K-LRFMD: Method of Customer Value Segmentation in Shared Transportation Filed Based on Improved K-means Algorithm

Hong Li , Xiaosheng Yang , Yao Xia, Lujie Zheng , Guoqing Yang and Pan Lv

Zhejiang University, Hangzhou, 310027, China

lihong@zju.edu.cn

Abstract. The advent of information age has transformed the focus of enterprise marketing from product-centric to customer-centric, and customer relationship management becomes the core problem of enterprises. Accurate customer value classification results are an important basis for enterprises to optimize marketing resources allocation, and customer value classification is becoming one of the key issues that need to be solved urgently in customer relationship management. In the face of the fierce market competition of the vehicle-sharing industries, each shared transportation company has introduced more preferential marketing methods to attract more customers. In this paper, with the aid of the vehicle-sharing platform in a domestic university campus, we established a reasonable customer value evaluation model called K-LRFMD. K-LRFMD did some clustering analysis with the customers based on specific feature engineering and improved K-means algorithm. In this paper, we compare different customer value derived from K-LRFMD. The analysis can formulate the corresponding marketing strategy to provide personalized customer service for different customers.

1. Introduction

In recent years, a lot of capital is flocking to make China's shared transportation filed blustery. Ride-sharing enterprises such as ofo and mobike and car rental company DiDi accelerate competition of traffic market share. Its rapid ascent, universal attention, also explain that people's a great demand of the mass travel. How to use these massive customer data to establish effective communication among relevant business units (such as marketing and customer service) about the segmentation and focus limited resources on high-value customers to achieve the goal of maximizing corporate profits is the current very popular research topic.

Although some literatures [1, 2] introduce some customer segmentation applied to fields of banking, insurance, network account, etc. There is no literature involving shared transportation filed.

There are a lot of data accumulated in the shared transportation filed. However, these data with many attributes is huge, making it difficult to extract valuable information from them. In order to ensure the accuracy of customer segmentation, we must choose the correct and reasonable classification indicators and classification methods. Currently the most widely-value model identifies customers by three indicators (Recent Frequency Monetary), referred to as RFM model [3].Monetary represents the sum of the amount of product purchased over a period of time in the RFM model in [4]. Due to the factors such as driving distance, promotion activities and other factors affecting the business of vehicle-sharing operation, the Monetary is different for different customers with the same consumption amount. Therefore, this indicator does not apply to customer value analysis. [5] proposed AFH customer classification model based on the analysis of RFM model and introduced customer value matrix model. The model adds an indicator of the average consumption amount, but in essence it still doesn't take any factors such as discount coefficient and mileage into account. In terms of classification, clustering analysis is commonly used nowadays. Currently the most widely used clustering algorithm is the K-means algorithm [6, 7]. The effect of K-means algorithm proposed in [8, 9] is greatly influenced by factors such as number of clusters and initial cluster centres. Studies show that the influencing factors mentioned above are related to the specific case and subjective experience. Therefore, this paper replaces the original consumption amount(M) with the two indicators, such as the travel distance(M) accumulated by the customers in a certain period of time and the average value of the discount coefficient(D) enjoyed by the customers in a certain period of time. In addition, considering the membership mechanism of the vehicle-sharing platform, the length of the joining membership can affect the customer value to a certain degree, so the length of the membership (L) is added into the model as another indicator for differentiating the customers. In order to solve the two shortcomings associated with the initial value of K-means algorithm, this paper will set the number of clusters K as 5, and proposed an initial clustering centre selection strategy based on the specific case and competent experience. Finally, we evaluated the customer segmentation based on the actual application scenario.

In short, the model we developed called K-LRFMD has the following key features:

- Special feature engineering: Based on the traditional RFM model, we propose features that are more consistent with customer segmentation in the area of vehicle-sharing.
- Reasonable initial clustering centre selection strategy: We implement a maximum probability of selecting initial clustering centre strategy based on Euclidean distance.
- The vehicle-sharing platform is already in use of the model: Our model has been used in vehicle-sharing platform operated in campus of some universities.

The rest of this article is organized as follows. The second section is an overview of our modelling process. The third section details the method we propose. In the fourth part, we systematically evaluate the performance of the model based on the actual application scenario. The fifth section is discussion and limits of the model. Finally, the sixth section concludes.

2. Overview of building a K-LRFMD customer segmentation model

This section will do a brief introduction to the process of building a K-LRFMD customer segmentation model. In this paper, five indicators of customer relationship: length L, consumption time interval R, consumption frequency F, driving distance M and average discount coefficient D are taken as the customer value index of vehicle-sharing platform(see Table 1) and recorded as LRFMD model.

Table 1. Index meaning

model	L	R	F	M	D
K-LRFMD	The number of days from of the Member Registration Time to the end of the observation	The number of days from the time of customers last driving to the observation window	The number of times customers drive in the observation window	Customer accumulated mileage in the observation window	The average discount factor enjoyed by customers in the observation window

According to LRFMD, if the traditional binning method of attribute of RFM model is used, as shown in Figure.1, it is classified according to the average value of the property, which is bigger than the mean value is expressed as ↑, on the contrary is expressed as ↓, although the model can also identify the most valuable customers, but the number of the customer segmentation is too much, and increase the cost of targeted marketing. Therefore, this paper adopts clustering method to identify customer value. Based on the improved K-means algorithm, five indicators of LRFMD model are clustered to identify the most valuable customers.

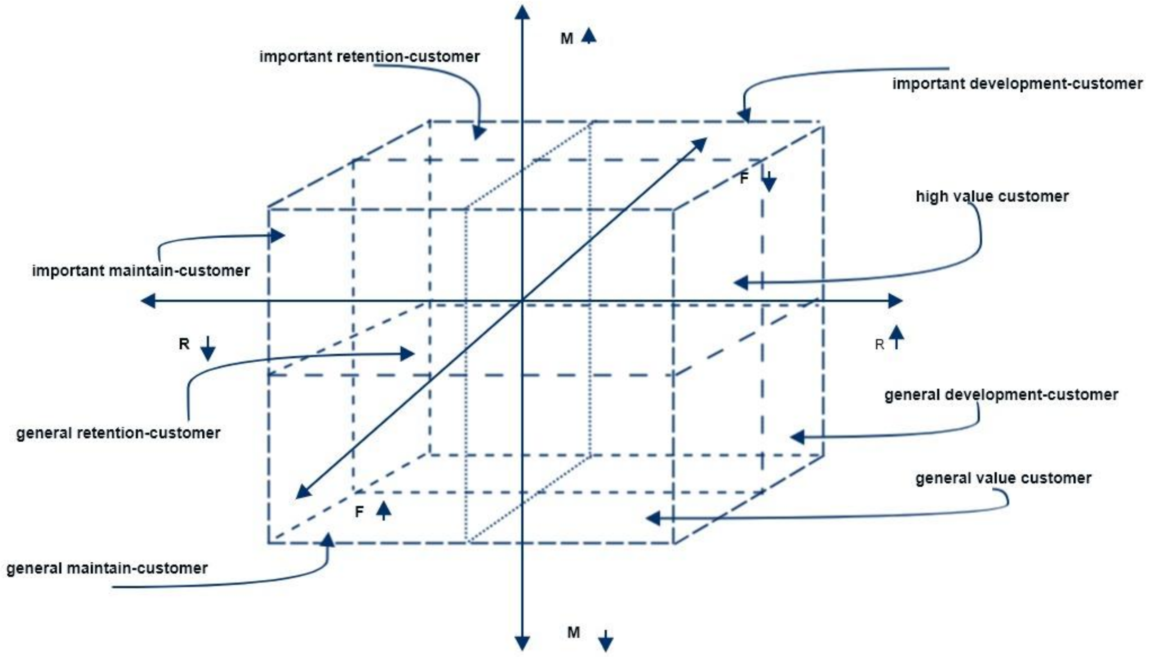
M
important retention-customer
important development-customer
F
high value customer
important maintain-customer
R
R
general development-customer
general retention-customer
general value customer
F
general maintain-customer
M

Figure 1. RFM Model Analysis

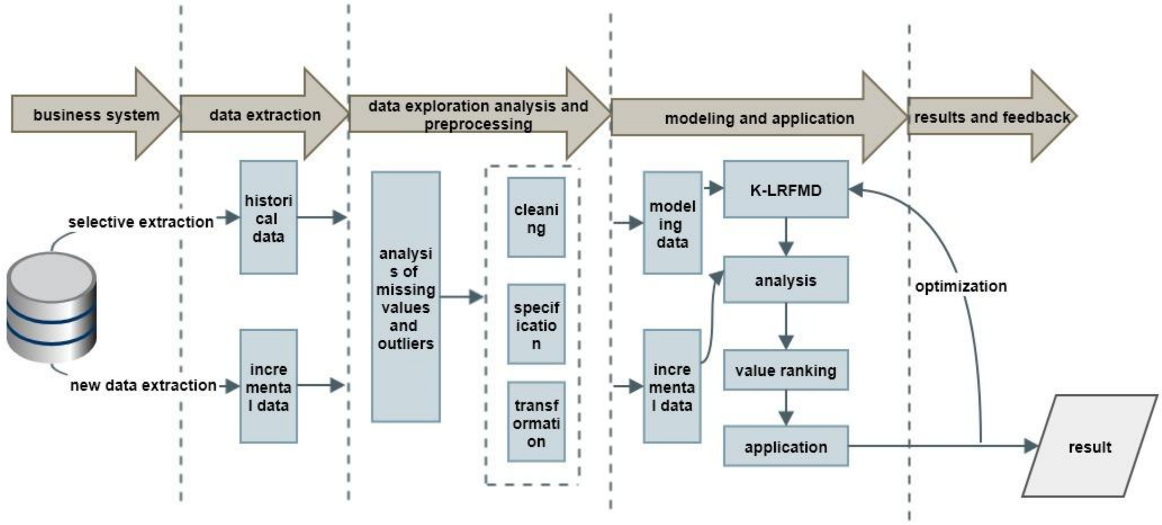

Figure 2. The Overall Process of Building K-LRFMD

The overall process of customer segmentation is shown in Figure.2 and the overall process of building K-LRFMD mainly includes the following steps.

- From the data sources operated by a vehicle-sharing platform in a university, do selective extraction and new data extraction to form historical data and incremental data, respectively.
- Perform Data exploration analysis and pre-processing of two datasets formed in step (1), including analysis of missing values and outliers, attribute specification, cleaning and transformation.

- Based on the modelling data of the completed data pre-processed in step (2), and the improved K-means algorithm, apply K-LRFMD model, and analyse the characteristics of each customer group to identify valuable customers.
- According to the results of the model, using different marketing tools to provide customized services.

3. K-LRFMD

In this section we will detail the specific processes and methods of building a model.

3.1. Data Extraction

For the ending time of November 30, 2017, the time window with a width of two years is selected as the analysis observation window, and the detailed data of all the customers are taken to form the historical data. For the subsequent added customer details, follow up the newest time point in the database as the end time, and extract the data in the same way as above to form incremental data.

From detailed information such as vehicle information and payment records in the database of vehicle-sharing platform in a university, the detailed data of all customers from November 30, 2015 to November 30, 2017 are extracted based on Last _Drive Time of the last driving date, for a total of 50715 records. Which contains the user ID, driving mileage, points, discounts and other 29 attributes.

3.2. Data Exploration and Analysis.

Data exploration and analysis is the analysis of the data missing value and outlier. Through the observation of the data, there is a record that the original data has a payment equal 0, a discount equal 0 and a total mileage more than zero. This phenomenon may generate because of free trial or redeeming points of the customer. The partial result of data exploration and analysis is in Table 2.

3.3. Data Preprocessing

This paper mainly uses the data cleaning, attribute specification and data transformation pre-processing method.

Table 2. Data Exploration Results Analysis Table

Attribute	Null value records	Maximum	Minimum
User id	0	47042	1939
Current miles	0	13710	0
...	...	...	...
Cost	11	78.1	-29.4
Car id	0	246	68

Data Cleaning Through data exploration and analysis, there exist missing value and abnormal value less than zero in the Cost and Money attributes. Due to the huge amount of raw data, such data occupy a relatively small proportion and have little effect on the result and therefore can be discarded. Specific treatment is as follows:

- Discard missing values
- Discard value less than 0 in the Cost and Money

Attribute specification There are too many attributes in the original data. According to the K-LRFMD model proposed in this paper, the six attributes associated with KLRFMD are selected, namely User id, Start time, Load time, Cost, Money, bonus.

Data transformation The data transformation is to transform the data into an “appropriate” format to accommodate the need for mining tasks and algorithms. In this paper, the data transformation method is used for attribute construction and data standardization. Since the original data does not directly give the five indexes of the K-LRFMD model, the five indexes need to be extracted through raw data. The specific calculation methods are as follows:

- L = Load time − Start time

The number of days from the Member Registration Time to the end of the observation window [unit: day]

- R = Last To End

The number of days from the time of customers last driving a shared car to the observation window while [unit: day]

- F = Drive Count

The number of times customers drive a shared car in the observation window [unit: day]

- M = SumCurrentMiles

Customer accumulated mileage in the observation window [unit: mile]

- D = Sum Bonus/Count

The average discount factor enjoyed by customers in the observation window [unit: None]

Table 3. K-LRFMD index range

Attribute	L	R	F	M	D
MAX	450	448	369	239968	111
MIN	1	1	1	0	0
AVG	257.39	168.54	14.57	9308.94	3.87

After five indicators are extracted, the data distribution of each indicator needs to be analysed. The range of the data is shown in Table 3.

From the data in the table, it can be found that the values of the five indexes vary greatly. In order to eliminate the impact of the magnitude of data, the data needs to be standardized. This paper uses Z-score approach and the result of some data processing is as shown in Table 4.

Table 4. Normalized data set

ZL	ZR	ZF	ZM	ZD
1.20882776	1.312803469	-0.185458018	-0.554793993	-0.526600821
0.956659445	1.497281062	-0.347784511	-0.554793993	-0.526600821
1.249094637	1.83109238	-0.510111005	-0.552708063	-0.322491976
0.872614763	1.463577819	-0.550692628	-0.554793993	-0.526600821
-1.581091175	-1.016106791	-0.550692628	-0.554793993	-0.526600821
1.033287101	1.625940291	-0.510111005	-0.554793993	-0.526600821
0.847914913	1.43860888	-0.510111005	-0.554793993	-0.526600821
1.230696342	-0.901523796	1.356643666	-0.091002535	2.085992401

3.4. K-means improved algorithm

K-means algorithm k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells [10, 11]. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms [12] that are commonly employed and converge quickly to a local optimum.

Improve K-means algorithm Two key features that make the K-means algorithm highly efficient are also seen as its biggest drawbacks:

- The number of clusters k is an input parameter. Choosing an inappropriate value of k may result in poor clustering results. This is why feature checking is needed to determine the number of clusters in a dataset.

- Convergence to a local optimal solution may lead to “counter-intuitive” erroneous results.

An important limitation of the k-means algorithm lies in its clustering model. The basic idea of this model is that we get spherical clusters separated from each other, in which the mean points tend to converge to the centre of the cluster. It is generally expected that the cluster sizes will be roughly equal, so assigning each observation to the nearest cluster centre (i.e. mean point) is the correct allocation.

Improve the initial clustering centre selection method The traditional k-means clustering algorithm obtains the last k centres through the initial central iteration. The relationship between the final clustering results and the initial clustering centres is still close, and different initial centres may get completely different results. [13] proposed the idea of data segmentation to choose the initial clustering centre. [14] proposed to determine the initial cluster centres based on distance estimation. In [15], K-means++ is proposed to determine the initial cluster centres based on the maximum probability. Although the K-means++ algorithm proposed in [16] can definitely initialize the clustering centre, it has one shortcoming in the scalability: its inherent ordering property: the choice of the next centre depends on Centre of choice. In response to this shortcoming, this paper changes the sampling strategy of each traversal based on K-means++, each traversal takes $O(k)$ samples, the sampling process is repeated about $O(logn)$ times, a total of $O(klogn)$ sample points obtained after repeated sampling, the set of constant factor approximation of the optimal solution, and then the $O(klogn)$ points are clustered into k points.

Table 5. Customer classification situation

ZL	ZR	ZF	ZM	ZD	num per
-1.26534934	-0.93586135	-0.24461487	-0.22321496	-0.68739483	1030 29.59%
0.56052897	0.81107607	-0.33850695	-0.35428804	-0.01618059	1415 40.65%
0.6649835	-0.90491723	4.04972263	4.07757598	-0.3143415	108 3.10%
0.3475817	0.87571874	-0.53503839	-0.46243334	2.44529302	376 10.80%
0.55734129	-0.75230912	0.89627725	0.84189442	-0.28001647	552 15.86%

Finally, these k points are sent to Lloyd iteration as the initial cluster centres. The actual experiments prove that $O(logn)$ sub-sampling is not needed; 5 times repeat sampling can get a better initial centre of clustering. The method is as follows:

- Randomly select $O(k)$ points from the set of input data points as the cluster centres and repeat the sampling five times to obtain a set of $O(5*k)$ sample points and then cluster them into k initial centres point;
- For each point x in the dataset, calculate the distance $D(x)$ between it and the nearest cluster centre (referring to the selected cluster centre), and select a new cluster centre based on the maximum probability criterion of Euclidean distance;
- Repeat step 2 until k clustering centres are found. In step 2, the distance between each data point and the nearest seed point (cluster centre) is calculated in turn, and a set D consist of $D(1), D(2), ..., D(n)$. n represents the size of the data set.

In D, to avoid noise, you cannot directly select the element with the largest value. You should select the element with a larger value, and then use the corresponding

The number of clusters K value selection When using K-means algorithm, it is necessary to specify the number of clusters k. The algorithm is based on the field of customer value segmentation and can take k as 5 based on engineering experience.

Clustering results The K-means algorithm is used to cluster the standardized data including L, R, F, M and D, and the clustering results are shown in Table 5 and Figure.3.

4. Customer Value Analysis

According to the result of clustering, the characteristic analysis is shown in Figure.4.

Customer group 1 has the smallest R attribute; Customer group 2 has the largest R attribute; Customer group 3 has the largest L, F, M attributes and R attribute is smaller; Customer group 4 has the largest D, R attributes. According to the specific business analysis of vehicle-sharing platform, the characteristics of a group were evaluated by comparing the size of each index among the groups. For example, customer group 3 has the largest attribute L, F, M, and the lowest attribute R, so it can be said that L, R, F and M are inferior features in customer group 3. And so on, in order to sum up the advantages and disadvantages of each group features, the specific results shown in Table 6.

The chart analysed by the above characteristics shows that each customer group has significantly different performance characteristics. Based on this feature description, the present case defines five levels of customer categories: important maintain-customer, important development-customer, important retention-customer, general customer, Low value customers. Each customer category is as follows:

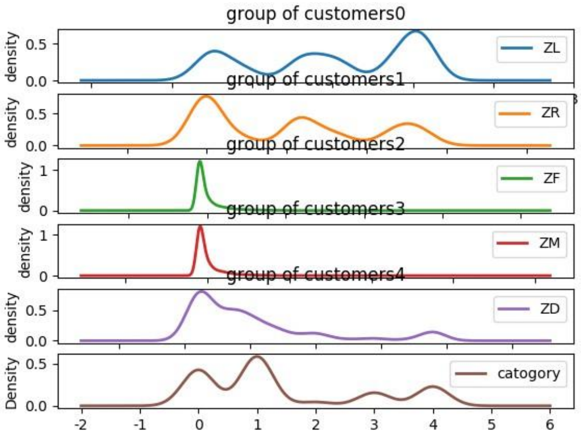

Figure 3. Cluster Diagram

Table 6. Customer Group Characteristics Description Table

Group category	Advantageous features	Weak features
Customer group 1	R L	
Customer group 2		R
Customer group 3	L R F M	
Customer group 4		R F M
Customer group 5		

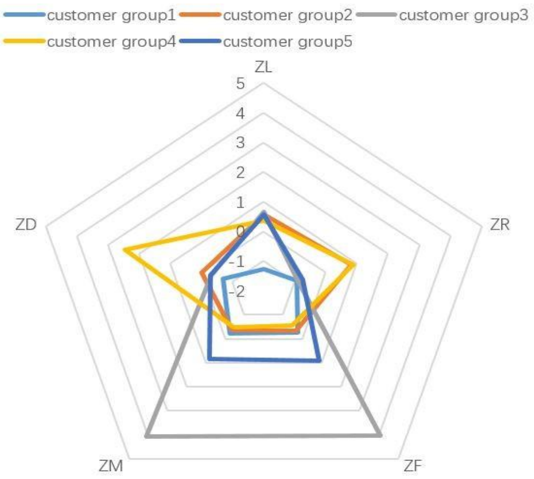

Figure 4. Analysis of Customer Characteristics

- Important maintain-customer: In general, such customers lower the average discount rate. Because the vehicle-sharing platform will hold some discount activities at certain time such as promotions. The lower average discount rate corresponds to more use times. The number of times that customer has used the driver recently (R) is low, the driving frequency (F) is high, and the driving mileage (M) is high, and the member time (L) is long. They are the ideal customer type and loyal customers of the vehicle-sharing platform, making the largest contribution to the operation of the vehicle-sharing platform, but the smallest proportion (3.10%).
- Important development-customer: Although these members have a short membership time (L), they have short R and driving range (M) and driving frequency (F) are relatively large, accounting for 29.59% of the total.
- Important retention-customer: These customers have a long membership time (L), a long R, but the total mileage (M) and driving times (F) are not low, accounting for (15.86%) which indicates the need for retain such customers. We should analyse why they have recently not use the service, and need to maintain more interaction with the customers.
- General customers and low-value customers: Such customers have not driven for a long time. Number (F) or mileage (M) less, accounting for (51.45%).

According to the characteristics of each customer type, customer value ranking of various customer groups, the results shown in Table 7.

Table 7. Customer Group Value Ranking

Customer group	Ranking	Ranking meaning
3	1	important maintain-customer
1	2	important development-customer
5	3	important maintain-customer
2	4	general customer
4	5	low value customers

5. Discussions and Limitations

The model proposed in this paper uses historical data modelling, with the change of time, the observation window of the analysis data is also changing. Therefore, it is recommended that the model should be run once a month considering the details of new customers and the actual situation of the business to judge clustering centre of new customers. Meanwhile, the characteristics of new customers are analysed. If the actual situation of incremental data and the results of the judgments vary widely, you need to focus on the business department to see the reasons for the changes and to confirm the stability of the model. If the stability of the model varies greatly, the model needs to be retrained to adjust.

6. Conclusion

This paper builds a reasonable customer value segmentation model K-LRFMD based on the shared transportation platform. Based on the improved K-means algorithm, Customers are clustered and divided into five types. We established customer segmentation table for customer value analysis, and put forward the corresponding marketing strategy. Empirical studies show that the proposed model and improved algorithm can effectively classify customers in shared transportation filed and differentiate between valueless and high value customers. The proposed K-LRFMD can also be extended to other areas such as aviation customer value segmentation, bank customer value segmentation and so on.

Acknowledgments

This work is partially supported by NSFC61602404, Zhejiang Provincial Natural Science Foundation LY16F020007, and the Fundamental Research Funds for the Central Universities 2017FZA5015.

References

[1] Li W, Wu X, Sun Y, et al. Credit Card Customer Segmentation and Target Marketing Based on Data Mining[C]// International Conference on Computational Intelligence and Security. IEEE, 2011:73-76.

[2] Jiang T, Tuzhilin A. Improving Personalization Solutions through Optimal Segmentation of Customer Bases[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(3):305-320.

[3] Wu J, Lin Z. Research on customer segmentation model by clustering[C]// International Conference on Electronic Commerce. ACM, 2005:316-318.

[4] Xu Xiangbin,Wang Jiaqiang,Tu Huan,et al. Customer classification of Ecommerce based on improved RFM model[J]. Journal of Computer Applications 2012,32(5):1439-1442.

[5] Wang K F. Applied Research on AFH Customer Classification based on Data Mining Technology[J]. Technoeconomics & Management Research, 2012.

[6] Berry M J, Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support[M]. John Wiley & Sons, Inc. 1997.

[7] Berkhin P. A Survey of Clustering Data Mining Techniques[J]. Grouping Multidimensional Data, 2002, 43(1):25–71.

[8] Ngai E W T, Xiu L, Chau D C K. Application of data mining techniques in customer relationship management: A literature review and classification[J]. Expert Systems with Applications, 2009, 36(2):2592-2602.

[9] Zakrzewska D, Murlewski J. Clustering Algorithms for Bank Customer Segmentation[C]// International Conference on Intelligent Systems Design and Applications. IEEE Computer Society, 2005:197-202.

[10] MacKay, David. Chapter 20. An Example Inference Task: Clustering. Information Theory, Inference and Learning Algorithms. Cambridge University Press.2003: 284292. ISBN 0-521-64298-1. MR 2012999

[11] Aurenhammer, Franz (1991). ”Voronoi Diagrams A Survey of a Fundamental Geometric Data Structure”. ACM Computing Surveys. 23 (3): 345405.

[12] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965, 21: 768769.

[13] Liu C, Zeng L, Zhang J, et al. An optimized K-means clustering algorithm for CMP systems based on data set partition[J]. Journal of Computational Information Systems, 2015, 11(13):4727-4738.

[14] Raed T. Aldahdooh, Wesam Ashour. DIMK-means Distance-based Initialization Method for K-means Clustering Algorithm[J]. International Journal of Intelligent Systems & Applications, 2013, 5(2074-904X):41-51.

[15] Bahmani B, Moseley B, Vattani A, et al. Scalable K-means++[J]. Proceedings of the Vldb Endowment, 2012, 5(7):622-633.

[16] Jdrzejowicz J, Jdrzejowicz P, Wierzbowska I. Apache Spark Implementation of the Distance-Based Kernel-Based Fuzzy C-Means Clustering Classifier[J]. 2016.