Hybrid Data Mining Method of Telecom Customer Based on Improved Kmeans and XGBoost

Yongge SHI 1*, Shaoyun YAN 2, Meibin HE 3, Xiangjun LI 2

1 Gongqing Institute of Science and Technology, Gongqingcheng, Jiangxi, 330029, China
2 Nanchang University, Nanchang, Jiangxi, 330031, China
3 Jiangxi branch of China Telecom Co, Nanchang, Jiangxi, 330047, China
*Corresponding author’s e-mail: ncsyg@163.com

Abstract: In order to mine specified telecom customers with special behaviours from vast voice communication records of Telecom company, a novel Hybrid Data Mining method of Telecom Customers (HDMTC) with special behaviours is proposed in this paper, which integrates Kmeans and XGBoost into one framework. First, AHP model helped to model the features of customers with special behaviours. Then, Due to semi-supervised methodology, Kmeans approach was improved by small amounts of tagged initial cluster centroids and weighted Euclidean distance. Improved Kmeans was utilized to tag the decision attribute of data to construct training dataset. After that, XGBoost model was trained via the training dataset. Finally, the efficiency of HDMTC was validated via real records from telecom company.

1. Introduction
Nowadays, the great development of communication technology substantially changes the style of people’s life and work, whereas more and more phone spams and telecom frauds also intrude into people’s life. Thus, it is critical need of telecom company that customers with special behaviors are discovered via data mining method from massive records of telecom communication and treated with specified solution.

Wang used fuzzy decision tree to recognize the behaviors of customers communication based on the records of telecom voice communication [1]. Liu analyzed the relation between customers to mining customers with strong association via semantic association algorithm based on fuzzy set. Without the analyzation of customers’ behaviors, recall rate of his method was not good enough [2]. Song built XGBoost model to classify the abnormal customers based on the correlation coefficient and temporal features with diverse scales [3]. Zhou and Zou used improved Kmeans Algorithms, which are unsupervised learning methods, to build customer data mining models. However, the precision rate and recall rate of their unsupervised learning methods are not good enough without the guidance of tagged data [4-5]. Henriques combined Kmeans and XGBoost approaches to detect abnormal log events from massive log records [6]. Du integrated Kmeans and SVM approaches to improve the accuracy rate [7].

2. Methodology
Based on improved Kmeans and XGBoost algorithms [10], a novel Hybrid Data Mining method of
Telecom Customers (HDMTC) with special behaviors is proposed in this paper. Based on small amount of tagged data, this method can not only decrease the cost of tagging, but also improve the performance of data mining. The procedure of Algorithm is represented as following figure 1.

![Figure 1. Flow chart of HDMTC Method](image)

2.1 Preprocessing
Raw data were selected from the voice communication records of China telecom company. There are 18 attributes in each record. These raw data were preprocessed with 3 stages, which included to fill the missing values, delete the inconsistent or abnormal data, and extract the related attributes to model features of special behaviors.

2.2 Modeling the features of customer’s special behaviors
In terms of special behaviors of telecom customer, the strong correlated features were selected to construct feature-based dataset of customer’s special behaviors. In this paper, the special behaviors of customer, which were figured out by experts of telecom company, were phone scam, advertisement, and business call. Furthermore, based on quantitative and qualitative analysis of Analytic Hierarchy Process (AHP) methods, decision maker balances the importance of specified objective, and set weights of objective features.

The stages of AHP model are described as following:
Stage 1: Build AHP model. In terms of decision attributes, all customers with special behavior t can be setup on the top level. The customers with the behavior of phone scam f, advertisement a, business b can be setup on middle level respectively. According to the decision rule, relevant conditional attributes of voice communication can be setup on bottom level. The AHP model is represented as following figure 2.
Stage 2: Construct Judgement Matrix, and optimize the weights of features. According to the experience of telecom experts, the initial weights of features can be setup. Based on the weights of features, judgement matrix of every level can be constructed, including judgement matrix of customer with special behavior $T$, judgement matrix of customer with advertisement behavior $A$, judgement matrix of customer with phone scam behavior $F$, judgement matrix of customer with business behavior $B$. Based on the sum-product method of judgement matrix, the weights of judgement matrix can be optimized with largest eigenvalue $\lambda_{max}$.

Stage 3: Check of consistency of judgement. The consistency can be checked by the rate of consistency $CR = CI/RI$, in which the Consistence Indicator of judgement matrix is $CI = (\lambda_{max} - n)/(n-1)$, and Random Indicator $RI$ can be searched in order table of matrix.

In order to figure out the best solution, all weights of features in judgement matrix need to be calculated and check to integrate into optimized synthetical weights as following table 2, including features’ weights of customer with special behavior $W_T$, features’ weights of customer with advertisement behavior $W_A$, features’ weights of customer with phone scam behavior $W_F$, features’ weights of customer with business behavior $W_B$.

| Table 1. Optimized weights of customers’ attributes |
|---------------------------------------------------|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ |
| $W_T$ | 0.210 | 0.268 | 0.035 | 0.035 | 0.053 | 0.193 | 0.204 |
| $W_A$ | 0.292 | 0.292 | 0.040 | 0.040 | 0.000 | 0.222 | 0.112 |
| $W_F$ | 0.290 | 0.290 | 0.041 | 0.041 | 0.074 | 0.000 | 0.262 |
| $W_B$ | 0.180 | 0.180 | 0.160 | 0.160 | 0.100 | 0.110 | 0.108 |

2.3 Improved Kmeans Clustering

2.3.1 Selection of Initial Cluster Centroid
Selection of initial cluster centroids of Kmeans algorithm effects the final result significantly. In this paper, the tagged samples were employed to indicate the selection of initial cluster centroids, compare to traditional random selection of initial cluster centroids. According to the experience of telecom experts, $S$ samples $X_i$ ($i \in [1,2,3,...,S]$) are selected from dataset $X$, and tagged as training dataset $X_i (X_i \subset X)$, in which at least one sample $x_i \in X_i$ in each cluster $C_k (k$ is number of clusters). Finally, the initial cluster centroids $m_k$ are calculated.
2.3.2 Weighted Euclidean Distance

Euclidean distance, which is utilized in classical Kmeans, doesn’t take diversity and correlation of samples into consideration, and use same weights to measure similarity for all samples [8-9]. Therefore, in this paper, according to the optimized judgement matrix and weights, Euclidean distance is adjusted with the importance of specified features.

2.3.3 HDMTC Method:

**HDMTC METHOD**

**INPUT:** Dataset $X = \{x_1, x_2, x_3, \ldots, x_n\}$, $x_i \in X$, $i \in [1, \ldots, n]$, Tagged dataset $X_s = \{x_1, x_2, x_3, \ldots, x_k\}$, $s \in [1, \ldots, k]$, Number of clustering $k$, and XGBoost Model $F$

**OUTPUT:** HMTSC model, and Predicted customer classes $Y'$ with special behaviors (advertisement behavior, phone scam behavior, business behavior)

**STEP:**

For $k=1,2,\ldots,n$ do

Locate initial cluster centroid $m_k = \frac{\sum_{x \in X_s} x_i}{S_k}$ with tagged dataset $X_s$

End for

Repeat

For all $x_i \in X$ do

Calculate Euclidean distance $d(x, m_k) = \|x, m_k\|

Partition the data into cluster $C_k$ with nearest cluster centroid $m_k$

End for

For $k=1,2,\ldots,n$ do

Update cluster centroid $m_k$

End for

Until cluster centroids don’t change

Obtain supervised training dataset $\{X, Y\}$ based on final cluster centroid $m_k$

Train XGBoost Model $Y = F(X)$ by supervised training dataset $X$

Predict final customer classes $Y' = F(X^{new})$ with new data $X^{new}$

3. Validation

3.1 Dataset of Experiment

Raw dataset was selected from 96,000 voice communication records of Telecom Company of China, including 600,000 online communication, 2,600,000 communication from mobile phone. 50,000 records were chosen randomly and preprocessed to construct training dataset. According to experience of experts, there are 100 tagged customers, including normal customers, customers with advertisement behaviors, customers with phone scam behaviors, and customers with business behaviors, which were setup as initial cluster centroid. Finally, there are 200 normal customers and 400 customers with special behaviors to be selected as testing dataset.

| Table 2. Indicators of HDMTC Model Testing |
|--------------------------------------------|
| XGBoost model | 89.3 | 90.5 | 89.6 | 95.1 | 95.8 | 95.4 |
| SVM | 88.2 | 89.1 | 88.6 | 97.5 | 94.4 | 95.9 |
| Naïve Bayesian model with Improved Kmeans | 93.4 | 90.3 | 91.8 | 93.6 | 90.1 | 91.8 |
| XGBoost model with Improved Kmeans | 95.8 | 94.1 | 95.0 | 98.2 | 99.1 | 98.6 |
3.2 Analysis of Experiment
The performance of XGBoost, SVM, naïve Bayesian model with improved Kmeans, and our XGBoost with improved Kmeans was compared as above table 2. According to F1 score of models for customers with special behaviors, 95.5% from our XGBoost with improved Kmeans, is better than the others. According to F1 score of models for customers with business behaviors, 98.6% from our XGBoost with improved Kmeans, is better than the others. Furthermore, precision of prediction for customers with business behaviors is higher than precision of prediction for customers with all special behaviors. Thus, the performance of our our XGBoost model with improved Kmeans is better in the analysis of customers with business behavior.

4. Conclusion
Based on the specified features of voice communication from telecom customers, the customers with special behaviors, including advertisement, phone scam and business, can be figured out to support specified solution of telecom company. In this paper, a Hybrid Data Mining method of Telecom Customer (HDMTC) combined with improved Kmeans and XGBoost was proposed to mine the customers with special behaviors. Semi-automatic tagging approach based on improved Kmeans was utilized to help organize training dataset. Then, XGBoost model can be trained well to classify the type of customers. Finally, the efficiency of our HDMTC method was validated via actual data from voice communication records of telecom company in the experiment.

Acknowledgments
Authors wishing to acknowledge financial support from project of science and technology department of Jiangxi province (No. 20192BBEL50041).

References
[1] Wang, J. (2013) Anomaly identification of user behavior based on complex network theory and mobile phone call records. University of Science and Technology of China.
[2] Liu, X. (2008) Application of Fuzzy-association Rule to Telecom Data Mining. Computer and modernization, 06:36-38.
[3] SONG, X., SUN, X., ZHAO, Y. (2018) Abnormal user identification based on XGBoost algorithm[J].Journal of Measurement Science and Instrumentation, 9(04):339-346.
[4] ZHOU, J., SHI, Y., HE, M. (2018) Application of K-means algorithm based on A-D model in calling abnormal customer mining. Telecommunication Science, 34(04):81-89.
[5] ZUO, G., ZHOU, R., LI, Z. et al. (2012) Improved K-means algorithm and its CRM in telecom customer segmentation application. Computer Systems & Applications, 19(11): 155-159.
[6] Henriques, J.; Caldeira, F.; Cruz, T.; Simões, P. (2020) Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. Electronics. pp.1159-1164.
[7] DU, Y., JIANG, Z., FENG, L. (2019) Novel learning algorithm combining support vector machine and semi-supervised K-means. Computer application, 39(12):3462-3466.
[8] MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1(14): 281-297.
[9] YANG, J., ZHAO, C. (2019) Survey on K- Means clustering algorithm. Computer Engineering and Applications, 55(23):7-14.
[10] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). pp. 785–794.