K-Means Algorithm:Fraud Detection Based on Signaling Data

Xing Min, Rongheng Lin
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications,
Beijing, China, 100876

***Abstract*—At present, the crime of telecom fraud, with advanced communications and Internet technologies, is growing rapidly and causing huge losses every year. The traditional fraud detection methods are less flexible. In this paper, we used the signaling data to train a clustering model, which can discover the hidden user characteristics of fraud phones. The paper puts forward the extraction method of behavior characteristics, reduce the dimension of features with principal component analysis and select the appropriate clustering parameters through grid search, then present the K-Means-based behavior identification system, which can help to distinguish the frauds and identify the fraud phone numbers. Finally, the feasibility of this model is verified by the actual sample dataset.**

Fraud detection; signaling data; K-Means; behavior identification;

I. INTRODUCTION

The continuous development of the communications industry not only brings more convenience, but also makes the telecom fraud more rampant. According to the Global Fraud Loss Survey released by the CFCA (Communications Fraud Control Association) in 2017, global fraud losses hit US $ 29.2 billion, reaching 1.27% of global telecommunication revenue [1].

Signaling data in the modern communication system has a very important position [2], through the analysis of the signaling data, we can find the characteristics of the scam phone to identify the fraud phone. Due to the huge amount of signaling data and its growing number, the traditional statistical methods alone can’t efficiently extract valuable information. Therefore, the use of feature engineering and data mining techniques is feasible and necessary.

Scholars have done a lot of research. Moreau and Bart combined rule-based systems and artificial neural networks [3] to detect abnormalities of users' calls efficiently, but due to the need of precise class annotation and not be widely used. Rosset established a user model using a rule-based approach and established a rule set using a greedy algorithm with a threshold [4]. However, the model based on rule analysis requires accurate data and long construction period. With the continuous increase of data, the performance drops sharply and it is difficult to build a complete system.

In this paper, K-Means clustering analysis of signaling data is mainly carried out, which does not need class annotation and can handle large data. Through the analysis and mining of call detail records, the model improves continuously.

II. SIGNALING DATA PROCESSING AND CLUSTERING

A. The main idea

We first clean the raw data and construct the features, and then reduce the dimension with PCA (Principal Component Analysis), evaluate the sample's clustering trend with Hopkins statistics [5] and select the appropriate clustering parameters through grid search. Finally, we train the model which based K-Means and analyze the results.

B. Data preprocessing

Approximately 130,000 call detail records were collected from the server in two days, with a total of 97 features, leaving 67 remaining after the blank and redundant features were removed. After its statistical analysis, we use source number as ID to convert it to 26670 pieces of data.

Based on the analysis of the differences between the characteristics of fraudulent calls and normal calls and a series of exploratory data analysis, we chose a series of statistical features, including the number of calls per number, the average of talk time, the variance of talk time, the number of roaming calls, and the number of homes of the called number, etc.

Too many features will cause the Euclidean distance to fail in the high-dimensional space, so in order to reduce the number of features, we used PCA algorithm to reduce dimension. The principle of PCA is to project a high-dimensional vector X through a special eigenvector matrix U into a low-dimensional vector space, characterize it as a low-dimensional vector y, and lose only a few minor information.

After PCA dimension reduction, the first six principal components already contain 92% of the original information. Considering the balance between feature number and the loss of information, the first six principal components are selected to use.

C. Clustering trend evaluation and parameter selection

The Hopkins statistics were calculated on the dataset obtained after the principal component analysis, and the result was 0.998. It can be seen that the data has a high tendency of clustering. Then we draw SSE(the sum of squared errors) graph under different clusters, as shown in the Fig. 1. It can be seen that k=7 is best according to elbow curve.

978-1-5386-7374-4/18/$31.00 ©2018 IEEE
DOI 10.1109/SERVICES.2018.00024

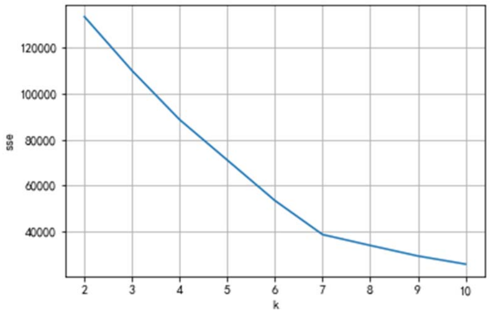

Figure 1. SSE

D. results analysis

K-Means algorithm is used to cluster the training data, and the seven clusters can be divided into a large cluster, five small clusters and an extremely small cluster.

With cluster labels as abscissa, the average of each feature of each cluster is ordinate, we draw Fig. 2. From the overall view of Fig. 2, the characteristic differences between different clusters are obvious and the clustering results have strong interpretability.

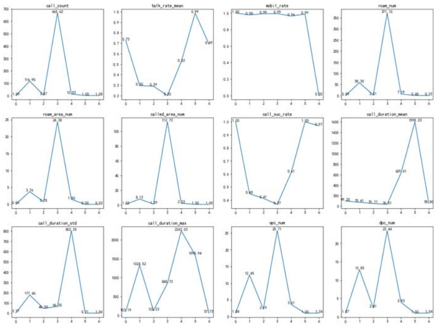

Figure 2. Mean of each feature of each cluster

It can be seen from the Fig. that the number of elements in cluster 0 reaches 75% of the total number, the average number of calls in two days is about 1.09 times, and the average duration of each call is 100 seconds, call time variance is low, indicating the difference between the time of each call is small, the average call success rate is very high, talk time accounted for 73% of the average call time, which is consistent with the normal situation. Based on above analysis, the elements in cluster 0 are more likely to be normal calls.

Cluster 1 has a total of 449 elements. The main features are the high number of calls, the low talk ratio, low call success rate, call duration, call duration has high peak value and large standard deviation, but average value is small. They are most likely fraud phones.

There are 3,868 elements in cluster 2. The call duration is short, and the standard deviation of call duration is large. Talk time percentage is only 0.29, while the number of source and target signaling points is big. We can believe that they are the manual call out fraud phone.

There are only 48 elements in Cluster 3, 0.18% of the total. The average number of calls in this cluster hits 678.15 times, and the number of roaming calls is the highest. Talk time ratio is only 0.2 and the percentage of successful calls is 0.37, which is the lowest among all cluster. it can be considered that the elements in this cluster are fraud or harassing calls made by an automatic dialing device.

There are 223 elements in cluster 4. The number of element calls in this class is slightly more than normal, and the call success rate is moderate. However, the standard deviation of call duration is large and the longest call duration is very long. The type judgment of this cluster needs to be further analyzed.

Cluster 5 has 661 elements. The elements in this cluster have the characteristics of high success rate and extremely high call duration and talk ratio, the elements in this cluster have a high probability of a normal call with an extremely long talk time.

There are 1295 elements in cluster 6, accounting for 4.9% of the total. All dialed numbers are fixed calls, and call success rate reached 0.97, in addition, other features are similar to cluster 0, it is initially believed that the elements in the cluster are normal calls whose called number is fixed.

III. CONCLUSION AND NEXT WORK

This paper uses CDR data to train K-Means model, then the potential category information is discovered and some feature patterns are found through the feature analysis of different clusters. Compared with other methods, the method in this article does not require class annotation, which reduces the difficulty of data acquisition. As the amount of data increases, the accuracy of fraud recognition increases.

In next work, different feature engineering methods will be used to process the data, such as constructing new features and adding polynomial features. At the same time, we will try other cluster algorithms and try to get better results.

ACKNOWLEDGMENT

This work was supported by the State Grid Corporation of China under the project title: “The Improved Core Analysis Algorithms and Utilities for Smart Grid Big Data” (520940180016) and the Beijing Natural Science Foundation (L171010,4174099).

REFERENCES

[1] Communications Fraud Control Association. “2017 Global Fraud Loss Survey,” Press Release, June 2017, available online: http://www.cfca.org/press.php

[2] Dan York. The Unified Communications Ecosystem[M].Elsevier Inc.:2010-06-15.

[3] Moreau Y, Bart P, Dept E, et al. Novel techniques for fraud detection in mobile telecommunication networks[A]. In: Proc of ACTS Mobile Summit[C]. Grenada, Spain, 1997

[4] Rosset S, Murad U, Neumann E, et al. Discovery of fraud rules for telecommunications -challenges and solutions[A]. In: Proc of 5th ACMAA SIGKDD Int Conf on Knowledge Discovery and Data Mining[C]. San Diego, California, United States, 1999:409-413

[5] BANERJEE A,DAVE R N. Validating clusters using the Hopkins statistic. Proc of IEEE International Conference on Fuzzy Systems . 2004