Journal of Applied Science and Engineering Innovation, Vol.4 No.4 2017, pp. 113-116
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

Application of Decision Tree Algorithm in Customer Churn Warning of Logistics Enterprises

Yang Jun[1], Li Yuhang[2]

[1] *Guangxi Vocational and technical college of communications, Guangxi, Nanning, China*
[2] *Beijing Logis Technology Development Co., Ltd, Beijing, China*

Abstract*:* With the increasingly fierce competitive environment, the competition among enterprises are gradually transferred to the customer as the center of the competition, while the customer churn phenomenon is a serious problem facing enterprises in the process of development, especially as a service industry of logistics enterprises. In this paper, the decision tree algorithm is used to analyze the third party logistics enterprise customer information, In order to predict how the behavior of the customers will be lost, and provide the basis for the logistics enterprises to take measures to retain customers.

Keywords C4.5 algorithm, Logistics enterprise, Customer churn warning

INTRODUCTION

With the rapid development of the logistics industry, the logistics enterprises are facing fierce competition, It is one of the challenges for logistics enterprises which how to make use of a large number of customer information resources accumulated under the background of big data age for customer relationship management and The enterprise has customer information data into effective information, on the basis of existing customers, to maximize customer satisfaction, to prevent customer churn.

In this article, we use the decision tree algorithm, According to the analysis of the third party logistics enterprise database of customer information, dig out what behavior customer churn intention, improvement strategy for enterprise customer management and management, And then effectively prevent the loss of enterprise customers, improve customer loyalty to the company, for future logistics enterprises to customer management and retention work to provide a scientific basis.

DECISION TREE ALGORITHM

Definition of decision tree

As a predictive model, decision tree is a tree model, which consists of three parts: decision node, branch and leaf node. Which represents a decision node attribute to classification data set, and represents a different attribute in the test set, the test results represent a branch, and branch said different values of a decision node. Each leaf node holds a category label that represents a possible classification result [Yiting et al., 2013]. The decision tree algorithm divides the training set into pure subsets and establishes the decision tree recursively. This paper mainly uses the C4.5 algorithm in the decision tree.

C4.5 algorithm idea

Ross Quinlan, a professor at the University of Sydney in Australia, proposed the C4.5 algorithm in 1993, which was proposed on the basis of the ID3 algorithm [Smith Tsang et al., 2011]. C4.5 algorithm is used to measure the rate of information gain, the information gain rate is a distortion of information entropy, it is in consideration of the split information "price" to partially offset the impact of attribute value the number, so the algorithm is a significant improvement on the classical ID3 algorithm, the algorithm uses the following formula, the first is the definition of information entropy.

Suppose D is a data set, the data set D contains d samples, the class attribute of this data sample can take n different values: then, for these n different categories Ai，i=（1，2，…… n），then the amount of information needed to categorize a given data object is:

$$Entropy(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad （1）$$

Among them, pi（i=1,2,3……n）the frequency of class attributes that appear in category Ai for n class labels.

Set the attribute A to {a1, a2, a3……，am}. Use attribute A to split the sample set D into D1, D2，……，Dm, If the Dm subset is, then the gain ratio of the attribute A is

$$GainRatio(A) = \frac{Gain(A)}{Splitl(A)} \qquad （2）$$

Corresponding Author: Yang Jun, Logistics cost management, Guangxi Vocational and technical college of communications, Nanning 530023, China

Gain(*A*) is the information gain of attribute A, information gain when choosing an algorithm for data sets for classification, classification of the data set information

entropy than before classification, the difference in the middle of that information gain is the denominator A attribute information division.

C4.5 algorithm can not only deal with discrete attributes, but also deal with continuous attributes, the basic idea is to discretize the values of the continuous attributes [Ma yue et al., 2009].The information gain rate is chosen as the testing attribute of the C4.5 algorithm, and the classification information is used as the denominator. The greater the number of attributes, the greater the value of splitting information, thus partially offsetting the influence of the number of attributes [Mengshangmin et al, .2014].

EMPIRICAL RESEARCH

Data preparation

The experimental data comes from the customer information data of a third party logistics enterprise, which contains part of the customer information of the enterprise, and contains 41138 customers' desensitization information（Remove client's name, address, contact, etc. from private information）。In this article, The 16 attributes which contains customer account number, singular number, and business volume, are used as the classification attributes of decision tree algorithm, and the distribution of attributes of some customers is shown in the following table1:

Table 1 Partial customer attribute information

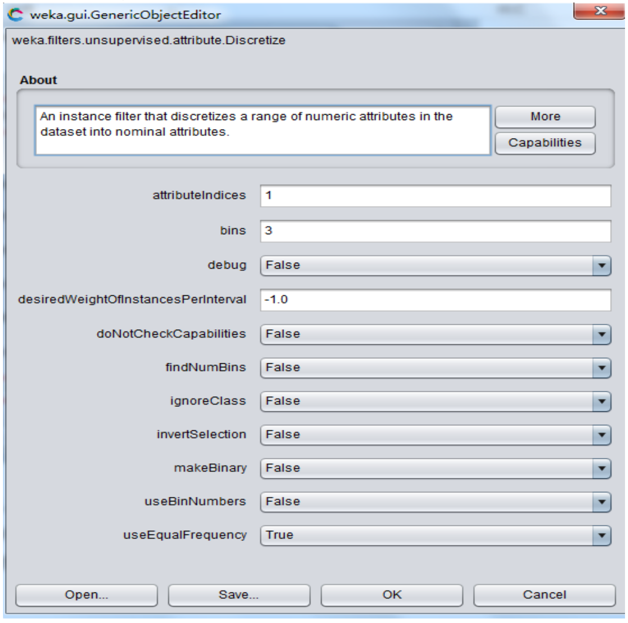
ID	waybill	volume of busi	volunme	weigh	tages	revenue	type	start city	end city	……	calss
1	344	6880	27.52	6880	Heavy cargo	3199.2	Express Service	shanghai	changsha		No loss
2	420	8400	33.6	8400	Heavy cargo	3738	Express Service	shanghai	zhengzou		No loss
3	65	1300	5.2	1300	Heavy cargo	604.5	Express Service	shanghai	changsha		No loss
4	489	9780	39.12	9780	Heavy cargo	2934	Express Service	shanghai	nanjing		No loss
5	107	2140	8.56	2140	Heavy cargo	1177	Express Service	shanghai	tangshan		No loss
6	155	3100	12.4	3100	Heavy cargo	1379.5	Express Service	shanghai	zhengzou		No loss
7	553	11060	44.24	11060	Heavy cargo	3318	Express Service	shanghai	nanjing		No loss
8	347	6940	27.76	6940	Heavy cargo	3088.3	Express Service	shanghai	zhengzou		No loss
9	184	3680	14.72	3680	Heavy cargo	1104	Express Service	shanghai	nanjing		No loss

Data cleaning

Data cleaning in data mining refers to the discovery of errors in data files, dealing with invalid values and missing values in data [Xuehua et al, .2012].the steps of the purpose is to remove the information data of the enterprise customer concentration does not meet the requirements as well as data not related

1)By analyzing the customer churn, what are the attributes of the customer? Therefore, the irrelevant attributes are deleted, the singular number, business volume, volume, billing weight, bubble mark, service type, settlement method, income and loss are taken as the final classification attributes

2)Classification algorithms require data attributes to be nominal data types, so numerical data need to be discretized. We divide the odd order, the quantity of business and the volume attribute into the boxes and divide them into 3 sections, as shown in figure 1:



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	1
bins	3
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
useBinNumbers	False
useEqualFrequency	True

Figure 1 parameter settings

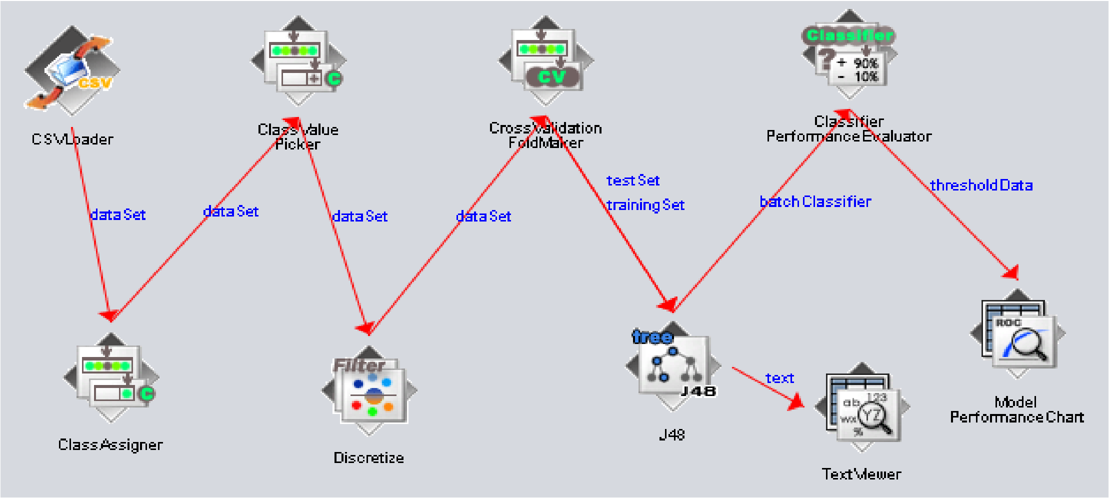


Figure 2 decision tree algorithm model

RESULT ANALYSIS AND APPLICATION

Result analysis

We click on the classifier function panel and select the J48 under the decision tree, that is, the C4.5 algorithm. In the parameter setting, we select the default setting, the training method, select "use training set training", "loss situation" as the classification target, and the results are as follows:

Table2 Summary

	Correctly Classified	Incorrectly Classified	Total Number of Instances
Instance	39068	1789	40857
Rate	95.6213%	4.3787%	100%

Table 3 Experimental result

Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
0.1181	0.0832	0.204	93.4711%	96.66945%

Table 4 Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1. 000	0. 934	0. 956	1. 000	0. 978	0. 244	0. 588	0. 962	NO LOSS
	0. 066	0. 000	0. 947	0. 066	0. 123	0. 244	0. 588	0. 130	LOSS
Weighted Avg	0. 956	0. 890	0. 956	0. 956	0. 938	0. 244	0. 588	0. 923	

Table 5 Confusion Matrix

a	b	Classified as
38942	7	a=NO LOSS
1782	126	b=LOSS

The results can be seen from the table2, examples of correct classification is 95.6%, from the table 3, we can see that the average absolute error is 0.0832, the root mean square error is 0.204, that of decision tree classification results accord with the ideal standard, better classification results, in the confusion matrix in all 38949 without loss of customers in the 38942 right, 7 forecast error.

Because there are many decision nodes in decision tree view, this paper describes the attributes of lost customers in terms of the results of text. According to the visual decision tree diagram, we can derive the loss and business volume of customers, such as billing weight, singular number, volume of goods, bubble mark, and so on. When the customer is a small enterprise, the enterprise needs to transport goods less light and less business, this kind of customer is easy to drain, And some large and medium-sized enterprises, their speed of goods is higher, this type of customers in this enterprise is easy to drain. In view of this situation, the corresponding countermeasures and suggestions are put forward for the future development of the logistics enterprise.

Suggestions

Through the analysis of the results, we put forward the following suggestions on the customer churn warning analysis of the customer relationship management in the following aspects

(1)Enterprises establish a complete set of sales incentive system for customers, Through the analysis of customer information, the enterprise customers are divided into different levels, different levels of management, and formulate different sales strategies, For example, the company adopts some preferential policies to some small customers to increase their satisfaction and attention to the company.

(2)The company establishes a full range of customer communication systems, for some key customers and key customers, logistics enterprises should pay a return visit and keep in touch with each other, and then find out their potential needs, and can provide them with timely solutions to improve customer loyalty to the company, to prevent customer churn.

CONCLUSION

Through the use of decision tree algorithm in data mining, combined with large data mining and analysis software, the customer churn of logistics enterprises is analyzed, From a large number of customer information in the enterprise, customers with loss tendency are found, which provides a strong support for the maintenance and retention of customer relations in logistics enterprises in the future, In the future work, how to further improve the forecast value of customer churn warning of logistics enterprises through other methods is the key work to be studied in the next step.

REFERENCES

Ma Yue, 2009, “Application of data mining decision tree method in rapid freight transportation”.

Mengshangmin. Analysis of customer churn in logistics enterprises and Countermeasures, 2014, “Enterprise technology development” Nol.30, pp25-27.

Smith Tsang,Ben Kao,Kevin Y.Decision, 2011, “Trees For Uncertain Data”,IEEE transactions on knowledge and data engineering, Nol.1, pp 64- 78.

Xuehua, 2012, “Research on Application of data mining in logistics customer relationship management”.

Yin Ting, Tan Xizhong, Jia Zhenhong, Ma Jun, 2013, “Research on customer churn prediction based on WEKA”, Journal of lasers, Nol.5, pp44-46.