Research on the Prediction of Customers’ Arrears Under Big Data Background

Xiyi Zhao, Jiajun Lang, Qian Zeng, Qingan Hu, Xingfu Shen, Fei Lyu and Jun Wu

ABSTRACT

This paper proposed a new prediction method to predict customers’ arrears by means of multivariable linear regression model and convolutional neural network (CNN) on Hadoop big data platform with mass of customers’ power consumption. The proposed method made the association between different type of customers’ data and related traits with customers’ arrear, and estimated the probability of customers’ arrear based on evaluated customers’ credit level and prediction of customers’ arrear. The application results illustrated that the method is available.

INTRODUCTION

With the development of informatization in power enterprises and construction of smart grid, the data resources grow rapidly. Meanwhile, electricity enterprises have accumulated mass data, including customer data from 95598, marketing system data, metering system data, etc. The white paper on the development of big data on China electric [1] pointed out that these data is of great value for guiding the electricity market consumption, managing customer relationship, optimizing the electrical power production. Under the situation, how to min these data resource to obtain the useful information becomes one of hot issues in electric power industry.

1Xiyi Zhao, Jiajun Lang, Qian Zeng, Qingan Hu, Duyun Power Supply Bureau, Guizhou Power Grid Co. Ltd, Duyun, Guizhou Province, China
Xingfu Shen, Fei Lyu, Jun Wu, Guizhou Qianchi Information Co., Ltd, Guizhou, Guiyang, China
In recent years, some researches have been done in this field. C3 Energy, USA, realized real-time monitoring of power grid and data analysis through performing deep analysis technologies on the integrated 22 kinds of data from public service companies and other third parties. Opower, USA, studied how to provide consumers’ energy saving programs to electricity enterprises based on the consumers’ energy consumption data.

In our country, there are also many unsolved critical issues in electric marketing management. Due to the lack of credit control, the business decisions had to be made based on the subjective judgment. The power supply enterprises cannot recover the electricity charges in time now and then. On the other hand, customers need the power supply enterprises to provide more precise and personalized service. The key to solving the problems is predicting the probability of customers’ arrears according to their credit level, which is contained in the massive data about customers’ power usage condition.

**TECHNOLOGIES REVIEW**

**Hadoop Big Data Platform**

Apache Hadoop [2] is an open-source software framework for reliable, scalable and distribute computing. It mainly includes MapReduce distributed architecture, Hadoop Distributed File System (HDFS), Zookeeper and HBase. MapReduce is responsible for splitting the input data set across several independent chunks, scheduling the task execution as monitoring and re-executing the failed tasks along with managing the required inter-machine communications. HDFS is designed for scalability and fault tolerance, which breaks the received files into items and stores them redundantly on three or more servers. Zookeeper is used to coordinate these distributed applications and manage the related configurations. HBase is a column-oriented NoSQL database, which is one of the non-relational databases.

**Multivariable Linear Regression**

Multivariable linear regression studies how to explain the relationship between multiple dependent variable set Y and multiple independent variable set X with linear polynomial model. Assuming \( Y = [y_1, y_2, \ldots, y_n]^T \) and \( X = [x_1, x_2, \ldots, x_m]^T \), then the regression model can be described as follows,

\[
y_i = [\beta_{i1} \beta_{i2} \ldots \beta_{im}] [x_1^T \ x_2^T \ldots \ x_m^T]^T + e_i (i = 1,2,\ldots, n)  \tag{1}
\]

Where \([\beta_{i1} \beta_{i2} \ldots \beta_{im}]\) is the regression parameter vector of \(y_i\) and \(e_i \sim (0, \sigma^2)\) is the regression error of \(y_i\). Least-squares regression (LSR) or partial least squares (PLS) can be used to estimate the parameters depending on whether the variables in X are linear independent.
Convolutional Neural Network

CNN is a kind of supervised machine learning algorithm inspired by the visual system structure [3]. Its first few stages are composed of a set of pairs of convolution and pooling layers that enable the model to extract and enhance implicit feature of the data set. The convolution layers convolute the input from the previous layer through a set of weights and compose a feature output. Pooling layer follows convolution layer to obtain a lower resolution representation of the convolution layer activations through sub-sampling. The last stage of CNN consists of a fully-connected layer which is similar to a traditional multilayer neural network and can be applied through different classification models. CNN uses forward path to mimic human visual neurons and process input data, and backward path for training.

PREDICTION OF CUSTOMERS’ ARREARS PROBABILITY

The probability of customer’s arrear has relationship with not only the status of customer’s arrear, but also the status of customer’s power consumption, the customer’s default and customer’s electricity theft. In this section, the model of customer’s power consumption, the model of customer’s arrear, the model of customer’s default probability, the model of customer’s electricity theft are firstly set up based on the treated historical data from Hadoop big data platform using CNN to obtain the third-level credibility indicators on customer’s power consumption, arrear, default, electricity theft in past and future. Based on them, the weight model of the second-level credibility indicators, covering commercial credit, behavior legal credit, economic legal credit and safety credit, can be respectively gotten by means of multivariable linear regression technologies. And these second-level credibility indicators can be changed depending on different situations. Then, the first-level credibility indicator is estimated based on the second-level credibility indicator.

According to the customers’ credit level evaluation system, the model for the predicting of customers’ consumption and rarer probability can be set up. The modeling process is as follows: (1) preprocessing the data of residents and non-residents; (2) selecting the training data set and testing data set in different proportions; (3) modeling the training data set in random forest method to obtain the alarm model of customers’ arrear; (4) demonstrating the model using testing data set; (5) predicting probability of customers’ arrear based on the model in step (3) and evaluated customers’ credit level.

APPLICATION

The proposed method was applied to predict the probability of the enterprises serviced by Duyun power supply bureau, GuiZhou Province. Besides the basic data
related to commercial credit, behavior legal credit, economic legal credit and safety credit, some external data which related to climate, socio-economic condition and finance condition. Based on these data, the customers’ credit evaluation system was set up and the customers’ credit score and grade were shown in Fig 1. Based on the results, the probabilities of the customers were predicted and shown in Fig 2. And the results were demonstrated to be fairly near the fact.

CONCLUSIONS

In this paper, a new prediction method to predict customers’ arrears is proposed based on Hadoop big data platform. This method set up a assessment system of customers credit rating by mean of CNN and multivariate linear regression, and made the association between different type of customers’ data and related traits with customers’ arrear through using CNN. This method is demonstrated to be available by a application of Duyun power supply bureau. This method will have a broad application prospect in industrial big data Era.

<table>
<thead>
<tr>
<th>No.</th>
<th>Enterprise name</th>
<th>Type</th>
<th>Industry</th>
<th>Address</th>
<th>Score</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>0827077010057666</td>
<td>100</td>
<td>C2620</td>
<td>66666666</td>
<td>61</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>08270770100427105</td>
<td>100</td>
<td>C4810</td>
<td>66666666</td>
<td>61</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>08270770500956666</td>
<td>100</td>
<td>C1210</td>
<td>66666666</td>
<td>61</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>08270270104427272</td>
<td>100</td>
<td>B0020</td>
<td>66666666</td>
<td>59</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09270770509981017</td>
<td>100</td>
<td>C1210</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0927077050985356</td>
<td>100</td>
<td>C1210</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09270770509516601</td>
<td>100</td>
<td>B1100</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09270770509744285</td>
<td>100</td>
<td>F5100</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09270770509744222</td>
<td>100</td>
<td>F5100</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09270770509128122</td>
<td>100</td>
<td>C4100</td>
<td>66666666</td>
<td>57</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. The evaluation of customers’ credit score and grade.
Figure 2 The probabilities of the customers.

REFERENCES