Pisco Med Publishing

ORIGINAL ARTICLE

Research on Telecom Fraud Detection Model Based on Cellular Network Data

Kaiyuan Guo, Wenbo Wang

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876

Abstract: With the rapid development of wireless communication technology, the use of mobile phones and other means of communication for telecommunications fraud has become a major problem that endangers user security. Aiming at this problem, this paper constructs a telecom fraud user detection model by in-depth analysis and mining of cellular network data. The model includes data processing, CNNcombine algorithm and model evaluation. First, in the data processing part, the data set is subjected to feature screening, coding, sampling, and the like. Secondly, the CNNcombine algorithm is a combination of a one-dimensional convolutional neural network and multiple traditional classification algorithms. The convolutional neural network is applied to solve classification problems other than text image signals. Finally, in the model evaluation part, it is proved that the CNNcombine algorithm has higher accuracy than the common machine learning classification algorithm such as XGBoost to detect telecom fraud users.

Keywords: Machine Learning; Cellular Network Data; Deep Learning; Classification Algorithm

Introduction

Telecommunications fraud refers to the criminal act of criminals sending false information, setting up fraud schemes, and conducting long-distance non-contact fraud on the victims by means of telephone voice, short messages and other means to induce the victims to pay or transfer money to the criminals, which affects the social security and stability for a long time. Telecom operators have a large amount of data, the data scale is far larger than that of other industries data, this paper focuses on how to analyze and model cellular network data, and use this information to realize the detection of telecom fraud.

1. Detection model framework and data preprocessing

1.1 Fraud user detection model framework

The research point of this paper is to build a detection model of telecom fraud based on the measured data of cellular network provided by operators, and analyze electricity letter cheats the user's characteristic, finds the telecommunication cheats the user, and promotes the restriction measure regarding the telecommunication cheats the user.

1.2 Data preprocessing

1.2.1 Data set introduction

The cellular network data used in this paper are collected from a telecom operator in a city in Guangdong Province. Data has been desensitized, ODS.

doi:10.18282/jnt.v2i1.835

Package information: package information refers to package information handled by the user, including package price, package type and package name, totaling 3 characteristics.

Terminal information: information based on the user's terminal model, including terminal model, terminal category, terminal brand, etc. with a total of 9 features.

Internet surfing behavior: the usage of traffic collected from users, including daily average traffic, and the proportion of active days of traffic totaling 6 features.

Call behavior: including daily average calling times, daily average calling duration, calling duration 5 seconds, 5 seconds -15 seconds accounting for all the main time periods

The proportion of calls and the proportion of calling calls in each time period are based on the characteristics of the user's call behavior, totaling 33.

Short message behavior: The number of short message objects and the number of up-and-down messages are recorded with 5 characteristics.

Base station data: the average number of calling base stations per month, the average number of cities in which the called are located per month, and the proportion of calls made in areas with high fraud incidence, etc.

1.2.2 Feature selection

The purpose of feature selection is to screen the model entry indexes and to reduce the training complexity. The main basis for feature selection based on the features of samples is as follows.

(1) Features are less important

Percentage of missing values: attribute missing values exceeding 50% are rejected as invalid features. Category Proportion in Typed Features: Calculate the proportion of category values in the total number of typed variables; if the proportion exceeds 50%, the feature is regarded as unimportant and the feature is eliminated. Proportion of categories in classification features: calculate the proportion of each category in classification variables to the total number; If the percentage of categories is greater than 90%, the feature will be regarded as unimportant and the feature will be eliminated. (2) Whether the characteristic variable is independent of the label variable: For classification problems, features independent of labels are irrelevant features, and the purpose of feature selection is to remove irrelevant features.

The specific method of 90 is related to the type of index: for classified indexes, chi square test is a commonly used method in statistics to evaluate whether two events are independent. through chi-square test, the probability of variable independence p can be obtained. the smaller the value of p, the more important the characteristic variable is. in this paper, the characteristic with p value less than 0.05 is regarded as an important characteristic. For numerical indicators, the F test is a kind of hypothesis test method based on the F distribution. It analyzes whether the average value of the indicators has significant difference under different values of target variables, and the variables with significant difference have strong correlation. The independent probability P of variables is obtained through the F test. If the value of P is less than 0.05, the feature is related to the label and is an important feature.

1.2.3 Coding, sample sampling and division

(1) One-Hot Encoding classifier has discrete features that are not easy to solve, and cannot deal with problems such as character strings. It needs encoding. This article adopts the independent method. Thermal coding is used to solve these problems.

Single-hot coding is also called one-bit valid coding. It mainly uses N bit status registers to code N states, each state is separated by its own register bits, and only one bit is valid at any time. The values of discrete features extend to European space and the eigenvalues of discrete features correspond to a point[1] in European space. For example, the auto-likelihood code is 000,001,010,011,100,101 and the single heat code is 000001,000010,000100,001000,010000,100000. Each features become mutually exclusive, making the distance calculation between features more reasonable.

(2) Sample sampling

The number of positive samples in the original sample is about 699,000, the number of negative samples is about 18,000, and the ratio of positive and negative samples is 1:37.8. In order to improve the modeling efficiency, the original samples should be sampled and the amount of modeling samples should be controlled. At the same time, in order to

improve the accuracy of the model, the proportion of positive and negative samples for modeling is generally controlled to be around 1:5. The model randomly samples positive and negative samples respectively. The model used in this paper randomly samples positive and negative samples respectively, and selects 67,000 samples into the model, in which the ratio of positive and negative samples is about 1:5.

(3) Data Set Partitioning In order to ensure the generalization ability of the model, the data set needs to be divided into training set and test set. In this paper, the data set is randomly partitioned points, with 70% as the training set and 30% as the test set.

2. CNN combine algorithm

The data set used in this paper contains a certain number of positive samples. Therefore, the whole detection model is designed based on the theoretical tools of classification algorithm in machine learning. This paper improves the classification model in machine learning based on convolution neural network.

2.1 Classification algorithm based on one-dimensional convolution neural network

2.1.1 One-dimensional convolution neural network

Convolution neural network is a kind of depth feed forward neural network, which is generally composed of convolution layer and pooling layer alternating with each other[2] Have Characteristics of local connection, weight sharing and sub sampling. These characteristics make the convolution neural network invariant to translation, scaling and distortion to a certain extent, and the convolution neural network needs fewer parameters, so it has higher training efficiency. The convolution neural network updates its weight by using back propagation algorithm.

One-dimensional convolution neural network (1D-CNN, 1D Convolutional Neural Network) has the same principle and characteristics as two-dimensional or three-dimensional convolution neural network. The key difference is that the input of 1D-CNN is a one-dimensional vector, and the convolution kernel and feature map in the neural network are also one-dimensional. Convolution neural network is usually used in the fields of image processing, natural language processing and voice frequency/speech processing. In this paper, the cellular network data set is innovatively used as the input of convolution neural network, the characteristics of each user are regarded as a one-dimensional vector, and the two-classification function can be realized through the output layer of 1D-CNN.

2.1.2 Neural network structure design

The CNN convolutional neural network designed in this paper consists of input layer, 5 hidden layers, and output layer, as shown in Figure 1.

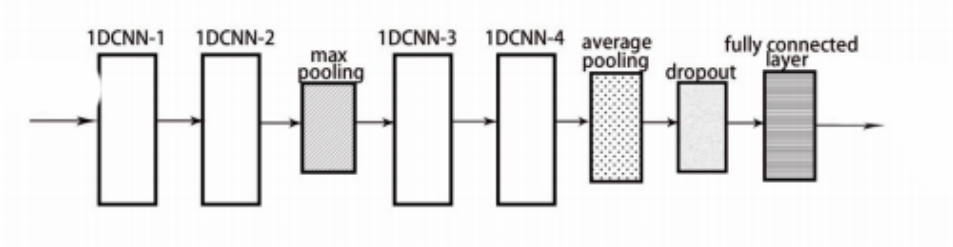

Figure 1. CNN convolutional neural network structure.

The first layer is the input layer: according to the data preprocessed previously, each data has a total of 63 features, so it is necessary to a vector with a length of 63 is transferred to the neural network, so here each user data needs to be reshaped to form A 21 x 3 matrix is used as the input of the neural network. The next 2-7 layer is the hidden layer and is the most important component of convolutional neural network.

The first 1D-CNN layer: defines a convolution kernel with a height of 2. The filter will allow the neural network to learn a single feature in the first layer, defining a total of 50 filters, which allows 50 different features to be trained on the first layer of the network. The output is 20×50 neuron matrix. Each column of the output matrix maintains the weight of a single filter. Use the defined kernel size and test. Considering the length of the input matrix, each filter will contain 20 weights.

Second 1D-CNN layer: single-layer convolution can only obtain shallow abstract information. In order to increase the depth of the neural network, a 1D-CNN layer is added. Results from the first CNN will be sent to the second CNN layer. Define 50 different filters again and train at this level. The second layer follows the same logic as the first layer, and the final output matrix size is 19 x 50.

Max Pooling Layer: pooling layer is usually used after convolution layer. The purpose of pooling layer is to reduce the complexity of output and prevent data over-fitting. The principle of maximizing the pooled layer is to take the largest eigenvalue as the value of the region in the neighborhood and set the pooled window size to 3. This means that the output matrix of this layer is only one-third of the input matrix.

Third and fourth 1D-CNN layer: add two convolution layers again to learn more global features, convolution kernel is large small is 2 and the number of filters is 80. The output matrices of these two layers are 5x80 matrix and 4x80 matrix. After a total of four convolution layers, we obtained neurons containing global features and realized feature extraction.

Global Average Pooling Layer: add another pooling layer to further avoid over-fitting. The global average pooling layer performs more extreme types of dimensionality reduction, and each feature detector has only one weight remaining in the neural network on the layer. The size of the output matrix is 1280 neurons, which is a one-dimensional vector and can be regarded as a feature vector after neural network feature extraction. It also provides the possibility to combine with other classification algorithms.

Dropout Layer: add a drop layer to solve the problems of overfitting and gradient disappearance. The discarding layer discards some neurons in the network on-board and sets the weight of 50% neurons to zero. The discarding layer averages and reduces the complex co-adaptive relationship between the divine elements, thus achieving the effect of over-fitting. The output of this layer is still the 1 x 80 divine element matrix.

Finally, the output layer is connected to the convolutional neural network in the form of a Full Connected. Since there are only two types of prediction targets: telecom fraud users and non-telecom fraud users, the vector with a height of 80 is reduced to a height of 2 at the last layer, which is completed by matrix multiplication at that layer. Using Softmax as the activation function, it can force the sum of all 2 outputs of the neural network to be one. The output value will represent the probability of each of the 2 classes.

2.2 Principle of CNN combine algorithm

In order to obtain better classification accuracy than 1D-CNN and traditional classification algorithms, we propose a classification model CNN Combination, which CNNcombine the 1D-CNN neural network designed in 2.1 with traditional classification algorithms. The advantage of 1D-CNN is that the convolution layer can extract high-level features from the input feature sequence and use the extracted features to improve the accuracy of the traditional algorithm model.

In addition, the prediction result of a single traditional algorithm model under the best parameters is the best performance of the model. Different parameters adopted by different models, in many cases, the prediction result of a certain model is not the best[3] but the prediction effect of the model on some samples in the sample set may be better than the original best prediction model. That is, the prediction results of each model are different, and it is possible to obtain better prediction accuracy by means of complementation. Therefore, a better model can be obtained by means of model fusion.

In this paper, the traditional algorithm for model fusion is to fuse multiple classifiers through meta-classifiers. After the secondary classifiers are trained, the prediction results are given based on the training data, and the meta-model is trained again based on the output of the prediction results of the secondary classifiers. Secondary classifiers can adopt many different classification algorithms, which can improve the generalization ability of the model and also reflect the advantages of each secondary classifier in different data. Therefore, the model is heterogeneous after fusion. The data set is divided into 4 parts, of which three are training sets and one is test sets. Adaboost[4], random forest[5], GBDT[6] and XG Boost[7] are integrated learning algorithms. Individual algorithms have achieved better prediction results on cellular network data sets. Therefore, the paper takes these four algorithms as the basis of integration. Firstly, AdaBoost, random forest and GBDT are selected as secondary classifiers, and the training set is subjected to K folding and cross training,

and a new data set is constructed according to the labels of the output results as new features of the data set. Then select XGBoost as the meta-classifier, train the meta-classifier according to the new data set, and finally output the prediction results of the comprehensive model.

The implementation method of combining 1D-CNN is to replace the last full connection layer in 1D-CNN with the classifier of comprehensive algorithm. The cellular network dataset is subjected to feature extraction by the neural network, and the length of the feature vector input into the synthesis algorithm is 80.

3. Model evaluation

In this section, data sets from telecom operators are used to verify the effectiveness of the model in detecting telecom fraud numbers. Experiments are designed to compare the performance differences between the proposed algorithm and the commonly used traditional classification algorithms.

3.1 Experimental design

The data set is randomly divided, with 70% as the training set and 30% as the test set. Use this dataset in Python. The training of classification algorithm is carried out under the environment. the selected algorithm includes the CNN combination classification model and design proposed in this paper.

The probability of being a telecom fraud user, the greater the value of the prediction result, the greater the probability that the sample belongs to a positive sample, i.e. a telecom fraud user, whereas the smaller the value, the greater the probability of being a normal user. Different thresholds can be set according to different requirements, and whether a sample belongs to a positive sample can be judged by comparing the results of the classification algorithm with the thresholds. The P-R curve (precision-recall)[8] and the average precision AP (average precision)[8] were used as evaluation criteria for the experiment.

In the P-R curve, P indicates the accuracy rate and R indicates the recall rate. With p as the abscissa and r as the ordinate, for the same model, a curve can be obtained according to the method of dividing points by taking samples one by one as the threshold value, which is the P-R curve. Assuming that the P-R curve of one model can completely surround another learner, the former has better performance.

3.2 Performance evaluation results

Compares the classification results of four different classification models, and the drawn P-R curve is shown in Figure 3 and Figure 4.

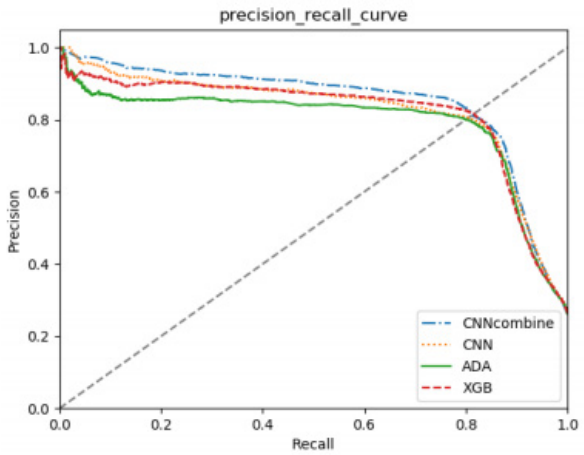

Figure 3. CNN combine algorithm structure.

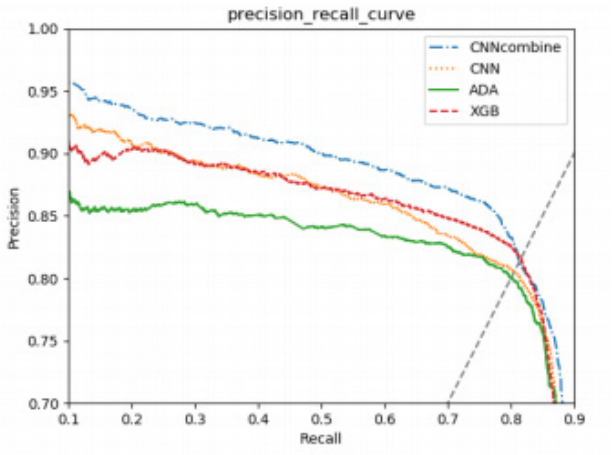

Figure 4. P-R curve comparison (upper right corner).

Comprehensive evaluation of the above model results can prove that the improved algorithm CNN Combination has better classification performance than other classification models in detecting telecom fraud users. On the other hand, when the 1D-CNN neural network designed in this paper is directly used to deal with the problem of detecting telecom fraud users in reality, the effect has no obvious advantage over traditional classification algorithms such as XGBoost. The actual evaluation proves that the telecom fraud detection model of CNN combine algorithm can more accurately detect fraudulent users from multiple users.

The output result of the telecom fraud detection model is the probability that the user is a fraud user. In practical application, it is necessary to set a classification threshold to determine the prediction result, i.e. those with a prediction probability greater than the threshold are classified as fraud users, and those with a prediction probability less than the threshold are classified as normal users. Generally, the threshold is determined by selecting the threshold corresponding to the same precision rate and recall rate, which can maintain a high precision rate and recall rate at the same time. When the precision rate and recall rate are the same, the classification threshold of the test set detection model is 0.41777, the precision rate and recall rate of the model are 0.8155, and 4275 fraud users can be correctly detected among 5242 fraud users in the test data set, and the classification accuracy rate for all 20133 users in the test set is 90.394%.

4. Conclusion

The content of this paper is to design and implement a detection model for telecom fraud users, and to propose a CNN combine algorithm that combines one-dimensional convolution neural network with traditional classification algorithm. Based on the experiment of the measured data of cellular network, this paper verifies that the CNN combination algorithm has better prediction results than the traditional algorithms XGBoost and Adaboost. The accuracy and recall rate are obviously improved. The average accuracy is 3% higher than XGBoost and 6% higher than Adaboost. This proves the effectiveness and feasibility of the model proposed in this paper in solving practical problems.

References

1. Xlinx. HDL Synthesis for FPGAs design guide[M]. XACT 1995: 3-13
2. Takahashi N, Nishi T, Hara H. Analysis of signal propagation in 1-D CNNs with the antisymmetric template[A].12th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA 2010)[C]. Berkeley:IEEE.2010.
3. Xiao JZ, Lei B, Wang CQ. Reclamation on building waste produced from Wenchuan Earthquake[A]. Shanghai: Tongji University Press; 2008. 64-65.
4. Sill J, Takacs G, Mackey L, et al. Feature-weighted linear stacking[J]. Computer Science 2009.
5. Yoav Freaud, Robert E. Schapire. Decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences 55: 119-139
6. L B. Random Forest [J]. Machine Learning 2001; 45: 5-32.
7. Friedman J H.Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics 2001: 1189-1232.
8. Chen T, Guestrin C. XGBoost: A scalable tree boosting system[J]. In Proceeding KDD ,16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco 2016; 785-794.
9. Zhou Zhihua. Machine Learning: Machine learning[M]. Beijing: Tsinghua University Publishing House; 2016. p.26-48.