Postprint: Railway Sporadic General Cargo Customer Churn Prediction Based on Parallel C4.5

Authors: Zhang Bin, Peng Qiyuan, Liu Fanxiao

Date: 2018-05-18T00:00:00+00:00

Abstract

To improve the accuracy and efficiency of customer churn prediction for railway less-than-carload general cargo customers, a customer churn identification method based on the CDL model is proposed according to the churn characteristics of such customers. On this basis, to address the problem of large data volumes, a C4.5 decision tree customer churn prediction model based on the Hadoop parallel framework is proposed. Simulation experiments demonstrate that the model exhibits favorable accuracy and predictive capability, and that as sample sizes increase, the efficiency of the Hadoop parallel framework is significantly improved without affecting the accuracy and predictive capability of the customer churn prediction model.

Full Text

Preamble

Title: Research on Railway Scattered Freight Customer Churn Prediction Based on Parallel C4.5 Decision Tree Algorithm

Authors: Zhang Bin, Peng Qiyuan, Liu Fanxiao (School of Transportation & Logistics, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: To improve the accuracy and efficiency of customer churn prediction for railway scattered freight, this paper proposes a customer churn identification method based on the CDL model according to the loss characteristics of railway scattered freight customers. Furthermore, to address the challenge of large data volumes, a C4.5 decision tree customer churn prediction model based on the Hadoop parallel framework is developed. Simulation experiments demonstrate that the model achieves good accuracy and predictive capability, and as sample size increases, the efficiency of the Hadoop parallel framework improves significantly without compromising the accuracy or predictive power of the customer churn prediction model.

Keywords: railway transportation; scattered freight; customer churn; C4.5 decision tree; parallel; Hadoop

0 Introduction

With the rapid development of the global economy and the deepening advancement of national supply-side reforms, the "Belt and Road" development strategy, and the orderly adjustment of economic structure, freight transportation market demands have undergone significant changes. The market has gradually shifted from focusing primarily on bulk cargo transportation to scattered freight transportation, with transportation organization patterns evolving from train formation plans to customer demand-oriented models. However, due to deficiencies in railway freight regarding timeliness and convenience, coupled with the continuous growth of alternative transportation modes such as highways and aviation, the railway scattered freight market faces fierce competition. Since 2005, the market share of railway scattered freight has been declining year by year, severely impacting the railway freight market' s position and revenue [1]. According to incomplete statistics, 80% of domestic express transportation currently uses highways, 15% uses aviation, and only 5% uses railways [2].

Retaining customers is key to ensuring core competitiveness in the railway freight industry [3], as acquiring a new customer costs 5-6 times more than retaining an existing one [4-5]. In this challenging environment for railway scattered freight, effective customer management to identify potentially churning freight customers and develop retention strategies is crucial for maintaining competitiveness.

Current research on customer churn prediction primarily employs statistical analysis and artificial intelligence methods [6], with widely used algorithms including Logistic regression [7], artificial neural networks [8], decision trees [9], and Support Vector Machines (SVM) [10]. Decision trees are data mining methods that induce learning from training sets, generating tree structures or decision rules from unordered, irregular cases for classification prediction on new datasets. Due to their high accuracy and robust tolerance for noise with strong interpretability [11], decision trees are widely applied in classification, prediction, and rule extraction. The C4.5 decision tree algorithm [12] improves upon the ID3 algorithm [13] by addressing the limitation of information gain bias toward multi-valued attributes. However, decision trees are constructed through iterative computation, which creates computational time and space limitations when facing large-scale data, severely impacting operational efficiency.

Google' s Hadoop distributed open-source computing framework can process massive datasets, providing the MapReduce programming model and Hadoop Distributed File System (HDFS) with fault-tolerant parallel computing capabilities that enable the construction of large clusters for big data processing. Literature [14] designed and implemented a parallel SPRINT classification algorithm based on the Hadoop platform, demonstrating good classification accu-

racy, low time complexity, and favorable parallel performance. Literature [15] proposed a Hadoop-based parallel shared decision tree mining algorithm, proving its good parallelism and scalability. Literature [16] proposed an uncertain probabilistic C4.5 algorithm based on the Hadoop platform, demonstrating its capability to process massive data. Literature [17] proposed an uncertain probabilistic error pruning algorithm based on Hadoop and applied it to the C4.5 algorithm, proving its ability to handle large-scale data and good scalability through MapReduce programming.

This paper extracts freight characteristics of scattered freight customers to establish a churn identification method and addresses the large volume of railway freight data by employing the C4.5 decision tree algorithm to propose a scattered freight customer churn prediction model based on the Hadoop distributed parallel architecture. Simulation experiments verify the high efficiency of the parallel algorithm and the effectiveness of the prediction model.

1 Scattered Freight Customer Churn Prediction Model Construction

1.1 Identification Method for Scattered Freight Customer Churn

Compared with bulk commodities, scattered freight commands higher prices, making it a premium product in the railway freight market. Additionally, scattered freight customers are more sensitive to market services and dynamics, exhibiting greater flexibility. Therefore, extracting freight customer churn features to judge the churn status of scattered freight customers represents an important challenge for churn prediction. This paper combines scattered freight transportation characteristics to assess customer churn propensity from three dimensions: delivery time compliance, cargo damage and discrepancy rates, and service quality.

Delivery time compliance reflects transportation time fulfillment rates. Since railway freight involves multiple operational stages—dispatch, en-route transportation, en-route decoupling, and arrival—each with numerous operations that often affect delivery times, whether delivery deadlines are met significantly impacts customer trust in railway transportation. Cargo damage and discrepancy rates are crucial factors in measuring scattered freight customer satisfaction. Unlike bulk cargo transportation, scattered freight customers impose higher requirements on cargo integrity and packaging integrity. Service quality reflects customer perception and experience during railway freight business processes, which can be gleaned from customer complaints and feedback.

Based on these scattered freight transportation characteristics, this paper proposes the CDL (Customer Defection Likelihood) model for scattered freight customer churn identification. In this model, represents the number of complaints within the observation window, denotes the delay hours for a single shipment, and represents the cargo damage and discrepancy rate for a single shipment.

represents the average delay time, average cargo damage and discrepancy rate, and average customer complaint rate within the observation window.

For customer churn identification using the CDL model, the Analytic Hierarchy Process (AHP) combined with the Delphi method is employed to assign weights to each indicator, yielding the calculation method for the scattered freight churn factor based on the CDL model as follows:

MATH_0

where represents the churn factor of the th customer based on the CDL model; , , and represent the weights of the , , and parameters; and , , and represent the standardized values of , , and for the th customer. This paper employs the Min-max standardization method to map each parameter' s standardized value to the [0,1] interval using the following method:

MATH_1

where is the standardized value of the th parameter for the th customer.

Based on the above analysis, this paper defines the scattered freight customer churn identification method as follows:

Definition 1: The churn customers discussed in this paper refer to scattered freight customers with churn propensity (i.e., potential churners). Customers who have not conducted business for an extended period are considered already churned and are not within the research scope of this paper.

Definition 2: This paper identifies scattered freight churn customers based on the churn factor from the CDL model and the standardized model parameters , , and . The identification method is shown in Equation (4).

If a customer' s , , or values in the CDL model exceed given thresholds, the customer is identified as churning. For customers not exceeding thresholds, if the churn factor exceeds a given threshold, the customer is identified as churning. Churned customers are labeled as 1, and non-churned customers as 0.

1.2 Railway Scattered Freight Customer Churn Prediction Model

In the scattered freight customer churn identification method, this paper identifies churn from delivery time compliance, cargo damage/discrepancy, and service quality dimensions, but cannot predict customers with churn propensity. This section combines freight customer characteristics to predict scattered freight churn customers using a parallel C4.5 decision tree model, examining four features within the observation window: customer registration duration ( ), customer shipping frequency ( ), recent shipping performance ( ), and customer shipping turnover ( ).

1.2.1 C4.5 Decision Tree The C4.5 decision tree approach determines the tree structure from root to leaf nodes by calculating the maximum information

gain ratio of variable attributes. The attribute with the maximum information gain ratio becomes the root node, with each leaf node representing a class decision rule. The key to determining the decision tree is calculating each variable attribute' s maximum information gain ratio.

First, the information entropy of the training sample must be calculated, expressed as:

MATH_2

where is the training dataset and represents the proportion of class in . If the training dataset is partitioned by attribute , the uncertainty of given attribute is expressed as:

MATH_3

where attribute divides dataset into classes, . By calculating the difference before and after partitioning, the information gain is obtained:

MATH_4

To compensate for information gain' s bias toward multi-valued attributes, C4.5 uses information gain ratio to overcome this deficiency:

MATH_5

where the split information is . C4.5 selects the attribute with the maximum information gain ratio to construct the decision tree from top to bottom.

1.2.2 Parallel C4.5 Decision Tree Customer Churn Prediction Model Based on Hadoop Constructing a decision tree is an iterative process. Faced with large-scale railway freight scattered freight customer information, serial computation wastes substantial resources in terms of computational time and space. This paper establishes a parallel C4.5 decision tree customer churn prediction model based on the Hadoop distributed platform using the MapReduce computing framework and distributed file system HDFS. The operational steps include data source integration and loading, data preprocessing, churn customer identification, and churn customer prediction, as shown in [Figure 1: see original paper].

a) **Data Source Integration:** Integrate customer data sources, including personal information and shipping information, and load them into HDFS to convert from multiple data sources to a single data source.

b) **Data Preprocessing:** Customer data information is extracted from HDFS and split into multiple Splits. MapReduce uses JobTracker to assign Splits to idle TrackTrackers, which then allocate them to Map and Reduce subtasks. Map receives data in <key,value> structure and performs data cleaning operations on customer information, including filtering duplicate data, removing illegal data, filtering irrelevant data, and handling incomplete and abnormal data. The cleaned data is then

passed to Reduce subtasks, which merge values (customer information) with the same key (customer ID), calculate customer shipping frequency , perform Min-max standardization on 各项数据, and finally return the processed data to HDFS.

c) **Customer Churn Identification:** Perform customer churn identification on sample data based on the CDL model. Customer sample data is extracted from HDFS, with Map subtasks sending data in <customer ID, CDL model parameters> structure to Reduce subtasks. Reduce merges data by customer ID as the key, aggregating multiple shipment records within the observation window for each customer to calculate and . Based on Definition 2, it identifies and labels customer churn status, then returns the processed customer information to HDFS.

d) **Churn Customer Prediction:** This component consists of two MapReduce processes. In the first MapReduce process, Map1 subtasks input data in <attribute name, (attribute value, class, primary key ID)> structure, where attribute names mainly consist of and churn status labels. Map1 sends data to Reduce1, which, since is a continuous attribute, performs k-means clustering discretization on attribute values (this paper sets) and counts classes. In the second MapReduce process, Map2 reads data as <(attribute name, class), (attribute value, primary key ID, class count)>, and Reduce2 calculates the information entropy and information gain of each attribute, selects the attribute with maximum information gain as the best splitting attribute, and progressively determines each decision tree node to finally complete decision tree construction.

2.3 Simulation Results and Analysis

To verify that the C4.5 decision tree achieves high computational efficiency and prediction effectiveness for railway scattered freight customer churn prediction, and to demonstrate the efficiency of the Hadoop-based parallel algorithm, three simulation experiments were designed.

The confusion matrix reflects model prediction effectiveness and forms the basis for model evaluation metrics [18]. The customer churn model prediction results confusion matrix is shown in , displaying classification results across real and predicted dimensions.

Experiment 1: Under single-node operation, compare the execution efficiency of C4.5, Logistic, and BP algorithms. Different sample data quantities were extracted from the simulation data, as shown in . The three algorithms were applied to calculate the sample data. shows the runtime of the three algorithms under different sample sizes, revealing that while runtimes are similar, C4.5 slightly outperforms the other two algorithms in speed, indicating good computational efficiency. shows the accuracy and coverage of customer churn prediction models obtained using the three algorithms combined with Equations (9)-(11). The results demonstrate that C4.5 algorithm exhibits advantages over

the other two algorithms in both accuracy and coverage across different samples, proving its superior prediction effectiveness for railway scattered freight customer churn prediction.

Experiment 2: Under the Hadoop platform, compare operation with different numbers of service nodes. shows the runtime required to run different sample quantities on varying numbers of service nodes. The table reveals that with small sample quantities, the difference between single-machine and parallel modes is minimal. However, as sample quantities increase, the efficiency of Hadoop-based parallel computation improves substantially, and increasing service nodes significantly reduces computation time as sample size grows. [Figure 2: see original paper] shows the speedup curves for different sample sizes across different node counts, where speedup (with representing single-machine runtime and representing multi-node parallel runtime) is an important parameter for measuring parallel algorithms [19]. The figure shows that speedup changes are not obvious with small data volumes but increases substantially with larger sample data, again demonstrating the advantages of Hadoop-based parallel algorithms for big data processing.

Experiment 3: Using the parallel C4.5 decision tree customer churn prediction model to evaluate the simulation data, the results are shown in . Parallel experiments with different numbers of service nodes show the model performs well in accuracy, hit rate, coverage, and lift coefficient, indicating strong predictive capability. Moreover, the Hadoop-based parallel C4.5 customer prediction model shows minimal differences across different service node configurations, suggesting that varying node counts have little impact on model accuracy and predictive capability but substantially affect runtime speed.

Applying the parallel C4.5 decision tree to predict customer churn on simulation data, the final decision tree reveals that customer average shipping frequency has the greatest impact on scattered freight customer churn, serving as the root node. Customers with standardized average shipping frequency less than 0.21 are identified as churning, those greater than 0.73 as non-churning, and others proceed to the second-level branch node. The second-level branch node is customer average turnover: customers with values greater than 0.65 are non-churning, those less than 0.13 are churning, and others proceed to the third-level branch node representing recent shipping performance. Customers with recent shipping performance greater than 0.82 are churning, those less than 0.09 are non-churning, and others proceed to the fourth-level branch node representing customer registration duration. Customers with registration duration greater than 0.75 are non-churning, and the remainder are churning. This analysis shows that factors affecting churn prediction can be ranked by importance as: customer average shipping frequency, average shipping turnover, recent shipping performance, and registration duration. Customers who ship frequently exhibit stronger stability, while long registration history does not necessarily indicate lower churn probability.

3 Conclusion

This paper addresses scattered freight customer churn in railway operations based on customer characteristics. By establishing a CDL model for customer churn identification and calculating churn factors, the paper defines a method for identifying scattered freight customer churn. Subsequently, using big data technology, a customer churn prediction model based on the Hadoop distributed platform and C4.5 decision tree was developed. Simulation using the MapReduce computing framework and HDFS demonstrated that the C4.5 algorithm achieves higher computational efficiency and accuracy for railway scattered freight customer churn prediction compared to Logistic and BP algorithms. The Hadoop-based parallel computation method significantly improves algorithm efficiency without affecting the accuracy or predictive capability of the churn prediction model, proving highly valuable for large-volume test samples. This approach can effectively guide railway freight departments in predicting scattered freight customer churn, enabling targeted retention strategies to increase railway freight volume and revenue.

References

[1] Wang Zhimei, Zhang Xingchen, Xu Bin. Integration of cargo organization and transportation organization issues for scattered freight [J]. Journal of Beijing Jiaotong University, 2016, 40(6): 43-49+56.

[2] Zhang Bomin. Thoughts on railway expansion from bulk cargo to scattered express freight under supply-side reform [J]. Journal of Transportation Engineering and Information, 2016, 14(4): 36-40.

[3] Zhou Xinjun. Theory and practice of introducing customer relationship management into railway freight services [J]. Railway Freight, 2008, 26(12): 25-28.

[4] Athanassopoulos A D. Customer satisfaction cues to support market segmentation and explain switching behavior [J]. Journal of Business Research, 2000, 47(3): 191-207.

[5] Bhattacharya C B. When customers are members: customer retention in paid membership contexts [J]. Journal of the Academy of Marketing Science, 1998, 26(1): 31-44.

[6] Xia Guoen, Jin Weidong. Customer churn prediction model based on support vector machine [J]. Systems Engineering—Theory & Practice, 2008, 28(1): 71-77.

[7] Chang Chengchang, Gong Dahchuan. A comparison of rohs risk assessment using the logistic regression model and artificial neural network model [C]// Proc of the 9th International Conference on Machine Learning and Cybernetics. 2010.

[8] Yu Lu. Combined forecasting model for telecom customer churn [J]. Journal of Huaqiao University: Natural Science Edition, 2016, 37(5): 637-640.

[9] Ye Zhilong, Huang Zhangshu. Modeling and prediction research on online membership customer churn [J]. Management Modernization, 2016, 36(3): 96-98.

[10] Yu Xiaobing, Lu Yiqun. E-commerce customer churn warning and prediction [J]. Systems Engineering, 2016(9): 37-43.

[11] Zhang Yu, Zhang Zhiming. Research on a customer churn prediction model based on C5.0 decision tree [J]. Statistics & Information Forum, 2015, 30(1): 89-94.

[12] Quinlan J R. C4.5: programs for machine learning [M]. San Francisco: Morgan Kaufmann Publishers, 1993: 17-42.

[13] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[14] Huang Gang, Sun Yuan. Analysis and research of SPRINT algorithm based on Hadoop platform [J]. Journal of Nanjing Normal University: Natural Science Edition, 2016, 39(4): 25-30.

[15] Chen Xiangtao, Zhang Chao, Han Xi. Research on parallel shared decision tree mining algorithm based on Hadoop [J]. Computer Science, 2013, 40(11): 215-221.

[16] Liu Yaqiu, Li Haitao, Jing Weipeng. Implementation of decision tree algorithm for massive noisy data based on Hadoop [J]. Computer Applications, 2015, 35(4): 1143-1147.

[17] Zhang Jingxing, Li Shijun. Improved decision tree pruning algorithm based on Hadoop [J]. Computer Engineering & Design, 2016, 37(7): 1942-1946.

[18] He Benlan. Application research of support vector machine model in bank customer churn prediction [J]. Financial Forum, 2014, 225(9): 70-74.

[19] Lu Qiu, Cheng Xiaohui. Parallelization of decision tree algorithm based on MapReduce [J]. Computer Applications, 2012, 32(9): 2463-2465, 2469.

Note: Figure translations are in progress. See original paper for figures.