Research Article

Research on the Innovative Marketing Model of E-Commerce Platform Driven by Big Data in the Era of Network Economy

Xiaofeng Zhou and Ying Zhou

Jiangxi University of Technology, Nanchang 330098, China

Correspondence should be addressed to Ying Zhou; zhouying@jxut.edu.cn

Received 10 July 2022; Revised 2 August 2022; Accepted 13 August 2022; Published 29 August 2022

Academic Editor: R. Mo

Copyright © 2022 Xiaofeng Zhou and Ying Zhou. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the booming development of Internet information technology, e-commerce platforms in the era of network economy have undergone great changes, triggering a new marketing model change. Innovative research on marketing models can help the transformation and development of small and medium-sized e-commerce companies, which has important practical significance and theoretical value. The prediction of e-commerce sales is one of the key aspects of the evaluation of innovative marketing models, and only an accurate prediction of future sales can lead to a reasonable marketing plan. Therefore, a big data-driven e-commerce sales forecasting method is proposed. First of all, for 1703 real e-commerce companies, a large number of relevant data that affect sales are selected, including sales records, product information, product evaluation, and other information. A knowledge graph was then used to preprocess the data samples to produce a sample set containing concepts, entities, and relationships. Next, the knowledge graph K-modes clustering model is established. By fixing the affiliation matrix and the clustering cluster matrix in turn, the minimum of the objective function is continuously solved to obtain the cluster centres. Finally, sales prediction is achieved based on the clustering results. The experimental results show that the proposed clustering model is able to obtain better performance in terms of cluster purity, NMI, and F-value. The proposed clustering model has high sales prediction accuracy and has certain reference value for e-commerce enterprises of different scales to formulate innovative marketing models.

1. Introduction

Of all the innovative modes of business management, the innovation of the marketing model is very important. Finding the right marketing model is fundamental to a company’s survival. Most scholars see the marketing model as a synergistic system consisting of several elements. The ultimate goal of marketing is to gain competitive advantage and achieve profitability. The marketing system therefore has to specify a holistic solution to achieve this goal [1–4]. The information technology revolution has provided space and impetus for marketing model innovation. In the current era of network economy, network and mobile communication technologies are changing the way of production and life of human beings in an all-round way. Therefore, marketing model innovation must also be coupled with information technology to play an important role. The Internet economy is developing rapidly and big data-driven innovation models have become a popular research direction at present [5–8]. Many companies are now using big data technology to achieve knowledge mining. At the same time, the full sharing of off-site resources is achieved through the network, shortening the distance between time and space. Real-time knowledge mining driven by big data can maintain the continuity and uninterrupted nature of innovation.

The China Internet Information Network Centre (CIINC) has published the 41st Statistical Report on the Development of the Internet in China [9–11]. This report showed that by December 2020, the number of online consumers in China would reach 772 million, with 40.74 million new people. Compared to 2019, the number of
online consumers has increased by 2.6%. The Internet penetration rate has reached 55.8%, as shown in Figure 1. According to the China Electronic Commerce Research Centre, e-tailing transactions reached 3.1 trillion yuan in 2020, up by 34.8% year on year compared to 2.3 trillion yuan in 2019, as shown in Figure 2.

In recent years, the number of small e-commerce companies engaged in marketing via the Internet is increasing year on year. E-commerce platforms such as Alibaba, Amazon, and eBay are home to a huge number of small e-commerce companies. These small e-commerce companies, like other traditional businesses, are faced with the problem of marketing model innovation. Among all the innovations, the innovation of the marketing model is fundamental to the business. There are three broad categories of marketing models [12, 13]: value creation models, ecosystem models, and profitability models. The three types of theories discuss the connotation of the marketing model from different perspectives. A marketing model is an ecosystem consisting of many elements. This study considers the core element of the marketing model to be the profitability model. Profitability is a necessary condition for the existence of a market player. A business cannot sustain its basic survival without profit. For the many small e-commerce companies in e-commerce platforms, profit model innovation is their main concern.

The inadequate information management systems of small and medium-sized e-commerce businesses make it more difficult for them to innovate their profit models. In addition, small and medium-sized e-commerce enterprises face huge risks in the process of marketing reform because they cannot afford the larger capital costs. As a result, small and medium-sized e-commerce enterprises can only rely on forecast data of future sales to monitor their own business risks in real time [14, 15]. Sales forecasting can greatly improve the flow of capital for small and medium-sized e-commerce enterprises, allowing them to use these funds to develop broader sales channels or to cope with turbulent changes in the market environment. The accuracy of sales forecasting is therefore relevant to all aspects of an e-commerce business and will have a direct impact on its profit and loss, its next steps in financing, and its survival.

There are many models on sales forecasting, such as linear regression models, time series exponential smoothing forecasting models, logistic regression models, convolutional neural networks, and clustering mining [16–18]. In the preparation stage of each of these forecasting models, a sufficient amount of historical data has to be collected in order to predict the results with high accuracy. Therefore, these prediction models are all big data-driven models. Clustering analysis is an important application technique in data mining and is widely used in practical problems. Due to the development of information technology, consumer business data have become very large. With the continuous accumulation of data and the emergence of new business behaviours, it becomes difficult to classify data based on a priori experience. Therefore, Erenko et al. [19] proposed the application of cluster analysis to business marketing problems. Kingsland et al. [20] proposed the use of cluster analysis for data mining of consumer behaviour to give valuable guidance for future business decisions. Elizabeth et al. [21] proposed a cluster analysis-based model for e-commerce platform sales risk prediction model. From the perspective of microfinance companies, the types of e-commerce financial risks and the causes of their formation are analysed.

Overall, clustering analysis can uncover new patterns in large amounts of data, which can be quite a powerful tool for data processing problems in the marketing field. By applying clustering algorithms, new patterns can be obtained that are not influenced by previous experience, thus allowing a more comprehensive exploitation of the information contained in the data. One of the most widespread applications of clustering analysis is the classical K-means algorithm [22]. However, the K-means algorithm is only applicable to datasets with ordered categorical attributes. To solve the clustering problem for unordered categorical attributes, Huang et al. [23] proposed the K-modes algorithm based on the classical K-means algorithm. Under the framework of the original K-modes algorithm, Saha et al. [24] proposed the genetic fuzzy K-modes algorithm, which improved the accuracy of data mining to a certain extent.

The aim of this study is to use the K-modes algorithm to mine the historical data of small and medium-sized e-commerce businesses to achieve an accurate forecast of
future sales, thus providing data support for e-commerce businesses of different sizes when developing innovative marketing models. In the sales forecast of goods, the demand for products in the e-commerce industry is unstable as people’s hobbies, consumption habits, and other factors are changing all the time. This phenomenon leads to no clear pattern in sales volume trends. In addition, sales are influenced by various external social factors such as climate and fashion trends. These changes can often affect the predicted outcome of a product, so it is essential to extract the key attributes of the data before clustering analysis.

The main innovations and contributions of this paper include the following:

1. The K-modes algorithm in cluster analysis is introduced into the field of e-commerce marketing forecasting research, thus providing data support for e-commerce enterprises of different sizes in developing innovative marketing models.

2. To further improve the clustering accuracy, the knowledge graph technique [25, 26] was used to extract the key attributes of the historical sales data before the clustering analysis of the historical sales data using the K-modes algorithm. After the knowledge graph analysis, K-modes algorithm was able to obtain a higher clustering accuracy.

The rest of the paper is organized as follows. In Section 2, the knowledge graph is studied in detail, while Section 3 provides the proposed knowledge graph K-modes clustering algorithm. In Section 4, the KGK-modes based e-commerce sales volume prediction model is studied in detail, while Section 5 provides experimental results and analysis. Finally, the paper is concluded in Section 6.

2. Knowledge Graph

For the marketing model of the e-commerce industry, accurate clustering of relevant sales data samples is more difficult due to the variability of the merchandise. This is because people’s preferences, consumption habits, and other factors change all the time when forecasting sales of products. This phenomenon leads to no clear pattern in sales volume trends. Sales are also influenced by external social factors such as weather, trends, and so on. These changes ultimately lead to highly unstable forecasting results for commodities. In addition, if the key attributes of the sample are not extracted correctly, this may lead to incorrect clustering results. The above analysis shows that if the core attributes contained in the data sample can be identified, the scope of the clustering can be further reduced, thus effectively improving the accuracy of the clustering. Therefore, before using the K-modes algorithm to cluster the historical sales data, this paper uses the knowledge graph technique to extract the key attributes of the historical sales data.

The knowledge graph uses quadruples to represent knowledge [27], mainly containing concept, entity, relation, and attribute. The structure of the knowledge graph is shown in Figure 3.

Suppose the set of all knowledge elements in knowledge domain \( d \) is \( KE_d \):

\[
KE_d = \{ke_1, ke_2, \ldots, ke_i, \ldots, ke_n\},
\]

where \( ke_i \) represents the \( i \)-th knowledge element. Each knowledge element contains the concept knowledge \( c_i \), entity knowledge \( e_i \), relation knowledge \( r_i \), and attribute knowledge \( a_i \).

The set of concepts, entities, and relationships within knowledge domain \( d \) is denoted as \( C_d \), \( E_d \), and \( R_d \), respectively.

\[
C_d = \{e_1, e_2, \ldots, e_k, \ldots, e_n\},
\]

\[
E_d = \{c_1, c_2, \ldots, c_k, \ldots, c_n\},
\]

\[
R_d = \{r_1, r_2, \ldots, r_k, \ldots, r_n\},
\]

where \( nc \), \( ne \), and \( nr \) represent the total number of concepts, entities, and relations, respectively.

The set of attributes corresponding to each concept \( A_{ci} \) is \( A_{ci} \):

\[
A_{ci} = \{a_1, a_2, \ldots, a_j, \ldots, a_{na}\},
\]

where \( na \) indicates the total number of attributes.

First, the complex commodity data are classified into knowledge sets. Next, knowledge unit parsing is performed. Finally, the knowledge elements and graphs contained in the knowledge units are extracted [28]. The knowledge graph is obtained by layer-by-layer analysis, where the scale structure of knowledge is shown in Figure 4.

3. The Proposed Knowledge Graph K-Modes Clustering Algorithm

3.1. Principle of the K-Modes Algorithm. The K-modes algorithm is a divisional clustering algorithm used to solve the problem of clustering categorical attributes [29–31]. The basic idea of the K-modes algorithm is the same as the classical K-means algorithm, but it introduces a different distance metric and a centroid selection method. Suppose the set of sample points for a classification attribute is
$X = \{x_1, x_2, x_3, \ldots, x_n\}$. Each sample point contains $m$ attributes. The samples are divided into $k$ clusters $z$; then, the minimisation objective function of K-modes clustering is $P(W, Z)$:

$$P(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \omega_{i,l} D(x_i, z_l),$$

$$\sum_{l=1}^{k} \omega_{i,l} = 1, 1 \leq i \leq n,$$

where $W$ is the binary subordination matrix of $n \times k$ and $Z = \{z_1, z_2, \ldots, z_k\}$ is a matrix containing $k$ centroid [32].

$D(x_i, z_l)$ is the distance from a sample point to a centroid. The distance is calculated in the K-modes algorithm.

$$D(x, y) = \sum_{j=1}^{m} d(x_j, y_j),$$

$$d(x_j, y_j) = \begin{cases} 1, & \text{if } x_j = y_j, \\ 0, & \text{if } x_j \neq y_j. \end{cases}$$

The objective function problem $P$ is transformed into two subproblems. Let $\tilde{Z}$ and $\tilde{W}$ be the current optimal solutions. After each solution, we need to update $(W$ and $Z)$ and save the result of this update to the database in order to continue solving for the minimum of $P(W, Z)$.

$$\omega_{i,l} = \begin{cases} 1, & \text{if } D(x_i, z_l) \leq D(x_i, z_t) \text{ for } 1 \leq t \leq k, \\ 0, & \text{for } t \neq l. \end{cases}$$

When $W = \tilde{W}$, set $z_{i,l}$ to the attribute value of the $j$-th component of the cluster $z_l$.

$$\sum_{i=1}^{n} D(x_i, z_l) = \sum_{i=1}^{n} \sum_{j=1}^{m} d(x_{i,j}, z_{l,j}) = \sum_{j=1}^{m} \left( \sum_{i=1}^{n} d(x_{i,j}, z_{l,j}) \right) = \sum_{j=1}^{m} n \left( 1 - \frac{n_{z,l,j}}{n} \right),$$

where $z_l$ is the attribute value of the cluster $Z_l$. Solve for the distance values in turn until the minimum value of $z_l$ is found, thus obtaining the centroid and the class of each centroid.

The original K-modes algorithm takes a simple matching similarity measure based on the Hamming distance [33]. When the values of an attribute of two sample points are equal, their similarity on that attribute is 1; otherwise, the similarity is 0. This distance metric method may randomly assign less similar objects when assigning sample points, resulting in weaker intra-cluster similarity [34]. Thus, the original K-modes algorithm has limitations when mining sales-related data of small and medium-sized e-commerce companies. This is because there is no obvious pattern in the trend of sales volume changes. In addition, the sales volume of a product is influenced by various external social factors, such as climate, fashion trends, and so on. This problem causes the original K-modes algorithm to rely heavily on a priori experience. Therefore, in order to further improve the clustering accuracy, this paper uses the knowledge graph technique to extract the key attributes of historical sales data before using the K-modes algorithm for clustering analysis.

3.2. The Proposed Knowledge Graph K-Modes (KGK-Modes) Algorithm. Firstly, the proposed KGK-modes algorithm will analyse the data related to e-commerce sales and generate a new sample set containing the knowledge graph quadrants. Then, K-modes clustering is performed on the new sample set. Firstly, $k$ cluster centres are determined, and $k$ clusters are randomly selected from the sample as cluster centres, thus forming the initial cluster set $Z_1$. A suitable $W_1$ is found so that $P(W_1, Z_1)$ is minimised. During the updating process, stop updating if $P(W_t, Z_{t+1}) = P(W_{t+1}, Z_{t+1})$ is satisfied; otherwise, continue updating $W_{t+1}$. The steps of the KGK-modes algorithm implementation are shown in Figure 5.

4. Big Data-Driven KGK-Modes-Based E-Commerce Sales Volume Prediction Model

4.1. Source of Data and Preprocessing. The data source for this study is the daily order volume, sales, number of customers, and number of reviews for 2,700 e-commerce companies on the Taobao website (https://www.taobao.com/). The time frame is from August 3, 2019, to April 30, 2020. The dataset was filtered using desensitisation rules as the data needed to be collected in a way that protected the privacy of the customer. The data deviate from the real e-commerce business data. However, this deviation did not affect the exploration and research of this solution. First, the three tables required (order, review, and product information) were merged into one table based on the e-commerce number in the SQL database. Then, the data were cleansed. The missing values in the fields were replaced by the median and mean values of the column. E-commerce numbers with only one or two items are removed directly. After a hierarchical process, the final selection of 1703 e-commerce companies with at least seven or more types of products was made.

Correction of missing data, duplicate data, and incorrect data contained in datasets related to e-commerce sales volume was done according to statistical methods. Records
4.2. Sales Forecasting Using KGK-Modes Clustering. In this paper, the proposed KGK-modes algorithm is used to cluster the data of 1703 e-commerce companies, so as to complete the sales forecast. Eight variables are first set in the data of these e-commerce companies: sale_amt (sales), offer_amt (offer amount), offer_cnt (number of offers), rtn_cnt (number of returned orders), ord_cnt (number of orders), bad_num (number of negative comment), and good_num (number of positive comment). Based on these 8 variables, the total data of 1703 e-commerce companies were calculated. The purpose of clustering with all variables is that the same type of e-commerce can be combined together, so as to forecast the sales of each type of e-commerce. The total daily data of the e-commerce companies are shown in Table 1.

Before clustering, we first performed a data check, i.e., descriptive statistics of the data [35], such as mean and standard deviation. It is found that the mean and variance are very small, so there is no need to standardize the data.

After analysing 1703 e-commerce companies by KGK-modes clustering algorithm, these e-commerce companies can be roughly divided into three categories. The first category is the e-commerce company with the largest variety of goods and a favorable rate of over 99%. In the first category, each company has more than 30 kinds of commodities. The number of the first type of e-commerce companies is 39, accounting for 2.3% of the total number. The second category includes e-commerce companies with 10–30 commodity types and 93%–98% favorable rate. The number of the second type e-commerce companies is 1,129, accounting for 66.3% of the total number. The third category includes e-commerce companies with less than 10 kinds of goods and a favorable rate of less than 93%. The number of the third type e-commerce companies is 535, accounting for 31.4% of the total number.

5. Experimental Results and Analysis

In order to verify the performance of KGK-modes in e-commerce sales forecasting, standard dataset tests and real case tests were conducted. The experimental hardware environment was a desktop computer with 64-bit Windows 10 operating system, Intel 7 CPU, 8G RAM, and GTX3060 graphics card, and the software used for the experiments was MATLAB R2018b. Firstly, the effect of the knowledge graph on the clustering of KGK-modes was verified on the commonly used machine learning dataset. Secondly, the clustering performance was compared between the commonly used clustering algorithms and the proposed KGK-modes algorithm, respectively. Finally, the effectiveness of the proposed KGK-modes algorithm was analysed on a dataset of historical sales of 1703 e-commerce companies. The main clustering evaluation metrics [36] were purity ($\gamma$), standard mutual information (NMI), and $F$-value ($F$). Commonly used machine learning data come from the published UCI dataset and Sogo laboratory news dataset, which are shown in Tables 2 and 3, respectively.

5.1. The Influence of the Knowledge Graph on K-Modes

5.1.1. Clustering Performance on the UCI Dataset. To verify the effect of the knowledge graph on K-modes, the UCI dataset was tested using the K-modes and KGK-modes algorithms, respectively, and the results are shown in Table 4.

It can be seen that the KGK-modes algorithm shows better performance for all four different datasets. The cross-sectional comparison shows that the K-modes algorithm has the highest clustering purity of 0.8061 on the Seeds dataset.
and the lowest clustering purity of 0.7662 on the Flowers dataset. This indicates that both algorithms obtained optimal performance on the Seeds dataset and the worst clustering on the Flowers dataset. Comparing the NMI and F performance, the K-modes algorithm showed better clustering performance after the knowledge graph analysis. This is because after the knowledge graph analysis, the data samples are accurately delineated in terms of concepts, entities, and relationships. The delineation of the data samples helps to determine the sample categories to a certain extent, thus reducing the difficulty of subsequent K-modes clustering.

Table 1: Total daily data for e-commerce.

| aleg_amt | offer_amt | offer_cnt | rtn_cnt | rtn_amt | ord_cnt | bad_num | good_num |
|----------|-----------|-----------|---------|---------|---------|---------|----------|
| 5324.47  | 27.55     | 23        | 2       | 187.25  | 67      | 0       | 0        |
| 6000.27  | 132.37    | 34        | 3       | 112.37  | 84      | 0       | 1        |
| 2957.59  | 48.13     | 17        | 1       | 9.2     | 52      | 0       | 17       |
| 4247.57  | 64.74     | 26        | 1       | 23.91   | 72      | 6       | 38       |
| 2888.33  | 77.01     | 18        | 0       | 0       | 60      | 6       | 122      |
| 6156.68  | 131.39    | 27        | 0       | 0       | 84      | 9       | 214      |
| 3286.77  | 84.31     | 31        | 3       | 132.07  | 65      | 17      | 210      |
| 3356.83  | 41.99     | 20        | 5       | 238.76  | 60      | 28      | 277      |
| 4227.12  | 129.61    | 34        | 2       | 30.53   | 80      | 23      | 270      |
| 6046.86  | 1201.02   | 32        | 3       | 119.19  | 73      | 3       | 315      |
| 2911.44  | 128.57    | 19        | 1       | 10.3    | 55      | 9       | 272      |
| 4245.03  | 57.82     | 19        | 1       | 0       | 62      | 31      | 232      |
| 3944.52  | 141.86    | 22        | 3       | 127.98  | 64      | 30      | 257      |
| 3706.83  | 127.55    | 29        | 3       | 86.82   | 71      | 17      | 516      |
| 4145.74  | 135.45    | 33        | 3       | 122.13  | 68      | 12      | 305      |
| 4138.08  | 606.01    | 31        |         | 39.73   | 65      | 11      | 472      |
| 4747.22  | 99.07     | 24        | 2       | 672.12  | 8       | 5       | 547      |
| 3587.7   | 83.18     | 17        | 1       | 146.78  | 61      | 10      | 157      |
| 6683.14  | 90.29     | 20        | 5       | 127.32  | 61      | 20      | 755      |
| 4328.59  | 1357.18   | 44        | 1       | 121.03  | 89      | 12      | 462      |
| 4431.47  | 255.46    | 56        | 3       | 268.55  | 85      | 10      | 556      |
| 3850.74  | 894.16    | 32        | 4       | 192.04  | 67      | 9       | 464      |
| 3887.83  | 203.74    | 25        | 1       | 61.07   | 57      | 7       | 638      |
| 3446.55  | 184.1     | 33        | 2       | 142.37  | 61      | 11      | 668      |
| 3080.63  | 107.5     | 20        | 2       | 41.94   | 51      | 6       | 862      |
| 3990.54  | 82.08     | 19        | 1       | 73.21   | 60      | 10      | 317      |
| 2918.88  | 94.95     | 21        | 0       | 0       | 55      | 17      | 731      |
| 2679.41  | 73.51     | 16        | 0       | 0       | 47      | 13      | 583      |
| 2957.46  | 48.71     | 16        | 2       | 60.33   | 61      | 13      | 548      |
| 4354.07  | 128.83    | 35        | 2       | 63.28   | 76      | 10      | 666      |

Table 2: UCI simulation set.

| Datasets | Sample size | Number of attributes | Number of clustering categories |
|----------|-------------|----------------------|-------------------------------|
| Flowers  | 385         | 34                   | 10                            |
| Wine     | 181         | 16                   | 5                             |
| Iris     | 1024        | 21                   | 8                             |
| Seeds    | 676         | 30                   | 6                             |

Table 3: News text set.

| Datasets | Sample size | Number of document categories | Category               |
|----------|-------------|-------------------------------|------------------------|
| 1        | 100         | 5                             | Finance, military      |
| 2        | 500         | 7                             | Education, sports      |
| 3        | 800         | 7                             | Tourism, culture       |
| 4        | 1000        | 7                             | IT                     |

Table 4: Clustering performance of two algorithms (UCI set).

| Datasets | Algorithms | P     | NMI   | F     |
|----------|------------|-------|-------|-------|
| Flowers  | K-modes    | 0.7662| 0.6311| 0.7283|
|          | KGG-modes  | 0.8824| 0.7017| 0.8592|
| Wine     | K-modes    | 0.7713| 0.6621| 0.7356|
|          | KGG-modes  | 0.8632| 0.7112| 0.8572|
| Iris     | K-modes    | 0.7894| 0.7177| 0.7812|
|          | KGG-modes  | 0.8927| 0.7809| 0.8798|
| Seeds    | K-modes    | 0.8061| 0.7249| 0.7651|
|          | KGG-modes  | 0.9046| 0.7964| 0.8955|
KGK-modes showed good performance in terms of $P$, NMI, and $F$-values in clustering the four-class dataset of UCI. To analyse the stability of clustering purity, the RMSE performance of clustering purity [37] was tested. A random sample of 1000 from the UCI dataset was tested for clustering, and the results are shown in Figure 6.

The RMSE of both the K-modes and KGK-modes algorithms gradually decreased as the clustering time increased. In comparison, it was found that the RMSE of the clustering purity obtained by KGK-modes decreased more rapidly. Eventually, the RMSE of KGK-modes converges to 0.5, while that of K-modes converges to 0.75.

In addition, the clustering times of the K-modes algorithm and the KGK-modes algorithm were further compared on the UCI dataset, and the statistical results are shown in Table 5.

It can be seen that the clustering time of the K-modes algorithm and KGK-modes algorithm on the UCI dataset is directly related to the sample size. The Iris dataset with the largest sample size required the longest clustering time, while the Wine dataset with the smallest sample size required the shortest clustering time. The comparison revealed that the knowledge graph analysis consumed some time, and therefore the clustering time for KGK-modes was longer than that for K-modes, but the difference between the two was smaller.

5.1.2. Clustering Performance on News Datasets. To further validate the effect of the knowledge graph on the K-modes algorithm, the performance of the news dataset was tested using the K-modes and KGK-modes algorithms, respectively, and the results are shown in Table 6.

It can be seen that as with the UCI dataset, after the knowledge graph analysis, the K-modes algorithm showed better NMI and $F$ performance. The analysis of the stability of the clustering purity on the news dataset is given in the following, and the results are shown in Figure 7.

It can be seen that the RMSE of the K-modes and KGK-modes algorithms decreases significantly as the number of clustering iterations increases. The RMSE of the clustering purity obtained by KGK-modes decreases faster than that of the K-modes algorithm and eventually converges to about 0.4, while that of K-modes converges to about 0.5.

Next, a comparison of clustering time performance was performed. The clustering times of the K-modes and KGK-modes algorithms on the four news datasets are shown in Table 7.

For the news set with the same sample size, the KGK-modes clustering time was slightly longer than the K-modes clustering time. A comprehensive analysis of the above results shows that the KGK-modes algorithm shows higher performance on the news dataset than the UCI dataset, mainly because the news dataset has significantly fewer feature dimensions than the UCI dataset, and therefore the data clustering effect is more significant.

5.2. Sales Forecast. To verify the effectiveness of the KGK-modes algorithm in e-commerce sales prediction, the sales-related datasets (containing 8 variables) of 1703 e-commerce businesses were tested using K-means, K-medoids, CNN [38], and KGK-modes algorithms, respectively. 90% of the dataset was used as the training set, and the remaining 10% was used as the test set. Eight variables are used as inputs, and the above four prediction models are, respectively, used for fitting. The loss function was mean squared deviation, the learning rate was 0.01, the maximum number of iterations was 1000, and the resampling rate was 50%.

Taking the first type of e-commerce as an example, the KGK-modes prediction model reached optimality when the optimal number of iterations was 367, as shown in Figure 8, by continuously performing iterative calculations. This indicates that the error has been minimised when the number of iterations reaches 367. The predicted values of the model obtained by the KGK-modes algorithm are more accurate at this point. The results in Figure 9 show that in the first

**Table 5: Clustering time of two algorithms (UCI set).**

| Sample set | Algorithms  | Running time (s) |
|------------|-------------|------------------|
| Flowers    | K-modes     | 25.13            |
|            | KGK-modes   | 29.25            |
| Wine       | K-modes     | 11.29            |
|            | KGK-modes   | 14.76            |
| Iris       | K-modes     | 61.52            |
|            | KGK-modes   | 67.19            |
| Seeds      | K-modes     | 42.67            |
|            | KGK-modes   | 48.73            |

**Table 6: Clustering performance of two algorithms (news sets).**

| Datasets | Algorithms | $P$       | NMI      | $F$       |
|----------|------------|-----------|----------|-----------|
| 1        | K-modes    | 0.8646    | 0.6473   | 0.8064    |
|          | KGK-modes  | 0.9275    | 0.7014   | 0.9012    |
| 2        | K-modes    | 0.8796    | 0.6725   | 0.8355    |
|          | KGK-modes  | 0.9418    | 0.7368   | 0.9137    |
| 3        | K-modes    | 0.8821    | 0.7073   | 0.8382    |
|          | KGK-modes  | 0.9437    | 0.7716   | 0.9176    |
| 4        | K-modes    | 0.8946    | 0.7312   | 0.8468    |
|          | KGK-modes  | 0.9526    | 0.8192   | 0.9235    |
category of e-commerce, the magnitude of the impact on sales is ord_cnt, offer_cnt, bad_num, offer_amt, rtn_cnt, and rtn_amt in that order.

The fitted model was applied to the test set, and the error was compared between the true and predicted values of the test set, and the results are shown in Figure 10.

It can be seen that the KGK-modes algorithm has the highest sales prediction accuracy of about 0.94. The CNN algorithm has the second highest sales prediction accuracy after the KGK-modes algorithm at about 0.91. The K-means algorithm has the worst sales prediction accuracy at about 0.81. In terms of running time, the K-means and K-medoids algorithms are the most efficient, reaching stability at around 60 s. On the other hand, the KGK-modes and CNN algorithms took 65 s and 70 s, respectively, to reach stability, and on balance, the KGK-modes algorithm had the highest sales prediction accuracy and the K-means algorithm had the best runtime.

5.3. Adjustment Strategies for Innovative Marketing Models.

For the first e-commerce, the main influencing factors for sales are ord_cnt, offer_cnt, and bad_num. The higher the order volume, the higher the sales amount. For the second category of e-commerce, the main influencing factors of sales are ord_cnt, rtn_cnt, bad_num and good_num. For the third category of e-commerce, which has the smallest range of products, they should focus more on the marketing model innovation in positive comment. The number of bad reviews has too great an impact on third category e-commerce, and some reasonable marketing strategies should be prepared in advance to reduce bad reviews. This is because the first thing that consumers browse for when they enter a third category e-commerce is the reviews of the goods. It is
recommended to offer some economical and affordable goods from the customer’s point of view and to improve the logistics management.

After a comprehensive analysis of the characteristics of these three types of e-commerce, it is found that ord_cnt is the most important influencing factor on e-commerce sales. Therefore, we can directly use ord_cnt to simply forecast the change of sales. The role of marketing activities such as appropriate price reductions is more obvious for the second category of e-commerce. For the third category of e-commerce, the impact of bad reviews is much greater than that of good reviews. This is because in the process of online shopping, in order to buy what they want, people will refer to other people’s reviews of this product. If this shop has more bad reviews, it will affect the consumer’s purchase intention. For e-commerce companies of different scales, the most important sales influencing factors are different. In the future, according to the priority of the factors affecting sales, different sizes of e-commerce companies need to adjust the innovative marketing model to ensure maximum profit.

6. Conclusion

In this paper, a big data-driven KGK-modes clustering algorithm is proposed and applied to e-commerce sales volume prediction, so as to reasonably adjust the required innovative marketing mode. Knowledge graph was used to preprocess the data samples to generate a sample set containing concepts, entities, and relationships so that the key attributes of the historical sales data could be extracted. After the knowledge graph analysis, K-modes algorithm was able to achieve better clustering performance. The KGK-modes clustering model was used to achieve e-commerce sales prediction. The experimental results show that the KGK-modes clustering model has high sales prediction accuracy. The proposed method has certain reference value for e-commerce enterprises of different scales to formulate innovative marketing models. Subsequent research will be carried out in terms of the efficiency of clustering, and invoking the Spark parallel platform should be considered to reduce the running time of KGK-modes clustering. On the other hand, an attempt is made to optimize the objective function of KGK-modes clustering in order to improve the efficiency of the clustering algorithm.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Humanities and Social Sciences Projects of Universities in Jiangxi Province (Research on the Compounding Mechanism of Changes in College Students’ Political Trust (issue no. SZXX21019); Practical Reflection and Strategy Optimization Research on Rural Cultural Construction from the Perspective of the Peasant (issue no. SH19104)).

References

[1] L. Y. Chen, “A knowledge marketing model: determinants of organisational resource-based capabilities on e-retail performance,” *International Journal of Business Information Systems*, vol. 9, no. 1, pp. 89–107, 2012.
[2] L. V. Lorente-Bayona, É. Gras-Gil, and M. Moreno-Enguix, “Internet penetration and international travel and tourism expenditure: the role of foreign exchange control,” *Tourism Economics*, vol. 2021, no. 2, p. 135481662110278, 2021.
[3] D. Shah and B. P. S. Murthy, “Marketing in a data-driven digital world: i,” *Journal of Business Research*, vol. 125, pp. 772–779, 2021.
[4] H. Q. Vu, G. Li, and R. Law, “Discovering implicit activity preferences in travel itineraries by topic modeling,” *Tourism Management*, vol. 75, no. 12, pp. 435–446, 2019.
[5] L. Liu, F. Zhang, H. Li, and C. Wen, “Research on the community group buying marketing model of fresh agricultural products in Jilin province from the perspective of internet marketing and retail,” *JOP Conference Series: Earth and Environmental Science*, vol. 769, no. 2, p. 022054, 2021.
[6] D. Klimanov and O. Tretjakov, “Linking business model research and marketing: new network-based approach to business model analysis,” *The Journal of Business and Industrial Marketing*, vol. 34, no. 1, pp. 117–136, 2019.
[7] D. Xu, T. Chen, J. Pearce, Z. Mohammadi, and P. L. Pearce, “Reaching audiences through travel vlogs: the perspective of involvement,” *Tourism Management*, vol. 86, no. 2, p. 104326, 2021.
[8] H. Gebauer, A. Arzt, M. Kohtamäki et al., “How to convert digital offerings into revenue enhancement – c,” *Industrial Marketing Management*, vol. 91, pp. 429–441, 2020.
[9] R. Dew and A. Ansari, “Bayesian nonparametric customer base analysis with model-based visualizations,” *Marketing Science*, vol. 37, no. 2, pp. 216–235, 2018.
[10] E. Kappe, A. Stadler Blank, and W. S. Desarbo, “A random coefficients mixture hidden Markov model for marketing research,” *International Journal of Research in Marketing*, vol. 35, no. 3, pp. 415–431, 2018.
[11] J. Martins, R. Gonalves, and F. Franco, “A multisensory virtual experience model for thematic tourism: a Port wine tourism application proposal,” *Journal of Destination Marketing & Management*, vol. 6, no. 2, pp. 91–106, 2017.
[12] J. J. Yang, H. W. Lo, C. S. Chao, C. C. Shen, and C. C. Yang, “Establishing a sustainable sports tourism evaluation framework with a hybrid multi-criteria decision-making model to explore potential sports tourism attractions in taiwan,” *Sustainability*, vol. 12, no. 4, p. 1673, 2020.
[13] S. Osama, N. Mohsen, H. Walayat, and M. Luis, “Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method,” *Computers & Industrial Engineering*, vol. 132, pp. 47–58, 2019.
[14] H. Y. Lee and N. J. Wang, “Cloud-based enterprise resource planning with elastic model–view–controller architecture for Internet realization,” *Computer Standards & Interfaces*, vol. 64, no. 5, pp. 11–23, 2019.
[15] H. Xie, R. T. B. Ma, and J. C. S. Lui, “Enhancing reputation via price discounts in E-commerce systems: a data-driven
approach,” ACM Transactions on Knowledge Discovery from Data, vol. 12, no. 3, pp. 1–29, 2018.

[16] S. Mu, Y. Wang, F. Wang, and L. Ogiela, “Transformative computing for products sales forecast based on SCIM,” Applied Soft Computing, vol. 109, no. 1, p. 107520, 2021.

[17] J. Zhao and C. Zhang, “Research on sales forecast based on prophet-SARIMA combination model,” Journal of Physics: Conference Series, vol. 1616, no. 1, p. 012069, 2020.

[18] A. Kumar, A. Adlakha, and K. Mukherjee, “Modeling of product sales promotion and price discounting strategy using fuzzy logic in a retail organization,” Industrial Management & Data Systems, vol. 116, no. 8, pp. 1418–1444, 2016.

[19] O. Erenko and O. Koppel, “Formal and informal macro-regional transport clusters as a primary step in the design and implementation of cluster-based strategies,” Transport and Telecommunication Journal, vol. 16, no. 3, pp. 207–216, 2015.

[20] M. Kingsland, L. Wolfenden, J. Tindall et al., “Tackling risky alcohol consumption in sport: a cluster randomised controlled trial of an alcohol management intervention with community football clubs,” Journal of Epidemiology & Community Health, vol. 69, no. 10, pp. 993–999, 2015.

[21] G. T. Elizabeth, W. Ian, and V. Bas, “Householders’ mental models of domestic energy consumption: using a sort-and-cluster method to identify shared concepts of appliance similarity,” PLoS One, vol. 11, no. 7, p. e0158949, 2016.

[22] Y. Fan, Y. Liu, H. Qi, F. Liu, and X. Ji, “Anti-interference technology of surface acoustic wave sensor based on K-means clustering algorithm,” IEEE Sensors Journal, vol. 21, no. 7, pp. 8998–9007, 2021.

[23] Z. Huang and M. K. Ng, “A fuzzy k-modes algorithm for clustering categorical data,” IEEE Transactions on Fuzzy Systems, vol. 7, no. 4, pp. 446–452, 1999.

[24] A. Saha and S. Das, “Categorical fuzzy k-modes clustering with automated feature weight learning,” Neurocomputing, vol. 166, no. 10, pp. 422–435, 2015.

[25] Y. Jia, Y. Qi, H. Shang, R. Jiang, and A. Li, “A practical approach to constructing a knowledge graph for cybersecurity,” Engineering, vol. 4, no. 1, pp. 53–60, 2018.

[26] L. Shen, R. He, and S. Huang, “Entity alignment with adaptive margin learning knowledge graph embedding,” Data & Knowledge Engineering, vol. 139, pp. 101987–101996, 2022.

[27] W. Zheng, L. Yan, C. Gou, Z. C. Zhang, and F. Y. Wang, “Pay attention to doctor-patient dialogues: multi-modal Knowledge Graph attention image-text embedding for COVID-19 diagnosis,” Information Fusion, vol. 75, no. 8, pp. 108–116, 2021.

[28] H. Wang, Y. Wang, J. Li, and T. Luo, “Degree aware based adversarial graph convolutional networks for entity alignment in heterogeneous knowledge graph,” Neurocomputing, vol. 487, pp. 99–109, 2022.

[29] C. Liu, X. Wang, Y. Huang et al., “A moving shape-based robust fuzzy K-modes clustering algorithm for electricity profiles,” Electric Power Systems Research, vol. 187, p. 106425, 2020.

[30] H. H. Nguyen, “Privacy-preserving mechanisms for k-modes clustering,” Computers & Security, vol. 78, pp. 60–75, 2018.

[31] C. Liu, X. Wang, Y. Huang et al., “A moving shape-based robust fuzzy K-modes clustering algorithm for electricity profiles,” Electric Power Systems Research, vol. 187, p. 106425, 2020.

[32] M. Faezirad, A. Pooya, Z. Naji-Azimi, and M. Amir Haeri, “Preventing food waste in subsidy-based university dining systems: an artificial neural network-aided model under uncertainty,” Waste Management & Research: The Journal for a Sustainable Circular Economy, vol. 39, no. 8, pp. 1027–1038, 2021.

[33] R. J. Kuo, Y. R. Zheng, and T. P. Q. Nguyen, “Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering,” Information Sciences, vol. 557, pp. 1–15, 2021.

[34] R. J. Kuo and T. P. Q. Nguyen, “Genetic intuitionistic weighted fuzzy k-modes algorithm for categorical data,” Neurocomputing, vol. 330, pp. 116–126, 2019.

[35] A. Ichenihi, W. Li, Y. Gao, and Y. Rao, “Feature selection and clustering of damage for pseudo ductile unidirectional carbon/glass hybrid composite using acoustic emission,” Applied Acoustics, vol. 182, no. 2, p. 108184, 2021.

[36] J. C. Forte, G. Yeshtamagbetova, M. Grinten, and B. Hiemstra, “Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering,” Scientific Reports, vol. 11, no. 1, p. 12109, 2021.

[37] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, “A novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots,” Applied Sciences, vol. 11, no. 23, p. 11202, 2021.

[38] F. Sun, G. Xie, and S. Li, “An artificial-neural-network based prediction of heat transfer behaviors for in-tube supercritical CO2 flow,” Applied Soft Computing, vol. 102, p. 107110, 2021.