precision and Recall of ISVD++_I_KM, RLRA, TRRA, and CF models are 85%, 73%, 70%, and 59% versus 87%, 75%, 69%, and 58%, respectively.

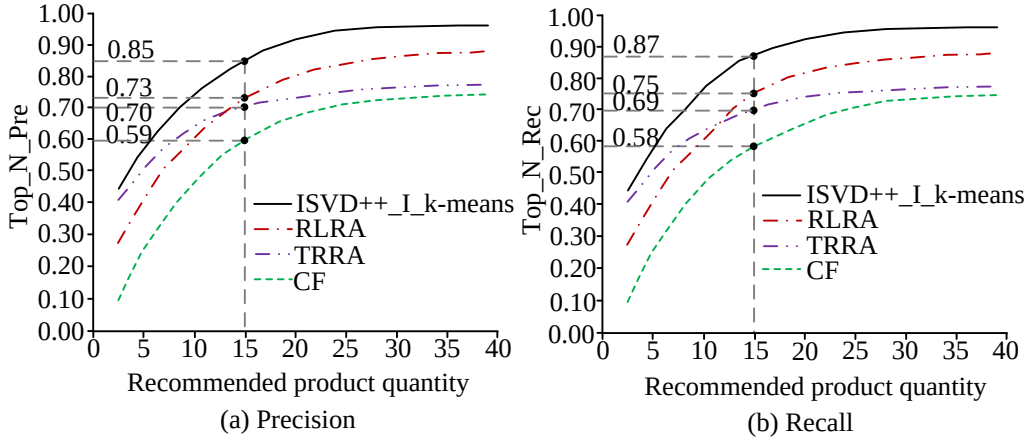


Figure 9: Comparison of recommendation accuracy among different models in the comparative experiment

The comparison of the ROC of each model and its AUC is shown in Figure 10. The ROC of ISVD++_I_KM, RLRA, TRRA, and CF models corresponds to an AUC of 0.83, 0.81, 0.78, and 0.76, respectively, and the former is significantly higher than the latter three.

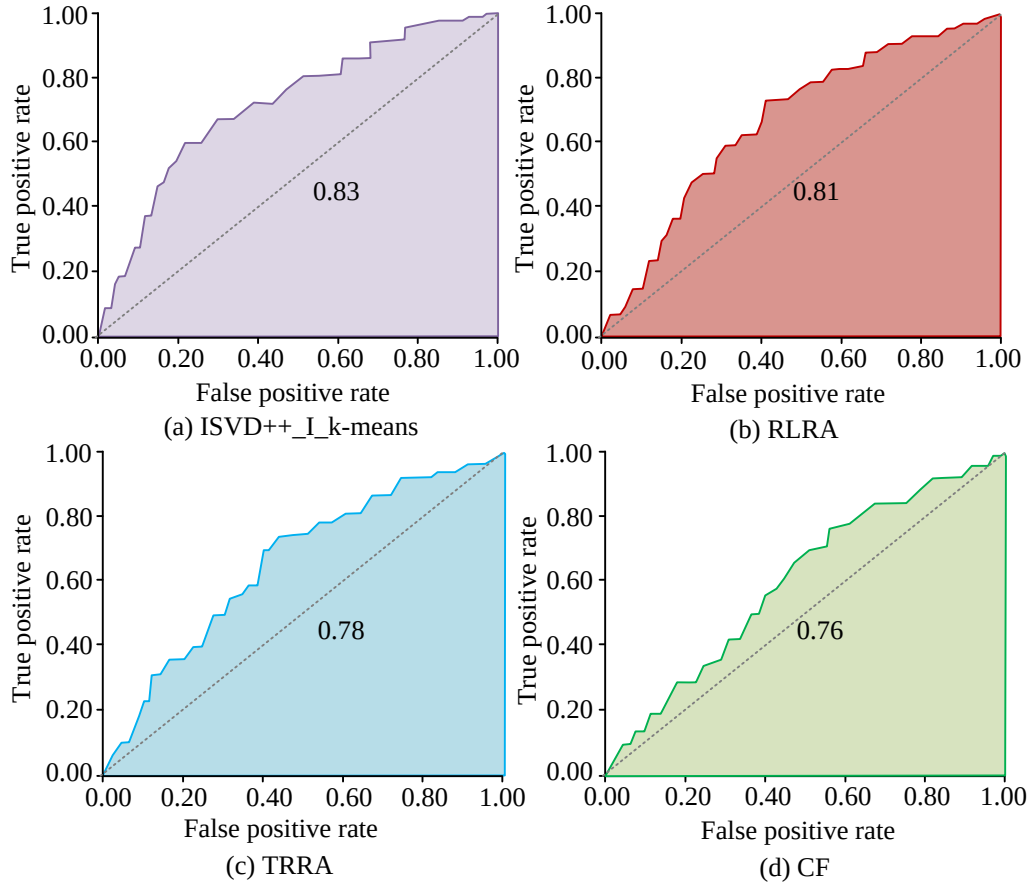


Figure 10: Comparison of ROC and AUC of various models

The computational elapsed time of each model is then counted and shown in Figure 11. The computational elapsed time of the ISVD++_I_KM model is only longer than the CF algorithmic model in different genders and different age groups. As a whole, the average computational elapsed time of ISVD++_I_KM, RLRA, TRRA, and CF models are 54.2s, 73.8s, 83.3s, and 58.7s, respectively. In terms of age segments, the computational elapsed time of recommendation for adult users is higher than that of minors, which is due to the fact that adults have a greater need for shopping and more relevant data. The reason why the recommendation consumption time

of female users is higher than that of males is also the same.

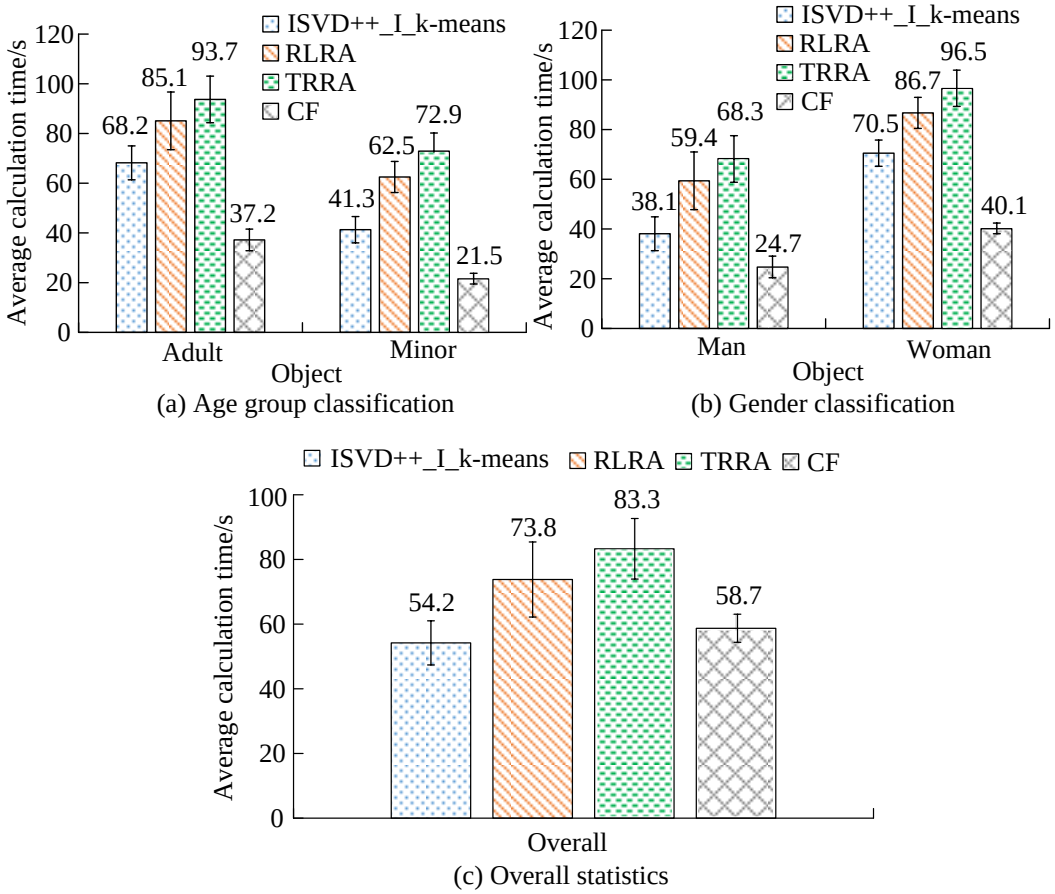


Figure 11: Comparison of computational time for various models

The computational memory consumption of each model is shown in Figure 12. To enhance the reliability of the statistical results, each experiment is repeated on multiple occasions. The data presented in Figure 12 represents the outcome of a single run. The median memory consumption of the ISVD++_I_KM, RLRA, TRRA, and CF models is 17.0MB, 12.6MB, 13.9MB, and 6.8MB, respectively. From the perspective of stability, the standard deviation of memory consumption for the ISVD++_I_KM, RLRA, TRRA, and CF models in Figure 12 are 3.24MB, 5.96MB, 4.88MB, and 4.57MB, respectively. This indicates that the models developed in this work operate with more stability, and that a higher level of redundancy in a commercialized product is not required to guarantee a smooth functioning of the models.

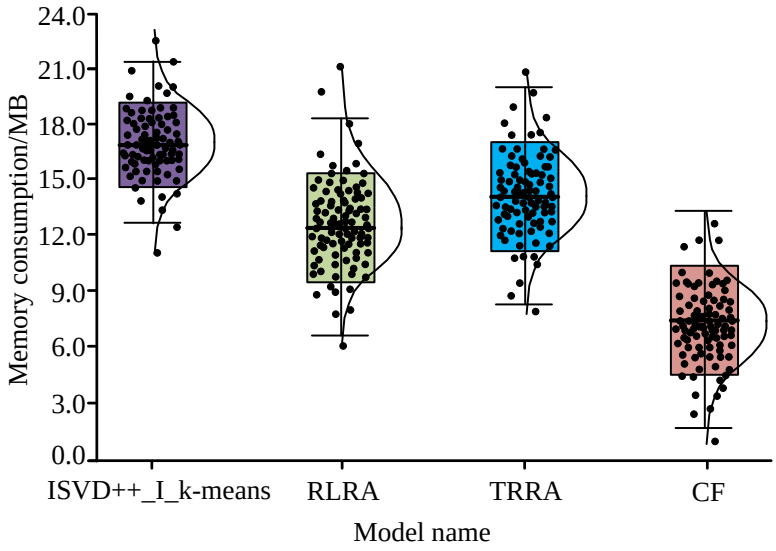


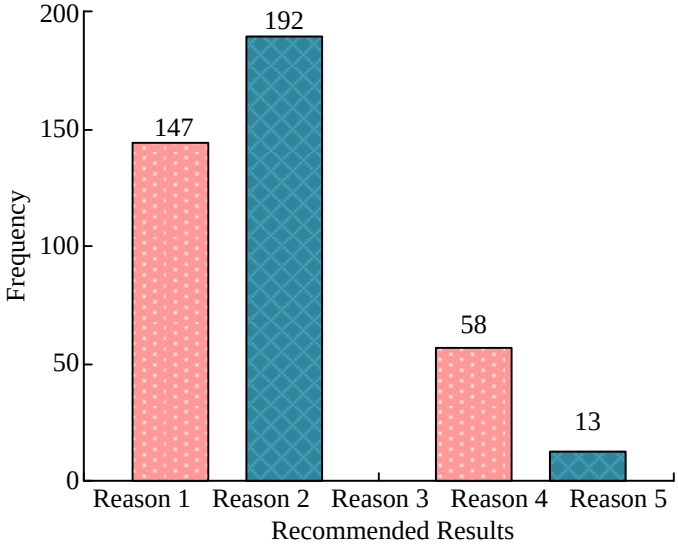
Figure 12: Comparison of computational memory consumption among different models

Next, the data of all performance evaluation indexes of each model in the control experiment are counted, and Table 3 is still counted in the way that the number of RIs is 15. Overall, the recommendation accuracy of the ISVD++_I_KM model developed in this work is substantially higher than that of the control model, and its computational time is slower than that of the more sophisticated RLRA and TRRA models. But the computational memory consumption is higher, with an average memory consumption of 16.8 MB, which is higher than that of all the comparison models.

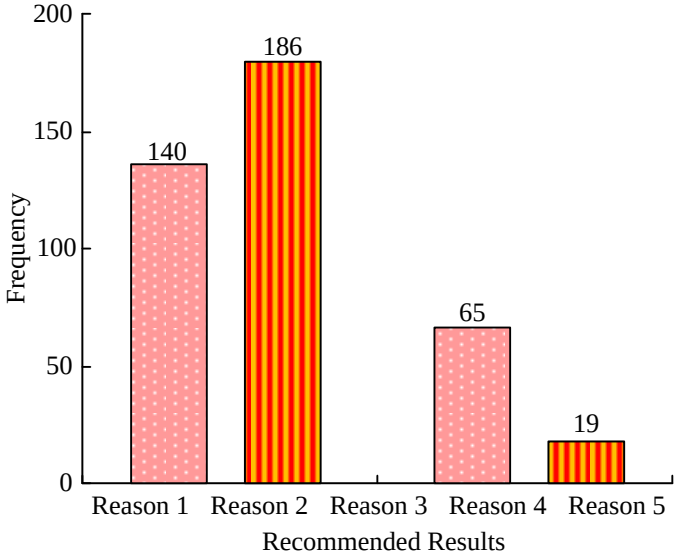
Table 3: Comparison of all performance evaluation indicators of each model in the control experiment

Index	ISVD++_I_k-means	RLRA	TRRA	CF
Top_N_Pre /%	85.4	73.1	69.8	59.2
Top_N_Rec /%	87.9	74.8	69.3	58.0
AUC	0.83	0.81	0.78	0.76
Average calculation time/s	54.2	73.8	83.3	58.7
Average memory consumption/MB	16.8	12.5	13.6	6.7
Maximum memory consumption/MB	17.1	16.6	20.5	13.9

Finally, to further validate the recommendation performance of the model, a comparison of ISVD++I_k means, RLRA, and TRRA models is conducted through A/B testing. A/B testing is a random testing method that allows for the comparison of hypotheses between two different objects. In A/B testing, two RMs are first deployed to the recommendation terminal simultaneously. Subsequently, user traffic from the spare parts platform website is allocated to the two RMs with a 50% probability for recommendation. Then, user behavior information is recorded on the recommendation results. In the final step, the recommendation indicators of the two models must be compared and analyzed based on the collected user feedback results. In the comparison between ISVD++ and RLRA, after parsing user recommendation logs, 70 successful recommendation records and 340 recommendation failure data are obtained. In the comparison between ISVD++I_k means and TRRA, 50 successful recommendation records and 310 recommendation failure data are obtained. The results are presented in Figure 13.



(a) Comparison between ISVD++I_k means and RLRA



(b) Comparison between ISVD++I_k means and TRRA

Figure 13: A/B test results of different models

Figure 13 (a) illustrates that the ISVD++I_k means model has 58 successful recommendation records, with a recommendation success rate of 28.8%. The RLRA model yielded 13 successful recommendation records, with a recommendation success rate of 6.34%. Figure 13 (b) illustrates that the ISVD++I_k means model has a successful recommendation record of 65, with a recommendation success rate of 31.7%. The RLRA model achieved a successful recommendation record of 19, with a recommendation success rate of 9.27%. In comparison to the other two models, the success recommendation rate of the ISVD++I_k means model has demonstrably improved, thereby corroborating the hypothesis that users are more satisfied with the model.

5 Discussion

The exponential growth of e-commerce has led to the accumulation of a vast quantity of intricate data, rendering it challenging for users to swiftly and accurately identify the information they are seeking amidst this deluge of information. The development of algorithms capable of providing personalized recommendations to users, assisting them in discovering products of interest and enhancing purchase rates and user satisfaction, has emerged as a pivotal research topic among professionals in this field. Although traditional CF algorithms and content-based recommendation algorithms have demonstrated improvements in recommendation performance, these algorithms often encounter limitations due to sparsity and cold start issues,

which result in average recommendation performance. In recent years, the application of deep learning algorithms in the domain of personalized recommendations has become increasingly prevalent. Conventional algorithms typically necessitate a substantial quantity of training data and exhibit high computational complexity. Consequently, the study integrates the conventional RM SVD++ with a straightforward and expedient unsupervised learning approach, K-means clustering, and optimizes them individually to develop an e-commerce data processing model, ISVD++I_KM, based on enhanced SVD++ and enhanced k-means.

The experimental results demonstrated that the accuracy, recall, and AUC of the ISVD++I_KM model are significantly superior to those of benchmark models such as SVD++ and k-means. This suggested that the ISVD++I_KM model is more effective than traditional methods in identifying complex interaction relationships between users and products on ECPs. In comparison to the session-based graph neural network RM proposed by Gwadabe and Liu [11], the ISVD++I_KM model was capable of directly modelling user-product interaction relationships, thus effectively circumventing issues that may arise due to data sparsity or model instability. In comparison to the ensemble model based on deep learning proposed by Choudhary et al. [13], the ISVD++I_KM model was not constrained by the requirement of ensuring that two outputs are within the same range. This demonstrated that it is more flexible and efficient. In terms of computational efficiency, the ISVD++I_KM model exhibited slightly higher time and space overhead than the benchmark model, yet it outperformed complex deep learning recommendation algorithms such as RLRA and TRRA. Furthermore, a comparison was made between the ISVD++I_KM model and the adaptive long short-term memory network designed by Rabiu et al. It was found that the ISVD++I_KM model does not require complex feature extraction processes, thus having significant advantages in computational efficiency.

In conclusion, although the ISVD++I_KM model exhibits some limitations in terms of memory usage, it effectively incorporates the strengths of its predecessors. The combination of the SVD++ and KMAs has led to significant improvements in key indicators such as accuracy and computational efficiency. These advances have the potential to expand the applicability of the model in the field of e-commerce recommendation.

6 Conclusion

This work used the SVD++ algorithm to create ECG RM, which was based on enhanced KMA. The ablation experiments' findings showed that while the growth rate was gradually slowing down, each model's suggestion accuracy generally displayed a monotonically improving trend as the number of RIs decreased. After the number of RIs increased from 15 to 30, the recommendation accuracy did not show a large increase. Under the condition that the number of RIs was 15, the accuracy and AUC of the ISVD++_I_KM model were higher than those of all the comparison models. The results of the control experiments revealed that when the RIs is 15, the Precision and Recall of the ISVD++_I_KM, RLRA, TRRA, and CF models are 85%, 73%, 70%, and 59% versus 87%, 75%, 69%, and 58%, respectively. The corresponding AUC of ROC for ISVD++_I_KM, RLRA, TRRA, and CF models were 0.83, 0.81, 0.78, and 0.76, respectively, and the former was significantly higher than the latter three. The average computation time for ISVD++_I_KM, RLRA, TRRA, and CF models was 54.2s, 73.8s, 83.3 s, and 58.7s. The median memory consumption of ISVD++_I_KM, RLRA, TRRA, and CF models was 17.0MB, 12.6MB, 13.9MB, and 6.8MB, respectively. On the whole, the ISVD++_I_KM model designed in this study was significantly better than the control model in terms of recommendation accuracy, and the computational speed was lower than the advanced RLRA and TRRA models, but the computational memory consumption was higher. The inability to develop commercial RS using the intended model and evaluate the system's effectiveness in actual application scenarios is the research's main shortcoming; this is an area that has to be explored more in the future.

References

[1] M. Ravakhah, M. Jalali, Y. Forghani, and R. Sheibani, "Balanced hierarchical max margin matrix factorization for recommendation system," Expert Systems, vol. 39, no. 4, pp. e12911.1-e12911.14, 2021. https://doi.org/10.1111/exsy.12911

[2] N. Jiang, L. Gao, F. Duan, J. Wen, T. Wan, and H. Chen, "SAN: Attention-based social aggregation neural networks for recommendation system," International Journal of Intelligent Systems, vol. 37, no. 6, pp. 3373-3393, 2021. DOI: https://doi.org/10.1002/int.22694

[3] A. Da'U, N. Salim, and R. Idris, "An adaptive deep learning method for item recommendation system," Knowledge-Based Systems, vol. 213, no. 8, pp. 106681.1-106681.12, 2021. https://doi.org/10.1016/j.knosys.2020.106681

[4] U. Yadav, N. Duhan, and K. K. Bhatia, "Dealing with pure new user cold-start problem in recommendation system based on linked open data and social network features," Mobile Information Systems, vol. 2020, no. 4, pp. 8912065.1-8912065.20, 2020. https://doi.org/10.1155/2020/8912065

[5] K. Benabbes, K. Housni, A. E. Mezouary, and A. Zellou, "Recommendation system issues, approaches and challenges based on user reviews," Journal of Web Engineering, vol. 21, no. 4, pp. 1017-1054, 2022. https://doi.org/10.13052/jwe1540-9589.2143

[6] N. Mohammadi and A. Rasoolzadegan, "A two-stage location-sensitive and user preference-aware recommendation system," Expert Systems with Applications, vol. 191, no. 4, pp. 116188.1-116188.25, 2022. https://doi.org/10.1016/j.eswa.2021.116188

[7] L. Li, Z. Zhang, and S. Zhang. "Hybrid algorithm based on content and collaborative filtering in recommendation system optimization and simulation," Scientific Programming, vol. 2021, no. 3, pp. 742709.1-742709.11, 2021. https://doi.org/10.1155/2021/7427409

[8] Y. Cui, "Intelligent recommendation system based on mathematical modeling in personalized data mining," Mathematical Problems in Engineering, vol. 2021, no. 9, pp. 6672036.1-6672036.11, 2021. https://doi.org/10.1155/2021/6672036

[9] M. C. Chiu, J. H. Huang, S. Gupta, and G. Akman, "Developing a personalized recommendation system in a smart product service system based on unsupervised learning model," Computers in Industry, vol. 128, no. 10, pp. 103421.1-103421.19, 2021. https://doi.org/10.1016/j.compind.2021.103421

[10] R. Shaw and B. K. Patra, "Cognitive-aware lecture video recommendation system using brain signal in flipped learning pedagogy," Expert Systems with Applications, vol. 207, no. 11, pp. 118057.1-118057.10, 2022. https://doi.org/10.1016/j.eswa.2022.118057

[11] T. R. Gwadabe and Y. Liu, "Improving graph neural network for session-based recommendation system via non-sequential interactions," Neurocomputing, vol. 468, no. 1, pp. 111-122, 2022. https://doi.org/10.1016/j.neucom.2021.10.034

[12] Z. Roozbahani, J. Rezaeenour, A. Katanforoush, and A. J. Bidgoly, "Personalization of the collaborator recommendation system in multi-layer scientific social networks: A case study of ResearchGate," Expert Systems, vol. 39, no. 5, pp. e12932.1-e12932.18, 2021. https://doi.org/10.1111/exsy.12932

[13] C. Choudhary, I. Singh, and M. Kumar, "SARWAS: Deep ensemble learning techniques for sentiment based recommendation system," Expert Systems with Applications, vol. 216, no. 4, pp. 119420.1-119420.8, 2023. https://doi.org/10.1016/j.eswa.2023.119420

[14] I. Rabiu, N. Salim, A. Da'U, and M. Nasser, "Modeling sentimental bias and temporal dynamics for adaptive deep recommendation system," Expert Systems with Applications, vol. 191, no. 4, pp. 116262.1-116262.15, 2022. https://doi.org/10.1016/j.eswa.2021.116262

[15] V. N. Pattwakkar, S. Kamath, M. K. Nanjundappa, and R. Kadavigere, "Automatic liver tumor segmentation on multiphase computed tomography volume using segnet deep neural network and k-means clustering," International Journal of Imaging Systems and Technology, vol. 33, no. 2, pp. 729-745, 2023. https://doi.org/10.1002/ima.22816

[16] Z. Sun, Y. Hu, W. Li, S. Feng, and L. Pei, "Prediction model for short-term traffic flow based on a k-means-gated recurrent unit combination", IET Intelligent Transport Systems, vol. 16, no. 5, pp. 675-690, 2022. https://doi.org/10.1049/itr2.12165

[17] H. M. Mohammed, Z. K. Abdul, T. A. Rashid, A. Alsadoon, and N. Bacanin, "A new k means grey wolf algorithm for engineering problems," World Journal of Engineering, vol. 18, no. 4, pp. 630-638, 2021. https://doi.org/10.1108/WJE-10-2020-0527

[18] S. F. H. Ziabari, S. Eskandari, and M. Salahi, "Clnf-fs_s: An efficient infinite feature selection method using k-means clustering to partition large feature spaces," Pattern Analysis and Applications, vol. 26, no. 4, pp. 1631-1639, 2023. https://doi.org/10.1007/s10044-023-01189-1

[19] J. Natarajan and B. Rebekka, "An energy efficient dynamic small cell on/off switching with enhanced k-means clustering algorithm for 5g hetnets," International Journal of Communication Networks and Distributed Systems, vol. 29, no. 2, pp. 209-237, 2023. https://doi.org/10.1504/IJCNDS

[20] L. Wang, Z. You, D. Huang, and J. Li, "MGRCDA: Metagraph recommendation method for predicting circrna-disease association," IEEE Transactions on Cybernetics, vol. 53, no. 1, pp. 67-75, 2021. https://doi.org/10.1109/TCYB.2021.3090756

[21] S. N. Amin, P. Shivakumara, T. X. Jun, K. Y. Chong, D. Zan, and R. Rahavendra, "An augmented reality-based approach for designing interactive food menu of restaurant using android," Artificial Intelligence and Applications, vol. 1, no. 1, pp. 26-34, 2023. https://doi.org/10.47852/bonviewAIA2202354

[22] C. Fan, "Evaluating employee performance with an improved clustering algorithm," Informatica, vol. 46, no. 5, pp. 123-128, 2022. https://doi.org/10.31449/inf.v46i5.4079

[23] T. Chen, "A fuzzy ubiquitous traveler clustering and hotel recommendation system by differentiating travelers' decision-making behaviors - ScienceDirect," Applied Soft Computing, vol. 96, no. 1, pp. 106585.1-106585.10, 2020. https://doi.org/10.1016/j.asoc.2020.106585

[24] Z. Abbasi-Moud, H. Vahdat-Nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," Expert Systems with Applications, vol. 167, no. 4, pp. 114324.1-114324.10, 2021. https://doi.org/10.1016/j.eswa.2020.114324

[25] P. Mazumdar, B. K. Patra, and K. S. Babu, "Cold-start point-of-interest recommendation through crowdsourcing," ACM Transactions on the Web (TWEB), vol. 14, no. 4, pp. 19.1-19.36, 2020. https://doi.org/10.1145/3407182

[26] S. Bin and G. Sun, "Matrix factorization recommendation algorithm based on multiple social relationships," Mathematical Problems in Engineering, vol. 2021, no. 9, pp. 6610645.1-6610645.8, 2021. https://doi.org/10.1155/2021/6610645

[27] Q. Liang, X. Zheng, Y. Wang, and M. Y. Zhu, "O3ERS: An explainable recommendation system with online learning, online recommendation, and online explanation," Information Sciences, vol. 562, no. 7, pp. 94-115, 2021. https://doi.org/10.1016/j.ins.2020.12.070

[28] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 5-9, 2022. https://doi.org/10.47852/bonviewJCCE2202192

[29] B. Chen, L. Zhu, D. Wang, and J. Cheng, "Research on the design of mass recommendation system based on lambda architecture," Journal of Web Engineering. vol. 20, no. 6, pp. 1971-1990, 2021. https://doi.org/10.13052/jwe1540-9589.20614

[30] Q. Zhu, "Network course recommendation system based on double-layer attention mechanism," Scientific Programming, vol. 2021, no. 13, pp. 7613511.1-7613511.9, 2021. https://doi.org/10.1155/2021/7613511.