RFM high-speed railway customer value classification model based on spark

RFM high-speed railway customer value classification model based on spark

WEI Zhengzheng[1], SHAN Xinghua[2]

[1]Railway Technology Research College, China Academy of Railway Science Corporation Limited, Beijing

[2]Railway Technology Research College, China Academy of Railway Science Corporation Limited, Beijing

[1] ninna.weina@163.com, [2]shanxinghua@vip.sina.com

Abstract. As implementing customer relationship management (CRM) is the future development trend of high speed railway, based on the customer value of customer classification research has important theoretical and realistic significance. Based on the huge number of the high speed railway's customers, this paper proposes a parallel RFM customer value classification model based on the Spark framework .First, based on the RFM customer value model, calculating the coefficient of customer's value, then based on the Spark framework, designing the parallel genetic k - means algorithm. The experiment proved this model has the quality of computing quickly and high precision. It has the obvious practical significance applied to the customer relationship management system.

1. Introduction

With the continuous construction of China's passenger high-speed railway network, the railway passenger transport capacity has been greatly released, the transportation market has been transformed from the seller's market to the buyer's market, and the increasingly fierce market competition between various modes of transportation has gradually formed. Therefore, paying attention to the relationship between enterprises and customers, implementing customer relationship management are the trend of high-speed railway passenger transportation development in the future. Customer classification research based on customer value has important theoretical and practical significance. Since customers of different values have different meanings to the company, it is necessary to classify customers based on customer value. Customer classification based on customer value is the first step for the company to conduct customer relationship management, which will lay the foundation for the implementation of customer strategic management.

The RFM model is a widely used and very important customer value segmentation method. In 1994, Hughes [1] proposed the RFM customer value analysis model, using three indicators: the recent consumption R (recency), consumption frequency F (frequency), consumption amount M (monetary) to determine customer value. Since the general business customer transaction data can provide these consumption information, the RFM model has become one of commonly used customer value analysis methods in enterprises.

For traditional customer relationship management (CRM), customer value research uses traditional serial clustering algorithms to classify RFM indicators, such as k-means, SOM, spectral clustering, etc[2]. However, in practical applications, the traditional customer value clustering method has the

problem that when the customer data volume is very large, the calculation time is long, and even the classification result cannot be calculated. Based on the above situation, this paper proposes a RFM customer value classification model based on Spark parallel framework. This paper takes China's high-speed railway customer as an example, and applied it to the model. Compared with the traditional method, the model has better classification experiment results.

2. Spark-based RFM high-speed railway customer value classification model

2.1. Model frame design

Based on Spark's RFM high-speed railway customer value classification model, the customer value parameters are calculated based on the RFM model, and the genetic algorithm is combined with k-means clustering and parallelized based on the Spark framework, so that the customer value classification results and calculation speed are better than traditional methods. Classification algorithms have obvious advantages. The specific model framework is shown in Figure 1:

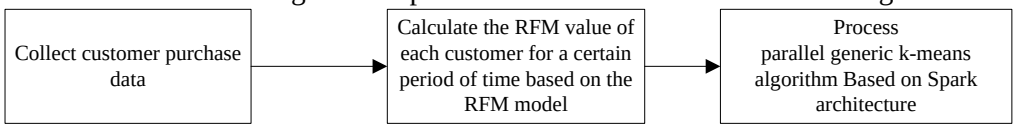

Figure 1. Spark-based RFM high-speed rail customer value classification model framework design

2.2. RFM model definition

The RFM method is a very important customer classification method that is widely used at present and is the most commonly used and practical predictors. It mainly judges the value of customers by analyzing the three important customer behavior indicators of “Recent consumption”, “Consumption frequency” and “Consumption amount”. Therefore, the RFM model [4] is based on the three dimensions of R, F, and M to distinguish customers and evaluate customer value. The RFM classification method has a set of fixed operational procedures in the application process because of its few indicators and simple data. However, the traditional “F” and “M” definition methods have multiple collinear problems. In this paper, "M" is changed to calculate the average amount of customers spent at a certain time .

In this paper, these three indicators R, F, and M for defining the high-speed railway customer value classification model are as follows:

R: given a time window, the time interval between the high-speed train customer taking the high-speed trains and the shortest time interval of this time window;

F: The frequency at which high-speed rail customers take high-speed trains during this time window;

M: The average amount of money spent by high-speed trains’ customers during this time window.

2.3. Spark-based parallelized genetic k-means algorithm

2.3.1. Spark architecture

With the emergence of distributed file systems such as the HDFS (Hadoop Distributed File System), it has become possible to store massive amounts of data. The main purpose is to use an algorithm platform that is applied to a large amount of data and many iterations of the algorithm. Different from MapReduce, the intermediate result value of the calculation needs to be written to the disk each time. Spark introduced the concept of RDD (Resilient Distributed Dataset) [6-8] (a parallel data structure based on distributed memory that allows users to store data in memory and control partitioning to optimize data distribution). It can cache results in memory during the cluster calculation process, which greatly increases the calculation speed. Spark's operating architecture is shown in Figure 2. Its calculation process can be summarized as conversion between different RDDs. Spark provides a variety of dataset operations, unlike Hadoop which only provides Map and Reduce operations. As

shown in Figure 2, flatMap, map, reduceByKey, saveAsSequenceFile and other types of operations, in which a variety of operational RDD types are divided into two categories: transformations and actions.

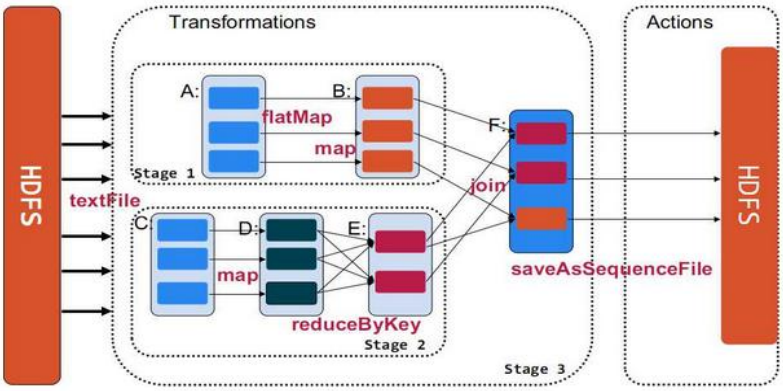

Figure 2. Spark operation architecture diagram

2.3.2. Genetic K-means algorithm

In cluster analysis, k-means algorithm is widely used in practical applications because of its simple implementation and easy parallelization, but its Achilles heel is very sensitive to initial values and easy to fall into local minimum values. The genetic algorithm is a method to search for the optimal solution by simulating the natural evolution process. Its distinctive feature is implicit parallelism and effective utilization of global information. Therefore, a k-means clustering method based on genetic algorithm is generated. It can not only play the global optimization ability of genetic algorithm, but also take into account the local search ability of k-means algorithm, so as to better solve the clustering problem. Because of its iterations, although the clustering accuracy is improved, k-means clustering also leads to inefficiencies, especially when dealing with large data sets. For the high-speed railway customer value clustering, due to the large base of its customer, if the traditional serial genetic k-means algorithm is used, it will inevitably lead to a long time of clustering, and even no calculation result. Considering that the genetic algorithm has natural parallelism and iterations, the Spark platform can be used to design and write the genetic clustering k-means algorithm. This not only improves the time efficiency of the algorithm but also increases the diversity of the population, which can prevent the occurrence of premature phenomenon.

The algorithm mainly performs genetic operations by defining p chromosomes as a population, and obtains optimal clustering results. First, each cluster center is used as a chromosome gene, and then the p chromosomes are initialized through the spark platform. Parallel k-means calculation is performed for each chromosome, and the optimal chromosome clustering result is selected by the fitness function. Finally, the judgment for the end condition is made. If the optimal solution or the maximum genetic algebra is not reached, select, cross, and mutate the chromosome to obtain a new chromosome, and then repeat the k-means calculation until it meets the end condition and obtains the final aggregation of classification result. Among them, the fitness function is defined as $\rho(K)$:

$$\rho(K) = \frac{1}{K}\sum_{n=1}^{K}\left(min_{1\le m\le k, m\ne n}\left\{\frac{\eta_n+\eta_m}{\delta_{nm}}\right\}\right) \quad (1)$$

$$\eta_n = \frac{1}{\|G^n\|}\sum_{G_i\in G^n} Sim(C_i, C^n) \quad (2)$$

$$\eta_m = \frac{1}{\|G^m\|}\sum_{G_i\in G^m} Sim(C_i, C^m) \quad (3)$$

$$\delta_{nm} = Sim(C^n, C^m) \quad (4)$$

Equation (2) defines the average similarity distance between a cluster center C_i and the middle cluster C^n in η_n; Equation (3) defines the average similar distance between a cluster center $C_i$and the middle cluster C^m in η_m; Equation (4) defines the similarity between C^n and C^m. Equation (1) is defined by equations (2), (3), and (4), where K is the number of clusters defined.

2.3.3. Implementation based on Spark genetic k-means algorithm

Spark implements k-means parallelization algorithm. The Spark cluster environment is initialized first; the original data is loaded from HDFS, and the RDD object is constructed to map and cache the RDD according to the definition of the RFM model. Then generate a new RDD object, and normalize each newly generated RDD object, store its data object in a partitioned form on each spark cluster node, and execute the algorithm in each node until the convergence condition is met or the maximum number of iterations is reached and the clustering operation ends.

Parallelized genetic k-means algorithm based on Spark framework:

Program input: Context, K, MaxNum, Max_generation. Where Context is the Spark environment parameter, K is the number of clusters, MaxNum is the number of populations, and Max_generation is the largest genetic algebra.

Program output: K cluster centers. The core steps are as follows:

1. Construct a sample instance value from the RFM model; // build a sample (R, F, M) according to the RFM definition in Section 2.2
2. Normalize the sample instances (R, F, M) on all RDDs of the cluster
3. Execute the k-means algorithm on all RDDs for each chromosome
4. $fitness = \rho(K)$ // Update the fitness value of each chromosome
5. Output results or cross or mutate to get a new chromosome
6. Go to step 3

3. Experimental analysis

The experiment mainly analyzes the two aspects of algorithm running speed and effect. Acceleration ratio refers to the ratio of the time calculated by the same algorithm after parallelization to the time calculated by serialization. It is a commonly used indicator to verify the performance of parallel computing. The higher the acceleration ratio, the less the relative time consumed by parallel computing, and the higher the parallel efficiency and performance. The experimental data is based on the customer data of China's high-speed trains in January 2017 using A to represent it, the customer data in the first quarter of 2017 (January to March) using B to represent it, and the customer data in the whole year of 2017, using C to represent it.

Experiment with a single machine on the Spark cluster to obtain the acceleration ratio of each data set. As shown in Figure 3: It can be clearly seen from Figure 3 that the data set A, that is, the cluster acceleration of customer data in January 2017 is not obvious, only a little higher than 1, and the acceleration ratio of data set B and data set C have a significant improvement, and as the number of computing nodes increases, the acceleration ratio becomes linear. The data set C obviously has a larger slope than the data set B, which indicates that as the computing node increases, the data set of the larger data volume accelerates more. When the data volume is smaller, the acceleration ratio is not obvious. Therefore, it is necessary to use the genetic k-means clustering for the whole year of 2017 or quarterly customer value data, it is suitable for parallel operation on the Spark platform. When the customer value data within one month will be analyzed, only the stand-alone version of the genetic k-means algorithm is needed.

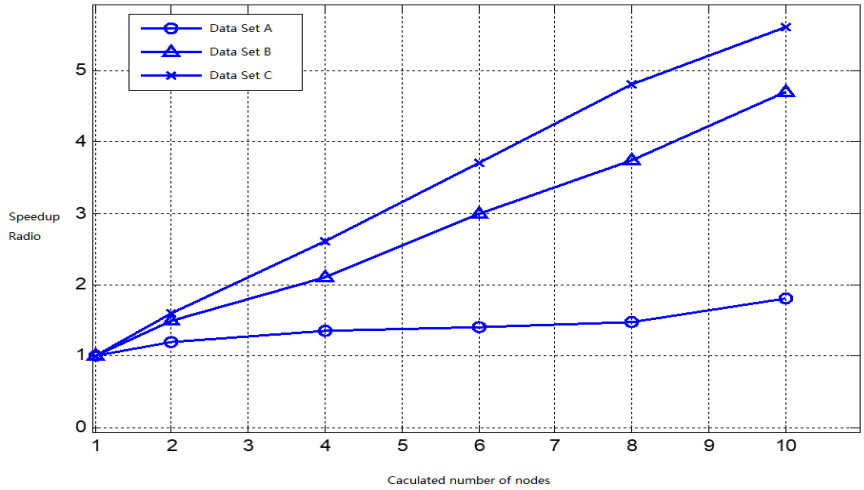

Figure 3. Calculation acceleration ratio of different data sets in different node numbers

The customer value data for the whole year of 2017 is selected as data set, and the parallel traditional k-means algorithm together with the proposed algorithm (Spark-based genetic k-means algorithm) are used to run on the Spark platform. The clustering effect is shown in Figure 4. As shown, the horizontal axis represents the number of clusters, and the vertical axis represents the fitness function value defined in Section 2.3.2. As can be seen from Figure 4, the Spark-based genetic clustering algorithm has higher fitness values than the traditional parallel k-means algorithm when the number of clusters is set to 1-18, indicating that the clustering effect is significantly better than the traditional parallel k-means algorithm.

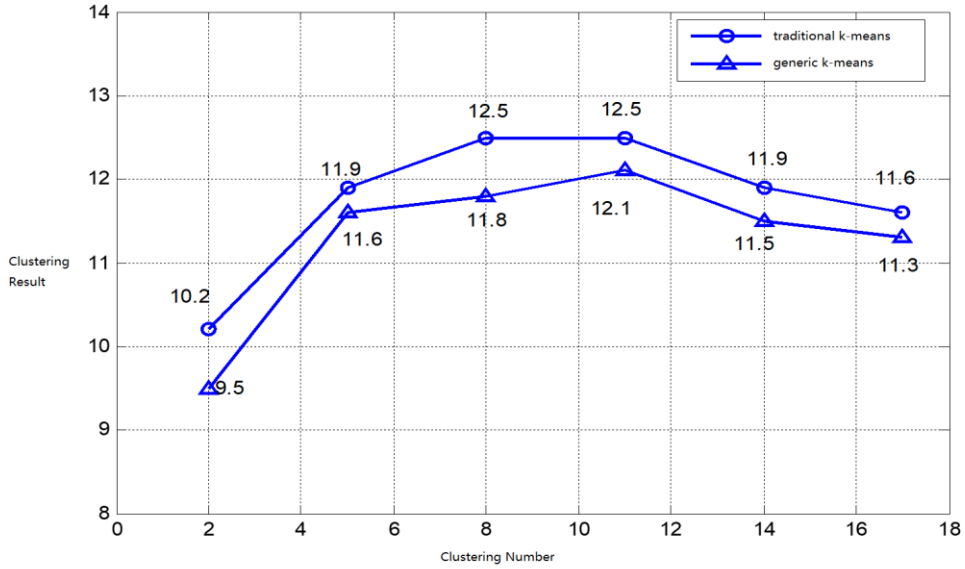

Figure 4. Comparison of traditional k-means algorithm and clustering effect of the algorithm

In the genetic clustering algorithm, two important parameters, chromosome length and genetic algebra, are needed to determine the number of clusters and the iterative termination conditions of the algorithm. In this paper, the 2017 high-speed train customer data is uses for genetic clustering algorithm experiment, as shown in Figure 5. As shown, the vertical axis represents cluster quality and the horizontal axis represents genetic algebra 5 to 35. Each of the curves in Figure 5 represents chromosomes of different lengths. When the maximum genetic algebra is set to increase, the clustering

quality also increases. However, when the maximum number of genetic algebras becomes greater than 25, the cluster quality grows slowly. When the genetic algebra is greater than 35, there is no significant difference in the clustering quality of chromosomes of different lengths.

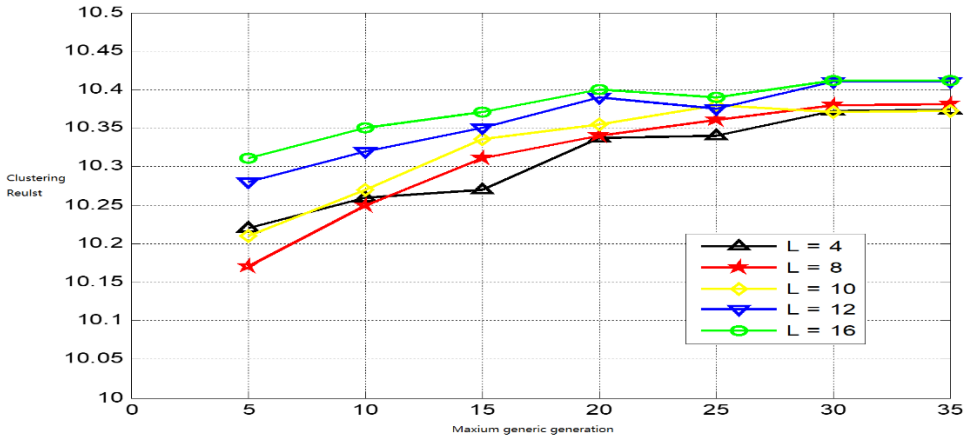

Figure 5. Clustering effect of the algorithm in different genetic algebra settings

According to the above experimental results, the parameter of the maximum genetic algebra in the algorithm can be set between 25 and 35.

4. Conclusion

The establishment of the railway passenger transport customer relationship management system, the implementation of the railway frequent flyer program, and the implementation of the member points and reward system are important measures to improve the loyalty of railway passengers, attract and stabilize customer resources, and enhance the overall competitiveness of the railway. Among them, the rational and efficient classification of high-speed train customer value is a key step to achieve the above system. Aiming at the problem of large customer base group of high-speed train and large amount of initial sample data, this paper proposes a genetic k-means clustering algorithm based on Spark, which not only improves the computational efficiency but also improves the clustering compared with the traditional k-means algorithm. The effect is to provide support for the railway to establish a passenger customer relationship management system, which has important practical significance.

Acknowledgment

Fund Project: Advanced Rail Transit Project of National Key R&D Program (2018YFB1201404)
Fund Project: Science and Technology Research and Development Program of China Railway Corporation (2017X004-C)

References

[1] Hughes. A.M. Strategic Database Marketing [M]. Chieago. IL: Probus Publishing ComPany. 1994:75-80.
[2] D.Sculley.Web Scale K-Means Clustering Proceedings of the 19th international conference on World wide web (2010).
[3] Buckinx W, Van den Poel, D. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting [J]. European Journal of Operational Research, 2005, 164 (1): 252-268.
[4] Zhao Xiaotong, Huang Xiaoyuan, Sun Fuquan. Optimization Model of Promotional Portfolio Strategy Based on RFM Analysis [J]. China Management Science, 2005, 13 (1): 60-64.
[5] Xia Junxi et al., Spark Data Processing Technology [M], Beijing, Publishing House of Electronics Industry, P15-19