symmetry

MDPI

Article

Research and Application of Improved Clustering Algorithm in Retail Customer Classification

Chu Fang [1] and Haiming Liu [2,*]

[1] College of Economics and Management, Zhaoqing University, Zhaoqing 526061, China; fangchu@zqu.edu.cn
[2] Business Administration College, Nanchang Institute of Technology, Nanchang 330099, China
* Correspondence: 2017010042@zqu.edu.cn

Abstract: Clustering is a major field in data mining, which is also an important method of data partition or grouping. Clustering has now been applied in various ways to commerce, market analysis, biology, web classification, and so on. Clustering algorithms include the partitioning method, hierarchical clustering as well as density-based, grid-based, model-based, and fuzzy clustering. The K-means algorithm is one of the essential clustering algorithms. It is a kind of clustering algorithm based on the partitioning method. This study's aim was to improve the algorithm based on research, while with regard to its application, the aim was to use the algorithm for customer segmentation. Customer segmentation is an essential element in the enterprise's utilization of CRM. The first part of the paper presents an elaboration of the object of study, its background as well as the goal this article would like to achieve; it also discusses the research the mentality and the overall content. The second part mainly introduces the basic knowledge on clustering and methods for clustering analysis based on the assessment of different algorithms, while identifying its advantages and disadvantages through the comparison of those algorithms. The third part introduces the application of the algorithm, as the study applies clustering technology to customer segmentation. First, the customer value system is built through AHP; customer value is then quantified, and customers are divided into different classifications using clustering technology. The efficient CRM can thus be used according to the different customer classifications. Currently, there are some systems used to evaluate customer value, but none of them can be put into practice efficiently. In order to solve this problem, the concept of continuous symmetry is introduced. It is very important to detect the continuous symmetry of a given problem. It allows for the detection of an observable state whose components are nonlinear functions of the original unobservable state. Thus, we built an evaluating system for customer value, which is in line with the development of the enterprise, using the method of data mining, based on the practical situation of the enterprise and through a series of practical evaluating indexes for customer value. The evaluating system can be used to quantify customer value, to segment the customers, and to build a decision-supporting system for customer value management. The fourth part presents the cure, mainly an analysis of the typical k-means algorithm; this paper proposes two algorithms to improve the k-means algorithm. Improved algorithm A can get the K automatically and can ensure the achievement of the global optimum value to some degree. Improved Algorithm B, which combines the sample technology and the arrangement agglomeration algorithm, is much more efficient than the k-means algorithm. In conclusion, the main findings of the study and further research directions are presented.

Keywords: clustering; k-means algorithm; customer segmentation; CRM

Citation: Fang, C.; Liu, H. Research and Application of Improved Clustering Algorithm in Retail Customer Classification. *Symmetry* **2021**, *13*, 1789. https://doi.org/10.3390/sym13101789

Academic Editors: Kuo-Hui Yeh, Chien-Ming Chen, Wei-Che Chien and Basil Papadopoulos

Received: 27 July 2021
Accepted: 17 September 2021
Published: 26 September 2021

1. Introduction

In the face of rapid change in information technology, people's ability to use information technology to produce and collect data has greatly improved. A large number of databases are used for business management, government office, scientific research, and engineering development. In order to make data truly become a company's resource, we

must make full use of it in order to serve the company's own business decision-making and strategic development [1]. Otherwise, a large amount of data may become a burden. Data mining and knowledge discovery came into being and flourished. The process entails the extraction of hidden, unknown, but potentially useful information and knowledge from the data. People regard raw data as a source of knowledge, just like mining from ore. The original data can be structured such as data from a relational database, or semi-structured such as text, graphics, image data, and even heterogeneous data distributed on the network. The method of discovering knowledge can be mathematical or non-mathematical [2]; it can be deductive or inductive. The mined knowledge can be used for information management, query optimization, decision support, process control [3], etc.; it can also be used for the maintenance of data itself. Therefore, data mining is an interdisciplinary subject, involving artificial intelligence technology, statistical technology, database technology, among others. It brings together researchers from different fields, especially scholars and engineers in database, artificial intelligence, mathematical statistics, visualization, parallel computing, etc. Cluster analysis divides a large number of data into subclasses with the same properties, which is convenient for the understanding of the distribution of data [4]. Therefore, it is widely used in many fields such as pattern recognition, image processing, data compression, and so on.

In market analysis, cluster analysis can help decision-makers identify customer groups with different characteristics and the behavioral characteristics of each customer group [5]. In bioengineering research, cluster analysis can be used to deduce the classification of animals and plants according to their functions; it can also be used in gene division and to obtain the inherent structural characteristics of the population [6]. In the field of non-relational database (such as spatial database), cluster analysis can identify the same geographical characteristics of an area as well as the environmental and human characteristics of the area. In the field of web information retrieval, clustering analysis can classify web documents and improve retrieval efficiency.

This paper applies cluster analysis technology to enterprise customer segmentation. The competition of enterprises is largely related to customer segmentation [7]. The competition for customer source, through the multi-angle quantitative analysis of customer value, can help enterprises effectively focus on the most valuable customers and customers with the most development potential, give priority to the allocation of resources, and thus cooperate with them [8].

2. Literature Review

Cluster analysis, as a branch of statistics, has been widely studied for many years [9], k-medoids (k-qb center points), and some other methods. Analysis tools have been added to many statistical analysis software packages or systems. In the field of machine learning [10], clustering is an example of unsupervised learning, which is different from classification. Guided learning does not rely on pre-defined classes and training instances with class labels [11]. For this reason, clustering is observational learning rather than example learning [12]. In concept clustering, a group of objects form a cluster only when they can be described by a concept, which is different from the traditional clustering based on geometric distance to measure similarity [13]. The research work on cluster analysis focuses on finding suitable solutions for an effective and practical cluster analysis of large databases. The current research directions include the following aspects [14].

Scalability of the algorithm: Many clustering algorithms involve the use of small data with less than 200 data objects. Ability to handle different types of attributes [15]. Many algorithms are designed to cluster the numbers of numeric types. However, in practical application, they may be required to cluster other types of data [16]. Such as classification/nominal type (categorical/nominal), ordinal, binary data, or the idean distance and the tendency to find clusters [17]. Spherical clusters can have similar density and size. However, a cluster may be of arbitrary shape, so it is very important to propose an algorithm that finds arbitrarily shaped clusters [18]. Finding clusters with arbitrary

shapes: Many clustering algorithms are based on the Euclidean method for determining input parameters; in cluster analysis, many clustering algorithms need to ask the user to enter certain parameters [19], such as the number of clusters you want. Clustering results are very sensitive to input parameters, which are usually difficult to determine, especially for data sets containing high-dimensional objects [20]. Requiring entry parameters not only increases the burden of users, but also makes the clustering quality difficult to control. Insensitive to the order of input records: Some clustering algorithms are sensitive to the order of input data. If the same data set is submitted to the same algorithm in a different order [21], it may produce very different clustering results. It is of great significance to do research and develop algorithms that are not sensitive to the order of data input [22]. High dimensionality: A database may contain several dimensions or attributes. Many clustering algorithms are good at processing. Low-dimensional data generally involve only two to three dimensions. Generally, the quality of clustering can be well-judged in three-dimensional cases at most. Clustering data objects in high-dimensional space is very challenging, especially considering that such data may be highly skewed and sparse [23]. Ability to process noise data: In practical applications, most data contain outliers, including missing, unknown, or incorrect data. Some clustering algorithms are sensitive to such data, which will lead to low-quality clustering results. Constraint based clustering: In practical application [24], it may be necessary to cluster under various constraint classes. It is a challenging task to find data groups that not only meet specific constraints, but also have good clustering characteristics. Interpretability and usability: Users usually want the clustering results to be interpretable [25] and understandable. Therefore, how the application target affects the selection of the clustering method is also an important research topic [26]. At the same time, there are the following main problems in clustering methods, which also need to be further studied and solved. Initial sensitivity [27]. The selection of the initial value and input order have a great impact on the final result of the clustering algorithm for big data. Measures that can be taken in the field of data mining: Multiple groups of different initial values can be used and iterated many times, until the best one is finally selected as the calculation result [28]. However, this cannot guarantee that the global optimal solution will be reached. Optimal solution: In essence, the clustering process is an optimization process, which creates the system through an iterative operation. The objective function of the system is what provides an optimal solution. However, this objective function is a nonconvex function in the state space. It has many minima, of which only one is the global minimum, while the others are the local minimum [29]. The goal of optimization is to achieve global optimization, therefore the optimization problem of a nonconvex function is a research topic to be solved. Efficiency of algorithm: Improving the efficiency of the algorithm is another important research topic in the field of clustering. Improving the existing clustering algorithm will enable the algorithm to perform incremental clustering and good scalability [30]. When dealing with large databases, the database is scanned only once to improve the efficiency of the algorithm. Research on wavelet transform clustering algorithm: At present, most of the research performed on clustering are on k-means. The generalization and improvement of the performance of the fuzzy c-means algorithm and the SOFM algorithm has been improved to varying degrees [31]. However, there is not a lot of literature on the wavelet clustering algorithm and its improved methods. Because it meets the many requirements of a good clustering algorithm, further research and development of the wavelet clustering algorithm will achieve unexpected results [32]. Mining based on different media: At present, most clustering algorithms in data mining are based on relational databases. The algorithm of the transaction database is designed and applied to other types of databases (such as the object-oriented database and the attribute-oriented database) [33]. The clustering algorithm for the mining of databases such as temporal database, text database, heterogeneous database, web database, multidimensional database, geographic database, data warehouse, etc., will also be very meaningful work.

We selected the attributes of customer data of a city branch of the Bank of China for a sample analysis, established a data mining model using the k-means method, and performed two-dimensional cross clustering on the current value and potential value of customers in the customer dimension [34]. We generated six customer groups (k = 6), and generated the initial customer group after one clustering. Then, the customer groups with small differences were clustered twice, and 11 customer types were obtained according to the customer value [35].

Starting with the existing results in the scientific literature, Table 1 shows the determinants that may significantly affect the factors of influence in the application of the data mining algorithm in customer segmentation.

Table 1. Factors of influence in the application of data mining.

Factor of Influence	Contents	Literature Review
Data acquisition	Bayesian classification is statistical	(Haiying 2010) [32]
Network classification	The task of data mining	(Sheshasaayee 2017) [33]
Clustering ways	Apply clustering to customer segmentation	(Srivastava 2016) [34]
Purchase frequency	Artificial neural network is an analysis	(Zhang 2002) [35]
Modified k-means	HK clustering network	(Zahrotun 2017) [36]
K-means algorithm	K-means clustering algorithm is a method based on centroid	(Tong 2017) [37]

The HK clustering algorithm based on the Hopfield network was developed and proved to be better than the MLPs and Kohonen networks. An outline of this algorithm is shown below [36].

Stage 1: Network Initial Design

A K × N matrix is established, where K is the number of preset clusters (obtained by using prior knowledge) and N is the number of customers in the data set.

It is randomly assigned to each element in the matrix, where VPI represents the strength of the group members of customer I in group P.

A small subset of customers is randomly selected to determine the initial group center of gravity.

Stage 2: Network Execution

Each n = k × n node in the network can accept the information from each node as the input value, where API represents the value input to node PI, which is a function of VPI, and vqi represents the strength of customer I assigned to groups outside group P. RPI is the square of the Euclidean distance between customer I and the center of gravity of group P [37].

In the T stage, after the first node in the network produces the result of VPI, it uses VPI to carry out a two-stage program.

The center of gravity of group P is recalculated by using the updated average weight of VPI, and RPI is recalculated.

In the T + 1 phase, the second node in the network updates VPI and recalculates the center of gravity and RPI of group P, as described in steps (5) and (6).

At t + 2, t + 3, t + 4 in the stage, all the remaining nodes in the network repeatedly update VPI and recalculate RPI, as mentioned in steps (5) and (6).

The network is continuously calculated until the output of each node remains unchanged. In this case, with VPI (T stage) = VPI (T + T stage), the optimal solution can be obtained.

Among them, a, B, and C denote coefficients greater than 0, while the first mathematical item restricts the same customer to be divided into two groups, the second item restricts each customer to belong to a certain group, and the third item restricts RPI (Euclidean distance between customer I and the center of gravity of group P) to be the minimum.

Therefore, when the minimum solution of this formula is obtained, it is the best solution for the achievement of the clustering result [38].

The main research work of this paper improves the shortcomings of the k-means algorithm in theoretical research. Some deficiencies include the fact that the k-means clustering problem is an NP-hard non-exponential problem, that is, the complexity of the problem varies with time, and the number of bits increases exponentially, which is closely related to many other clustering and location problems. However, the algorithm has the following disadvantages [39]:

(1) The algorithm requires the user to specify the value of parameter K, and different K values affect the efficiency and result of cluster analysis;

The results have a great impact:

(2) The selection of the initial center point of the algorithm is closely related to the operation efficiency of the algorithm. It may lead to a large number of iterations or limited to a local optimal state;
(3) It is only good at dealing with spherical data;
(4) It is sensitive to abnormal deviation data;
(5) The algorithm is only suitable for the data of numerical attributes, and the data processing effect of classification attributes is poor;
(6) The efficiency of the search strategy is low, and the overhead of the algorithm is large when dealing with large databases. In this paper, the defects of point (1), (2), and (6) for the typical k-means algorithm are analyzed because the clustering results of the traditional k-means algorithm are affected by the selection of the initial clustering center. This algorithm is improved on the basis of the traditional k-means algorithm. The improved algorithm A effectively solves the problem of the algorithm. Depending on the initial value of K, the number of classes K can be automatically generated; at the same time, the algorithm compares the selection of the initial center point. Strictly, the distance between each central point is far, which prevents the initial clustering center to select a class and have a certain process. It overcomes the limitation of the algorithm to the local optimal state. Algorithm B is improved by point (6), combined with sampling technology and a hierarchical aggregation algorithm that improve the original algorithm; thus, the new algorithm B is more effective. In the application of the algorithm, clustering technology is used in customer segmentation, and customer segmentation is established by the analytic hierarchy process value system and by quantifying customer value. Based on this, cluster technology is applied to divide customers into different categories. Therefore, it has a certain practical significance in that it effectively carries out customer management. At present, there have been some customer value reviews of the price system, but the measurement model is not mature enough [40]. The first measurement index is the direct profit contribution of customers to the enterprise. It is also difficult to quantify. This paper will use the method of data mining to start from the actual situation of the enterprise. Through a series of operable customer value evaluation indexes, a customer value evaluation model suitable for enterprise development is established. Based on this, we can measure customer value, segment customers, and establish a decision support system for customer value management. As for the part on theoretical research, this paper focuses on the ideas of discovering problems, raising problems, analyzing problems, and solving problems based on clues, adopts the research method of combining empirical analysis and theoretical analysis, and combines theoretical research with applied research [41]. Organically combined, through the enlightenment of the improved k-means algorithm, the aim is to resolve the existing problems of the algorithm. The deficiency is studied and a new algorithm is obtained. Based on the previous research results, this paper follows the basic requirements of clustering and gradually completes the research work.as can be seen in Table 2.

Table 2. Methods of cluster analysis by adopting the k-means algorithm.

Hierachical Clustering Methods	K-Means Clustering	Fuzzy C-Means
(Cho 2013) [40] Hierarchical clustering is a method of classifying classes from variable to less. The steps of classification are as follows: Each sample falls into one category, and there are different categories at this time. Calculate the distance between each sample and classify the nearest two samples into one category; Calculate the distance between the new class and other classes, and then merge the two nearest classes. It is still greater than 1, continue to repeat the above steps until all samples are classified into one category, then stop. The distance between samples has different definition methods, and there are also different definitions of distance between class methods, which produces different hierarchical clustering methods.	(Jiang 2009) [41] K-means algorithm adopts iterative updating method: K clustering centers are used in each iteration. Form the surrounding points into K clusters, and recalculate the centroid of each cluster (i.e., the plane of all points in the cluster). The average (i.e., geometric center) will be used as the reference point for the next iteration. Iteratively generates the selected reference point. The closer it gets to the real clustering centroid, the objective function becomes increasingly smaller, and the clustering effect progressively better.	(Lu 2014) [42] Fuzzy c-means is a clustering method that allows specific data to appear in multiple clusters. It does not determine the member history of the data point in a given cluster. Instead, a specific data point calculates data belonging to the cluster. The advantage of fuzzy c-means over k-means is that the results obtained for large, similar data sets are better than those of the k-means algorithm, where data points must completely exist in a cluster.

3. Materials and Methods

3.1. Aim and Hypotheses

Establishment of market and customer segmentation indexes and data processing.

Combined with the principle of selecting the market and customer segmentation indicators, and according to the characteristics of the retail industry and market customers, this paper selected eight customer segmentation indicators: gender, age, education, occupation, income, purchase frequency, purchase amount, and shopping satisfaction.

The initial hypothesis is that the variables mentioned above have an important impact on the research and application of improved the clustering algorithm with regard to retail customer classification.

Considering the specialized studies summarized in Table 1, previous studies considered only some of these factors or small groups of influencing factors. Our research is important and relevant because it performed an exhaustive analysis of the variables that have a significant influence on the retail industry. Our study also classified these indicators for the first time in scientific literature. The result of the research consists in selecting and highlighting the indicators that have a specific level of significance. The selected indicators, starting from the previous partial results from scientific literature, were analyzed as individual hypotheses (H1–H7), as can be seen in Table 3.

Table 3. Research hypotheses-indicators.

Hypothesis Number	Alternative Hypothesis	Previous Research
H1	Association rule analysis	(Haiying 2010) [32]
H2	Cluster analysis	(Sheshasaayee 2017) [33]
H3	Classification square analysis	(Srivastava 2016) [34]
H4	Customer value evaluation method	(Zhang 2002) [35]
H5	Customer classification model	(Zahrotun 2017) [36]
H6	Implementation of cluster analysis algorithm	(Tong 2017) [37]
H7	Model application example analysis	(Shah 2012) [38]

In our research, we also aimed to analyze whether various demographic aspects could generate a significant impact on the improved clustering algorithm for retail customer classification. Based on this additional objective, we proposed the following additional hypotheses detailed in Table 4.

Table 4. Additional research hypotheses.

H	The Hypothesis
Ha1	The phenomenon of improved clustering algorithm in retail customer classification depends significantly on age.
Ha2	The phenomenon of improved clustering algorithm in retail customer classification depends significantly on gender.
Ha3	The phenomenon of improved clustering algorithm in retail customer classification depends significantly on marital status.

3.2. Variables and Instruments

In order to confirm or refute the research hypotheses proposed in Tables 2 and 3, the study included several variables. Therefore, the independent grouping variables were age, gender, and marital status. The independent variables were interval type (value scale), including education, occupation, income, purchase frequency, purchase amount, and shopping satisfaction.

In a questionnaire containing 11 questions, the phenomenon of the improved clustering algorithm in retail customer classification was evaluated: four questions considered the criteria of the filtered respondents and their demographic characteristics, while seven questions referred to the analyzed independent variables.

In research, the improved clustering algorithm analyzed the phenomenon of each index in retail customer classification. We used the Likert scale to score the respondents' decision to overwork from 1 (the influence of this factor is very small) to 5 (the influence of this factor is very strong).

3.3. Sample

Our study analyzed the responses of 178 respondents. The overall sample of the improved clustering algorithm in retail customer classification was calibrated by using [42]. In order to achieve a 95% confidence level, the sample size had to be at least 139 for a response distribution equal to 90%. Among all the respondents surveyed, 37.64% were over 25 years old (millennials), 62.36% were under 25 years old (generation z), 64.41% were male, and 35.59% were female. In terms of marital status, 22.59% were married and 77.41% were single.

3.4. Statistical Analysis of Data and Procedure

The survey was conducted in China from January to March 2020. The survey was distributed electronically to 178 employees in China's banking industry, which is one of the representative centers in southeast of Europe [43]. The participation of respondents was voluntary. The respondents did not involve any financial or material rewards. Respondents were aged between 18 and 45; in the literature [36], two representative and recognized generations—millennials and generation Z—achieved age separation. Descriptive statistics (mean, bias) were used for all data collected from respondents. The tool we used to calculate statistical indicators and coefficients was IBM SPSS statistics v. Based on a detailed literature review, we set the research hypotheses. The reliability of the questionnaire was analyzed by Cronbach alpha test. At different stages, the research hypotheses were verified by using several statistical analyses: Pearson correlation (analyzing the correlation between independent variables and dependent variables), multiple analysis of variance (determining the demographic group and its impact on workaholic phenomenon), and multiple linear regression analysis model (determining the correlation coefficient of multiple linear regression equation).

The reliability of the questionnaire was verified by calculating the Cronbach alpha coefficient. Therefore, for n = 7, the $\alpha = 0.715$, which verified the internal consistency of the survey used to obtain respondents' answers according to [39].

4. Results

Descriptive statistics are shown in Table 5. We considered all seven independent factors selected from the scientific literature and dependent variables. As can be seen in Table 4, respondents believed that the level of the improved clustering algorithm in retail customer classification was mainly affected by the following variables: days since the last purchase (M = 4.64), days since the first purchase (M = 4.42), and intrinsic fun of work (M = 4.21). According to the average value of the survey used (M = 3.91), and the total number of orders to be proved (M = 4.15). These numerical results partially confirm some previous studies on the improved clustering algorithm for retail customer classification. To test the validity of the H1–H7 hypotheses, we used an analysis based on multiple linear regression. In this econometric model, the dependent variable was account age, and the independent variables were education, occupation, income, purchase frequency, purchase volume, and shopping satisfaction. All variables were included in the multiple linear regression model by enter method; we also used linear correlation. After calculation using IBM SPSS statistics, the data in Table 5 were obtained. The beta column contains coefficients with normalized values. In order to illustrate and verify the econometric model in this paper, the significance threshold of the relevant coefficient value was considered to be less than 5%.

Table 5. Independent and dependent variables in the banking industry—descriptive statistics.

Segment	(Haiying 2010) [32]	178	2.00	5.00	3.9101	0.67726
Education	(Sheshasaayee 2017) [33]	178	2.00	5.00	4.2191	0.53786
Account age (months)	(Srivastava 2016) [34]	178	1.00	5.00	2.4326	1.27490
Occupation	(Zhang 2002) [35]	178	2.00	5.00	4.4213	0.59136
Income	(Zahrotun 2017) [36]	178	1.00	5.00	4.1503	0.65167
Purchase frequency	(Tong 2017) [37]	178	1.00	5.00	4.1742	0.79405
Purchase amount	(Shah 2012) [38]	178	2.00	5.00	4.6404	0.62431
Shopping satisfaction	(Liu 2014) [39]	178	3.00	5.00	4.3090	0.61067

Table 6 was obtained. The beta column contains coefficients with normalized values. In order to illustrate and verify the econometric model in this paper, the significance threshold of the relevant coefficient value was considered to be less than 5%.

Table 6. The independent variables' impact on the description of the banking industry.

Indicators	**Beta**	**T**	**Sig.**
(Constant)	0.772 *	2.097	0.037
Occupation	0.334 **	5.323	0.000
Income	0.215 *	2.522	0.013
Purchase frequency	0.032	1.071	0.286
Purchase amount	0.201 **	2.854	0.005
Shopping satisfaction	−0.070	−1.060	0.291

Note: * Significant for the 5% level, ** Significant for the 1% level.

It should be noted that according to the calculation results, the value of the R square indicator is 0.444, f = 19.360, $p < 0.01$. The values in Table 5 lead us to conclude that workaholics are statistically significantly affected by some analysis indicators. The independent variable multicollinearity test value (VIF = variance expansion coefficient) also supports this confirmation, and the analysis results are shown in Table 7.

Considering that the VIF values of all independent variables are less than 2.00, we can safely conclude that the variables are not collinear. The results of multicollinearity analysis are strong evidence that support the effectiveness of the model.

According to the index contribution in Table 5, we note that retail customers in the banking industry are classified by occupation (β = 0.334), income (β = 0.215), purchase amount (β = 0.201), and shopping satisfaction (β = 0.150).

Table 7. Multicollinearity test (VIF) statistics.

Indicators	Tolerance	VIF
Education	0.676	1.480
Occupation	0.581	1.721
Income	0.843	1.186
Purchase frequency	0.705	1.418
Purchase amount	0.656	1.523
Shopping satisfaction	0.744	1.344
Account age (months)	0.607	1.647

Based on the research conducted and the results obtained, we can conclude that the research hypotheses H1, H2, H4, and H6 have been confirmed. Therefore, we can be sure that the variables related to income, purchase frequency, purchase amount, and shopping satisfaction as well as the desire for grade promotion have been confirmed. These assumptions do not establish the support indicators of life partners for the classification of retail customers, nor the ability to prove gender, age, and education level. To test the hypotheses on demographic variables (Ha1–Ha3), we performed a multivariate analysis of variance. The purpose was to emphasize whether the classification composition, income, purchase frequency, purchase amount, and shopping satisfaction level of retail customers depended on gender, age, education, and occupation.

According to the results presented in Table 8, there is no significant difference between the HK clustering method and the k-means method when the input value types are different, but the results of rational seeds are better for the mixed mode. When using random seeds, the HK clustering method is better than the k-means method and mixed mode, while when using rational seeds, HK clustering method is the same as the k-means method and better than the mixed mode. Therefore, on the whole, the HK means method is better than the k-means method and the mixed model no matter what kind of data is used.

There are several reasons for the difference. (1). The data in the real world are too large, complex, and disturbed, and lack sufficient definition. (2). In the process of calculation, the HK clustering method and the mixed mode use the partial assignment of group members as the operation criterion, while the k-means method uses all assignments. (3). The hybrid mode is based on the error when adjusting the group members, while the HK cluster rule is to adjust the group members only after each node is updated.

Table 9 can be divided into two parts. Firstly, it can be seen clearly in part (a) that when the group density is the same, the result of HK cluster method is better, but when there is a certain group density of 60%, the result of K-means method is better. When there is no outlier in the group, the results of the three clustering methods are good, but when the outlier is 20%, the values of the three methods decrease, especially in the mixed mode. In addition, if there is only one disturbance variable, the solution of HK cluster method is similar to that of K-means method and is better than that of mixed mode. After adding another disturbance variable, the difference between the three methods will be reduced.

The result of the HK clustering method is better, but when there is a certain group density of 60%, the result of the k-means method is better. When there is no outlier in the group, the results of the three clustering methods are good, but when the outlier is 20%, the values of the three methods decrease, especially in the mixed mode. In addition, if there is only one disturbance variable, the solution of the HK clustering method is similar to that of the k-means method and is better than that of the mixed mode. After adding another disturbance variable, the difference between the three methods is reduced.

Table 8. The following two tables are the results of 4317 real data operations.

(a) HK Random Seeds (Minimum Variation Solution)

Segment	Number of Credit Cards Used	Account Age (Months)	Days since First Purchase	Days since Last Purchase	Total Number of Orders	Total Dollars Spent	Segment Size
One	1.6	39.8	797.2	319.2	4.0	425.67	603
Two	1.2	54.8	824.0	659.4	1.6	137.02	648
Three	0.9	31.7	541.0	452.5	1.4	113.53	1173
Four	1.5	48.0	454.4	335.2	1.6	144.16	645
Five	1.0	33.8	858.3	795.8	1.4	139.85	1248

(b) K-Means Rational Seeds

Segment	Number of Credit Cards Used	Account Age (months)	Days since First Purchase	Days since Last Purchase	Total Number of Orders	Total Dollars Spent	Segment Size
One	1.1	56.8	712.1	591.9	1.5	133.00	684
Two	1.0	35.0	480.3	417.8	1.4	115.83	1296
Three	0.9	34.7	839.2	794.7	1.4	138.48	1383
Four	2.0	38.2	787.2	333.8	2.4	207.74	605
Five	1.4	41.5	765.6	322.3	4.7	536.43	351

(c) Normal Mixtures Rational Seeds

Segment	Number of Credit Cards Used	Account Age (months)	Days since First Purchase	Days since Last Purchase	Total Number of Orders	Total Dollars Spent	Segment Size
One	1.5	40.5	754.8	273.2	4.3	429.05	404
Two	3.1	45.9	725.3	557.4	1.6	163.77	141
Three	4.0	51.5	871.0	188.5	6.0	787.52	12
Four	1.0	38.9	689.4	577.6	1.5	136.20	3729
Five	1.2	41.6	809.7	483.0	4.9	1044.47	31

The rational seeds for K-means and normal mixtures were based on the centroid locations obtained from hierarchical clustering using the average method.
Total within-segment variation for a real-world data set using standardized variables (N = 4317)

Seed-type	Clustering algorithm		
	HK	K-means	Normal mixtures
Rational (based on hierachical clustering using average method)	19.29	19.66	29.21
Random (mean across 250 analyses)	19.23	20.47	27.31

Table 9. Manual data results.

(a) Average HARI by Clustering Algorithm and Data Complexity for N = 72 Artificial Data Sets						
Algorithm	Cluster density Equal 0.60		Outliers None 0.20		Variables and noisy Variables	
					8 + 1	8 + 2
					Noisy	Noisy
HK	0.92	0.74	0.93	0.72	0.84	0.82
K-means	0.80	0.87	0.92	0.75	0.82	0.85
Normal mixtures	0.66	0.73	0.92	0.48	0.68	0.72
(b) Average Total within-Segment Variation by Clustering Algorithm and Data Complexity for N = 72 Artificial Data Sets						
Algorithm	Cluster density Equal 0.60		Outliers None 0.20		Variables and noisy Variables	
					8 + 1	8 + 2
					Noisy	Noisy
HK	21.82	24.76	20.29	26.31	21.39	25.20
K-means	23.87	29.13	20.34	32.66	26.24	26.77
Normal mixtures	43.95	43.64	20.40	67.20	42.17	45.42

In part (b), because the total difference between groups is the smallest, the smaller results of the three clustering methods are better, and the results of the comparison of (a) and (b) are similar.

The results show that the HK clustering method is better than the k-means method and the mixed model in the customer clustering of retailers. Because the HK clustering method can provide higher isomorphism within the group and is less sensitive to the initial solution of the calculation process, it can obtain good results regardless of whether random or rational seeds are used, so it is more flexible in actual operation. When we use the HK clustering method to cluster market customers, we can avoid the complex operation of hierarchical clustering. The disadvantage of this method is that when the density of a group is too large, the k-means method can get better results.

Because improper selection and application of market clustering technology will seriously affect the company's financial structure, and more correct clustering results can save the company's resources and money, when creating a marketing strategy, we must select a more suitable technology for customer clustering, and then adjust the marketing strategy according to the characteristics of each different group of customers, so as to achieve the most effective strategic result.

5. Conclusions

5.1. Main Work of the Paper

The improved Algorithm A overcame the disadvantage of the k-means algorithm, where the users must give the clusters to be generated in advance. However, Algorithm A required the user to input parameter H. Parameter h was different for different situations, which was still difficult for users. Thus, facing the trend of a growing data scale, the time measure of the algorithm became increasingly higher, and Algorithm B was combined. The hierarchical aggregation algorithm and sampling technology were used to improve the k-means algorithm. The improved algorithm was more effective. Only one TB in the clustering was obtained through the data analysis of the sample set, which saved time and cost. At the same time, Algorithm B maintained the advantages of Algorithm A and could automatically generate k-class numbers. The improved Algorithm A and Algorithm B still have their own shortcomings. When encountering practical problems, the corresponding algorithm can be selected according to the actual situation.

Based on the cluster analysis in data mining, this paper explored the theory and application of the algorithm. Firstly, it made a systematic and complete analysis of clustering, including the concept of clustering, clustering algorithm, the quantitative (clustering criterion function) and qualitative (algorithm fitness) evaluation of clustering, and clustering in other fields.

Secondly, the advantages and disadvantages of existing typical clustering algorithms (k-means) were analyzed so as to facilitate further development and improvements. Aiming at the different defects of the typical k-means algorithm, Algorithms A and B were proposed. The improved Algorithm A effectively solved the problem of the initial value K automatically generating the number of classes K; the algorithm was strict in the selection of initial center points. The distance was long, which prevented the initial clustering center from selecting a class, which overcame the limitation of the algorithm to a certain extent and allowed it to enter the local optimal state. In order to further improve the computational efficiency of the algorithm, an improved Algorithm B was proposed. The original algorithm was improved by combining sampling technology and hierarchical aggregation algorithm; the new Algorithm B was then made to be more effective. Finally, in the application of the algorithm, clustering technology was applied to customer segmentation, which was established by the analytical hierarchy process for the customer value system, quantifying customer value. Based on this, clustering technology was applied to divide customers into different groups, which had a certain practical significance to effectively carry out customer management. At present, there are some customer value evaluation systems, but the measurement model is not mature enough. The measurement index is generally the direct profit the enterprise makes from the customers. There are also some difficulties in quantitative contribution. This paper used the method of data mining from the actual situation of enterprises. Based on the situation, through a series of operable customer value evaluation indicators, customer values suitable for enterprise development were established. The evaluation model was used to measure customer value and segment customers, and the decision support system of customer value management was the established system.

5.2. Further Research Directions

The two improved algorithms have not been tested on large data sets, and the new algorithm has not been fully confirmed. The effectiveness of the algorithm in clustering massive data sets cannot be fully exposed in the current test. Therefore, this part of the work needs to be deepened further. On the other hand, there are a lot of coalescence nowadays. With regard to class algorithm, users are often confused when choosing clustering algorithms and do not know which algorithm to choose. It is therefore necessary to classify the existing clustering algorithms to provide theoretical guidance for users.

6. Discussions

Problem analysis and existing improvements on the k-means algorithm. The k-means algorithm has many disadvantages. The main problems are as follows:

The number of clusters K in the k-means algorithm needs to be given in advance. The selection of this K value is very difficult to estimate. Many times, we do not know the given value in advance. How many categories of data sets should be divided into is the most appropriate, which is also a deficiency of the k-means algorithm. Indeed, the algorithm is meant to obtain a more reasonable K number of types through the automatic merging and splitting of classes, such as the 1sodata algorithm. In order to determine the K value of the number of clusters in the k-means algorithm, the variance score was used in document 17L. The mixed F statistics was used to determine the optimal classification number, and the fuzzy partition entropy was used to verify the optimal classification number. In the literature, it was proposed for "the effectiveness of clustering".

A competitive learning rule called "the second winner penalty" was used to automatically determine the appropriate number of classes. It was thought for each input; not

only the weight of the winning unit was modified to adapt to the input value, but also for the second time. The winning unit used a penalty method to keep it away from the input value. The algorithm was highly dependent on the selection of the initial value and the algorithm often fell into the local minimum solution. Different initial values often led to different results. The k-means algorithm firstly randomly selected K points as the initial clustering seed was used, and then the iterative relocation technology was used until the algorithm converged. Therefore, the difference of the initial value may be different the unstable clustering effect of the algorithm. Moreover, the k-means algorithm often uses the sum of squares of the error criterion function. As a clustering criterion function (objective function), the objective function often has many local minima, with only one belonging to the global minimum. Because the initial clustering center selected at the beginning of each algorithm falls into the center of the non-convex function, surface "position" often deviates from the search range of the global optimal solution. Therefore, the objective function often reaches the local minimum, not the global minimum. To solve this problem, many algorithms use a genetic algorithm (GA); for example, in document [9i], the genetic algorithm GA was used for initialization, and an internal clustering criterion was used as the evaluation criterion—price index. It can be seen from the framework of the k-means algorithm that the algorithm needs to continuously adjust the sample classification. Therefore, it becomes necessary to analyze the time complexity and improve the application scope of the algorithm. According to the literature, the time complexity of the algorithm can be analyzed and considered, and the clustering can be removed by a certain similarity criterion [39]. The k-means algorithm is used to cluster the sample data. Both the selection of initial points and the adjustment of data at the completion of an iteration are based on random selection. Based on the sample data, this can improve the convergence speed of the algorithm because the centroid (i.e., the mean point) of the cluster is used as the cluster center for a new round of cluster calculations, which is far away from outliers and noise points in data-intensive areas and will cause the cluster center to deviate from the real-data-intensive areas. As the k-means algorithm is sensitive to noise points and outliers. To solve this problem, the early k-medoid method will be of use. That is, the average value of the objects in the cluster is not used as the reference point, which can be selected. The most central object in the cluster is the center point. Generally, the k-means algorithm can only find spherical clusters and often uses the sum of squares of the error criterion function as the clustering criterion function (objective function). Moreover, it has a similarity measure based on Euclidean distance. It is found that if the difference between clusters is obvious, the data can be divided into different clusters. If the distribution is dense, the error square sum criterion function based on Euclidean distance is more effective.

Author Contributions: Conceptualization, C.F. and H.L.; methodology, C.F.; software, C.F.; valida tion, C.F. and H.L.; formal analysis, C.F.; investigation, C.F.; resources, C.F.; data cu-ration, C.F.; writing—original draft preparation, C.F.; writing—review and editing, C.F.; visualization, C.F.; supervision, C.F.; project administration, C.F.; funding acqui-sition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Research project of Humanities and Social Sciences in Colleges and universities of Jiangxi Province, China (Grant No. GL18107).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Research project of Humanities and Social Sciences in Colleges and universities of Jiangxi Province, China (Grant No. GL18107).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gurley, J.W. The one Internet metric that really matters. *Fortune* **2000**, *2*, 141–392.
2. Bickerton, P. 7 technologies that are transforming the hospitality industry. *Hosp. Mark.* **2015**, *1*, 14–28.
3. Pui, L.T.; Chechen, L.; Tzu, H.L. Shopping motivations on Internet: A study based on utilitarian and hedonic value. *Technovation* **2007**, *27*, 774–787.
4. Lee, E.Y.; Soo, B.L.; Yu, J.J. Factors influencing the behavioral intention to use food delivery. *Apps. Soc. Behav. Pers.* **2017**, *45*, 1461–1474. [CrossRef]
5. Armstrong, G.; Kotler, P. *Marketing*; Prentice-Hall: Englewood Cliffs, NJ, USA, 2000; pp. 1–98.
6. Ozkara, B.Y.; Ozmen, M.; Kim, J.W. Examining the effect of flow experience on online purchase: A novel approach to the flow theory based on hedonic and utilitarian value. *J. Retail. Consum. Serv.* **2017**, *37*, 119–131. [CrossRef]
7. Park, J.; Ha, S. Co-creation of service recovery: Utilitarian and hedonic value and post-recovery responses. *J. Retail. Consum. Serv.* **2016**, *28*, 310–316. [CrossRef]
8. Anderson, K.C.; Knight, D.K.; Pookulangara, S.; Josiam, B. Influence of hedonic and utilitarian motivations on retailer loyalty and purchase intention: A facebook perspective. *J. Retail. Consum. Serv.* **2014**, *21*, 773–779. [CrossRef]
9. Chiu, C.M.; Wang, E.T.; Fang, Y.H.; Huang, H.Y. Understanding customers' repeat purchase intentions in B2C E-Commerce: The roles of utilitarian value, hedonic value and perceived risk. *Inf. Syst. J.* **2014**, *24*, 85–114. [CrossRef]
10. Lin, K.Y.; Lu, H.P. Predicting mobile social network acceptance based on mobile value and social influence. *Internet Res.* **2015**, *25*, 107–130. [CrossRef]
11. Huang, J.H.; Yang, Y.C. Gender difference in adolescents' online shopping motivation. *Afr. J. Bus. Manag.* **2010**, *4*, 849–857.
12. Grandom, E.; Mykytyn, P. Theory-based instrumentation to measure the intention to use electronic commerce in small and medium sized businesses. *J. Comput. Inf. Syst.* **2004**, *44*, 44–57.
13. Chan, T.K.H.; Cheung, C.M.K.; Shi, N.; Lee, M.K.O. Gender differences in satisfaction with Facebook users. *Ind. Manag. Data Syst.* **2015**, *115*, 182–206. [CrossRef]
14. Meyers-Levy, J.; Sternthal, B. Gender differences in the use of message cues and judgments. *J. Mark. Res.* **1991**, *28*, 84–96. [CrossRef]
15. Jo¨reskog, K.G.; So¨rbom, D. *LISREL 8: User's Reference Guide*; Scientific Software International: Chicago, IL, USA, 1996; pp. 1–98.
16. Arnold, M.J.; Reynolds, K.E. Hedonic shopping motivations. *J. Retail.* **2003**, *79*, 77–95. [CrossRef]
17. Schwab, D.P. *Research Methods for Organizational Studies*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2005; pp. 1–30.
18. Byrne, B.M. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2001; pp. 1–28.
19. Swaminathan, V.; Lepowska, W.E.; Rao, B.P. Browsers or buyers in cyberspace: An investigation of factors influencing electronic exchange. *J. Comput.-Mediat. Commun.* **1999**, *5*, 224–234. [CrossRef]
20. Stützle, T.; Hoos, H. Improvements on the Ant System: Introducing MAX—MIN ant System. In Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms, East Lansing, MI, USA, 19–23 July 1997; Springer: Vienna, Austria, 1997; pp. 245–249.
21. Yang, Z.; Li, H.; Haodong, Z. An improved k-means dynamic clustering algorithm. *J. Chongqing Norm. Univ. Nat. Sci. Dep. Acad. Ed.* **2016**, *33*, 97–101.
22. He, Q.; Wang, Q.; Zhuang, F.; Tan, Q.; Shi, Z. Parallel CLARANS Clustering Based on MapReduce. *Energy Procedia* **2011**, *13*, 3269–3279.
23. Karypis, G.; Han, E.H.; Kumar, V. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *IEEE Trans. Comput.* **1999**, *32*, 68–75. [CrossRef]
24. Guha, S.; Rastogi, R.; Shim, K. CURE: An efficient clustering algorithm forlarge databases. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, Seattle, WA, USA, 2–4 June 1998; ACM Press: New York, NY, USA, 1998; pp. 73–84.
25. Yang, X. *Application of Improved Birch Algorithm in Telecom Customer Segmentation*; Hefei University of technology: Hefei, China, 2015; p. 77.
26. Zhang, J.; Wang, L.; Yao, Y. A traffic classification algorithm based on options clustering. *Zhengzhou Inst. Light Ind. J. Nat. Sci.* **2013**, *28*, 83–86.
27. Hinneburg, A.; Keim, D. An efficient approach to clustering large multimedia database with noise. In Proceedings of the 4th ACM SIGKDD on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 August 1998; AAAI Press: New York, NY, USA, 1998; pp. 58–65.
28. Wang, D.; Wang, I.; Fan, W. An improved clique high dimensional subspace clustering algorithm. *Semicond. Light Dian* **2016**, *37*, 275–278.
29. Sheikholeslami, G.; Chatterjee, S.; Zhang, A. WaveCluster: A Multi—Resolution Clustering Approach for Very Large Spatial Database. In Proceedings of the 24th Conference on VLDB, New York, NY, USA, 24–27 August 1998; pp. 428–439.
30. Liang, J. *Research on Ant Colony Algorithm and Its Application in Clustering*; South China University of Technology: Guangzhou, China, 2011.
31. Tan, S.C.; Ting, K.M.; Teng, S.W. Simplifying and improving clustering ant-based. In Proceedings of the International Conference on Computational Science 2011, Dalian, China, 24–26 June 2011; pp. 46–55.

32. Haiying, M.; Yu, G. Customer Segmentation Study of College Students Basedon the RFM. In Proceedings of the 2010 International Conference on E-Business and EGovernment, Guangzhou, China, 7–9 May 2010; pp. 3860–3863. [CrossRef]
33. Sheshasaayee, A.; Logeshwari, L. An efficiency analysis on the TPA clustering methods for intelligent customer segmentation. In Proceedings of the 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 21–23 February 2017; pp. 784–788.
34. Srivastava, R. Identification of customer clusters using RFM model: A case of diverse purchaser classification. *Int. J. Bus. Anal. Intell.* **2016**, *4*, 45–50.
35. Kamel, M.H. Topic discovery from text using aggregation of dlfferent clustering methods. In *Advances in Artificial Intelligence. Canadian AI 2002*; Lecture Notes in Computer Science; Cohen, R., Spencer, B., Eds.; Springer: Berlin, Heidelberg, 2002; Volume 2338, pp. 161–175.
36. Zahrotun, L. Implementation of data mining technique for customer relationship management (CRM) on online shop tokodia-pers.com with fuzzy c-means clustering. In Proceedings of the 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 1–2 November 2017; pp. 299–303.
37. Tong, L.; Wang, Y.; Wen, F.; Li, X. The research of customer loyalty improvement in telecom industry based on NPS data mining. *China Commun.* **2017**, *14*, 260–268. [CrossRef]
38. Shah, S.; Singh, M. Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. In Proceedings of the 2012 International Conference on Communication Systems and Network Technologies, Rajkot, India, 11–13 May 2012; pp. 435–437.
39. Liu, C.C.; Chu, S.W.; Chan, Y.K.; Yu, S.S. A Modified K-Means Algorithm—TwoLayer K-Means Algorithm. In Proceedings of the 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kitakyushu, Japan, 27–29 August 2014; pp. 447–450. [CrossRef]
40. Cho, Y.; Moon, S.C. Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system. *J. Converg.* **2013**, *4*, 36–40.
41. Jiang, T.; Tuzhilin, A. Improving personalization solutions through optimal segmentation of customer bases. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 305–320. [CrossRef]
42. Lu, H.; Lin, J.; Lu, J.; Zhang, G. A customer churn prediction model in telecom industry using boosting. *IEEE Trans. Ind. Inf.* **2014**, *10*, 1659–1665. [CrossRef]
43. He, X.; Li, C. The research and application of customer segmentation one-commerce websites. In Proceedings of the 2016 6th International Conference on Digital Home(ICDH), Guangzhou, China, 2–4 December 2016; pp. 203–208. [CrossRef]