Prediction of Tariff Package Model Using ROF-LGB Algorithm

NaiSong Zheng†
School of Mathematics and Statistics
Xi’an Jiaotong University
Xi'an 710049, China
naisong.zheng@stu.xjtu.edu.cn

XiaoWei Jiang
School of Mathematics and Statistics
Xi’an Jiaotong University
Xi'an 710049, China
jiangxv@stu.xjtu.edu.cn

Yibo Ao
School of Electronic and Information Engineering
Xi’an Jiaotong University
Xi'an 710049, China
perayb@stu.xjtu.edu.cn

Xi Zhao
School of Management
Xi’an Jiaotong University，Xi'an 710049, China
The Key Lab of the Ministry of Education for process control & Efficiency Engineering， Xi'an 710049, China
zhaoxi1@hotmail.com

ABSTRACT

With the slowing growth of the telecommunication market and the intense competition for existing customers, Customer Churn Management has become a crucial task for all mobile network operators. Recommendation models based on customer behaviors are widely used by operators to provide diverse telecom tariff packages for suitable people and thus improve customer satisfaction. To address the low precision rate and data granularity of prior studies, this study combined rotation forest (ROF) and LightGBM and construct a hybrid algorithm (ROF-LGB). Grid search method was used in parameter tuning, and ten-fold cross-validation method was used to prevent overfitting. Using mobile data generated by operators, ROF-LGB method was tested and compared with other five traditional machine learning methods. The results showed that ROF-LGB method achieved better performance with better precision rate and execution efficiency in telecom tariff package recommendation.

CCS CONCEPTS

•Computing methodologies→Combinatorial algorithms.

KEYWORDS

Tariff Package, Prediction, ROF-LGB, LightGBM

ACM Reference format:

NaiSong Zheng, XiaoWei Jiang, Yibo Ao and Xi Zhao. 2019. Software Defect Prediction Based on Ensemble Learning. In *Proceedings of 2019 2nd International Conference on Data Science and Information Technology (DSIT’19). Seoul, Republic of Korea, 5 pages.* https://doi.org/10.1145/3352411.3352421

1 Introduction

With the increasingly fierce competition in the mobile network service market, customer churn management has become an important task for mobile communication operators. In order to improve customer loyalty, the operators devote themselves to designing all kinds of telecom tariff packages suitable for different types of people. However, it’s an urgent problem for operators to recommend different telecom tariff packages to suitable customers.

Many studies have analyzed the customer churn issue from the perspective of the operators. Huang B. et al. [1] constructed a customer churn prediction model based on machine learning algorithm such as decision tree and SVM by using users’ Cell Detail Records (CDR) data and bill information. Vafeiadis T. et al. [2] compared several machine learning methods and adopted the SVM-POLY using AdaBoost with an accuracy of nearly 97% and an F-measure of over 84%. In order to improve the competitiveness of telecom operators and improve user loyalty, Aheleroff S. divided regular users into four loyal groups by using clustering method by the CDR data of users and their consumer data, thus implementing different marketing strategies [3]. Kolarovszki P [4] et al. found the factors influencing brand loyalty and user involvement in the telecom service field based on the questionnaire data. In terms of the pricing of telecom packages, Basaran et al. [5] focused on the factors that influence users’ choice of mobile operator and the pricing strategy of telecom tariff package. Su Z X [6] built the telecom package evaluation model by using AHP, so as to put forward specific suggestions for operators' marketing management. These studies evaluate and improve the telecom tariff packages based on the users’ choice behavior and the factors that affected users’ the choice of operator. However, it didn’t mention recommending suitable telecom tariff packages for users by their usage behavior data. Gao X et al. used the improved ARIMA model to predict users’ future consumption

DSIT '19, July 19–21, 2019, Seoul, Republic of Korea
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7141-4/19/07…$15.00
https://doi.org/10.1145/3352411.3352421

data, and helped them choose suitable package based on consumption [7]. Shuochen X et al. [8] introduced a matching model that combines customers' usage behavior and telecom tariff packages by calculating the similarity. Miao Y H et al. [9] built multiple logit model (MNL) with multiple attributes to predict the probability of users' choice of telecom package. In the end, it obtained an accuracy of 53.3% on the dataset of 1000 people. The above literatures are all experimented based on small-scale samples. On the one hand, due to the limitation of time complexity, they are not suitable for large-scale samples. On the other hand, except for the low accuracy of the model in the last literature, other literatures lacked the accuracy of the model.

In this paper, we proposed a telecom tariff package recommendation method by using the ROF-LGB (Rotation Forest and LightGBM) method which combined Rotation Forest algorithm and LightGBM algorithm. We applied the ROF-LGB method to the real operator's data sets and compared with other machine learning methods such as Random Forest and XGboost. The result shows that the ROF-LGB method with an accuracy of nearly 92% in the tariff package recommendation.

The structure of this paper is as follows. In Section 1 we introduce the background and literature review of telecom tariff package recommendation. We then explain the implementation process of ROF-LGB methods in Section 2. We provide our dataset, experiment and result in Section 3. Finally, we draw conclusions, and suggest directions for future research in Section 4.

2 ROF-LGB Algorithm

In this section, we introduce the detail of the ROF-LGB method which combines ROF algorithm and LightGBM algorithm. This section is divided into two parts: the basic principles of LightGBM algorithm and Rotation forest respectively and the implementation process of the ROF-LGB method combining above algorithms. [10]

2.1 LightGBM Algorithm

LightGBM is an improved algorithm based on Gradient Boosting Decision Tree (GBDT). Comparing to the traditional GBDT, LightGBM has faster running speed and better accuracy, and supports parallel learning to handle large-scale data. LightGBM did two optimizations: using the Grandient-based One-Side Sampling (GOSS) algorithm to filter samples with smaller weights and using the Exclusive Feature Bundling (EFB) algorithm to reduce the feature dimension. Two optimization algorithms and the overall flow of LightGBM are described in detail below.

2.1.1 GOSS. GOSS is an algorithm that balances the amount of data and guarantees accuracy. The main idea of the algorithm is to preserve large gradient samples and randomly select small gradient samples and make up a constant weight, so as to achieve fewer samples without reducing the precision. The main steps are as follows:

1. Select the gradient value in the first a% sample as the large gradient value sample;
2. From the remaining samples, randomly select b% samples as small gradient value samples;
3. Consider the impact of small samples on the model, the weight of the small gradient sample (1-a)/b is given.

2.1.2 EFB. EFB is an algorithm that improves the training speed by combining mutually exclusive features. The algorithm is mainly divided into two questions: determining which features can be merged and how to merge the features.

The first problem is an NP-hard problem: the feature is treated as a point, and the feature of the conflict is always considered as an edge. The problem is to find the merged feature and minimize the feature dimension. This method can be solved using a greedy algorithm.

The second problem arises because the different feature ranges in the bundled feature are different, so the merged feature range needs to be rebuilt. In a set of bundled features, these features can be safely merged by adding offsets to some of the features.

2.1.3 GBDT. After the above two steps, we use the GBDT model to train the acquired features. The main process of the model is shown below:

1. Initialize the weak classifier according to the following formula to minimize the loss function.

$$f_0(x) = \arg min_\gamma \sum\nolimits_{i=1}^{N} L(y_i, \gamma)$$

2. The following process loops M times
3. Calculate the residual estimate of the model.

$$r_{mi} = -[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}\ (x)}$$

4. Fitting the regression tree to get the leaf nodes of the m tree.
5. For $, j = 1,2,\ldots,J$ compute

$$C_{mj} = \arg\min_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)$$

6. Update regression tree

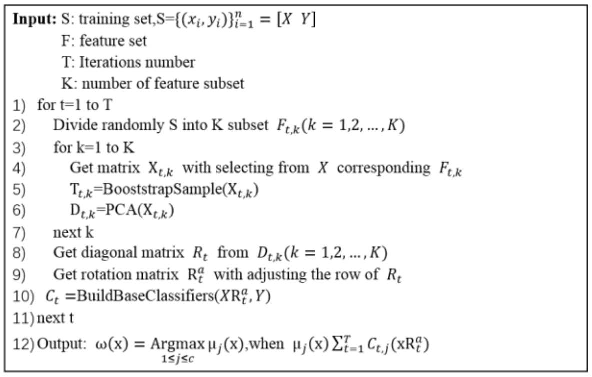
$$f_m(x) = f_{m-1}(x) + \sum\nolimits_{j=1}^{J} c_{mj} I(x \in R_{mj})$$

7. Output the final Model

$$\hat{f}(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$$

2.2 Rotation Forest Algorithm

Rotation forest algorithm could randomly divide the original feature set into several feature subsets, use principal component analysis to perform feature axis rotation on the original data set, and then input the obtained new data set into the base classifier. Finally, for each base classification, the predicted probability of the device takes the principle of maximum confidence to determine the final prediction category. Figure 1 shows the steps of Rotation forest algorithm.



```
Input: S: training set,S={$(x_i, y_i)$}$_{i=1}^{n}$ = $[X\ \ Y]$
        F: feature set
        T: Iterations number
        K: number of feature subset
for t=1 to T
    Divide randomly S into K subset $F_{t,k}(k = 1,2, \dots, K)$
    for k=1 to K
        Get matrix $X_{t,k}$ with selecting from $X$ corresponding $F_{t,k}$
        $T_{t,k}$=BooststrapSample($X_{t,k}$)
        $D_{t,k}$=PCA($X_{t,k}$)
    next k
    Get diagonal matrix $R_t$ from $D_{t,k}(k = 1,2, \dots, K)$
    Get rotation matrix $R_t^a$ with adjusting the row of $R_t$
$C_t$ =BuildBaseClassifiers($XR_t^a, Y$)
next t
Output: $\omega(x) = \underset{1 \le j \le c}{\mathrm{Argmax}}\, \mu_j(x)$,when $\mu_j(x) \sum_{t=1}^{T} C_{t,j}(xR_t^a)$
```

Figure 1: Pseudo-code for Rotation Forest Algorithm

2.3 ROF-LGB Model

As Figure 2 shows, the basic idea of Integrated learning prediction model based on ROF algorithm and LightGBM algorithm (ROF-LGB) is to divide the feature set and principal component analysis (PCA) as well as perform feature axis rotation on the data set, which makes each base classifier LightGBM get different training sets. On the one hand, this approach not only ensures the diversity between the base classifiers, but on the other hand, since only a part of the features is used for each training, this can prevent over-fitting of the model.

Compared with the other six integrated learning algorithms (Random Forest, Adaboost, Bagging, GBRT, Xgboost, LightGBM), ROF-LGB has the following advantages:

1. It can achieve a relatively high prediction effect even though in the case where the amount of basic classifiers is small;
2. It can be applied to deal with the unevenly distributed data of class samples in a great effcet.

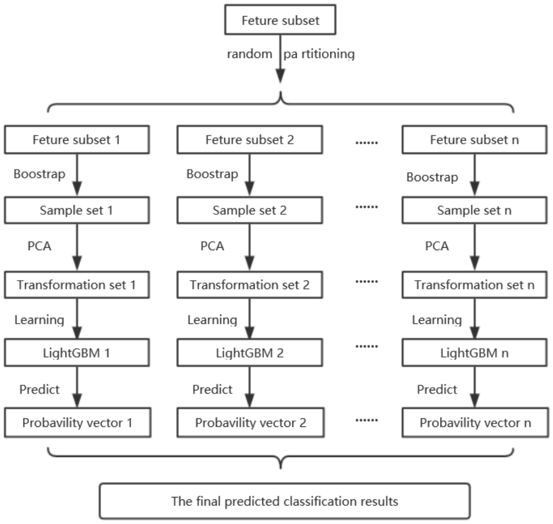


Figure 2: Schematic Diagram of the ROF-LGB Model

3 Experiments

3.1 Dataset

The dataset used in this study was from mobile phone signaling data, consumption and account information of a Chinese operator. A total of 870,000 users were collected for data within three months. The mobile phone signaling data includes the user's CDR data and SMS data, and the consumption situation includes the user's package usage status and the detailed transaction bill record, and the account information includes the user's gender, age, and duration of the network. Specific characteristics are shown in Table 1.

Table 1: Dataset Features

Data Source	Data Type	Data Description
Signaling Data	CDR	Incoming Call
		Outgoing Call
	SMS	Receive SMS
		Send SMS
Consumption Data	telecom tariff package	Package Name
		Package Type
		Whether the phone broadband integration package
		Contract Type
		Contract Duration
		Local Voice Call Length
		Voice Call Length Outside the Package
		Carryover Data Last Month
		Monthly Cumulative Usage Data
		Continuous Beyond the Package Range
		Whether to promise the lowest consumer users
	Transaction Record	Number of transactions
		The transaction amount
		Current monthly bill amount
		The amount of the bill in the month before the month
		The amount of the bill in the first two months of the month
		The amount of the bill in the first three months of the month
account information		Gender
		Age
		Online Time
		Net Service Type

3.2 Experiment Process

3.2.1 Model Training

In our research, the data was randomly divided into two groups, of which 70% were used as training sets and 30% are were as verification sets. We apply the gridded tuning method to find the optimal parameters, and the ten-fold cross-validation method to prevent the model from over-fitting. The result shows that the accuracy is significantly improved by combining two strong classifiers in our model. The optimal parameters of the model are presented in Table 2.

Table 2: The Best Parameter List of ROF-LGB Model

model	parameter	Best value
ROF part	B	5
LGM part	learning_rate	0.2
	num_iterations	100
	max_depth	5
	min_data_in_leaf	70
	feature_fraction	0.7
	bagging_fraction	0.7
	num_leaves	100
	bagging_freq	5
	lambda_l1	0.8
	lambda_l2	0.8

3.3 Model Evaluation

3.3.1 Compared with Other Algorithms

So as to measure the advantage of ROF-LGB method in the tariff package forecasting problem, this study compares the accuracy and the operating efficiency of Random forest, Adaboost, Bagging, GBRT, Xgboost, LightGBM on the same data set while ensuring the depth of the tree and the random tree seeds are consistent. The result is shown as Table 3. Based on Table 3, it can be seen that the ROF-LGB method has better performance than the other six algorithms in accordance to accuracy. Although the accuracy of Xgboost is not much different, but the operating efficiency is significantly higher than it. The pre-processing of datasets by ROF part can ensure a higher accuracy rate with fewer number of trees applied in LGM part, thus our method is relatively improved than LightGBM in accuracy and operation efficiency.

Table 3: Comparison of Accuracy and Operational Efficiency of these Algorithms

Method	Accuracy	Running Time
ROF-LGB	0.9168	489.94
Random Forest	0.7456	821.42
Adaboost	0.6797	1062.66
Bagging	0.8499	2022.26
GBRT	0.8567	14558.43
Xgboost	0.9065	7379.28
LightGBM	0.9147	585.35

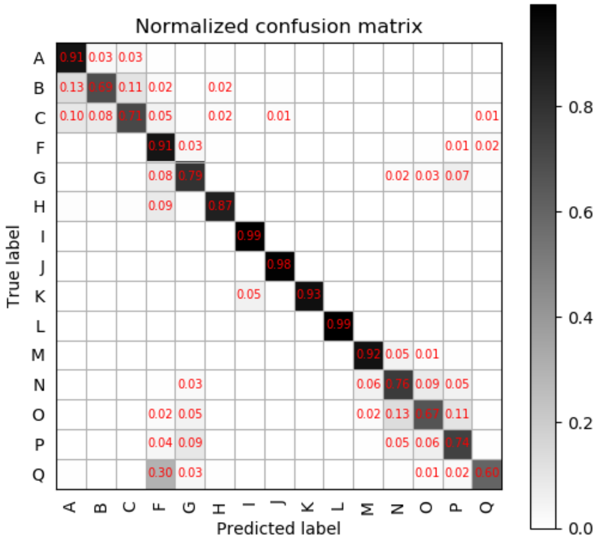


Figure 3: The Confusion Matrix of ROF-LGB Model

The final prediction result of this study is 0.9168, the recall rate is 0.8334, and the F1 value is 0.8371. Judging from the confusion matrix in Figure 3, this study has strong predictive ability for I, J, and L telecom tariff packages and weaker predictive ability for O and Q telecom tariff packages.

3.3.2 Model Sensitivity

In order to obtain the sensitivity of the ROF-LGB method to the size B of the feature subset and the influence on the accuracy of the model, the other parameters of the fixed model are unchanged. The value range of B is set to [2, 13], and the step size is 1. A change in the sensitivity of the model is obtained in Figure 1. It is found from Figure 4 that the accuracy of the model is decreased with the increase of the parameter B.

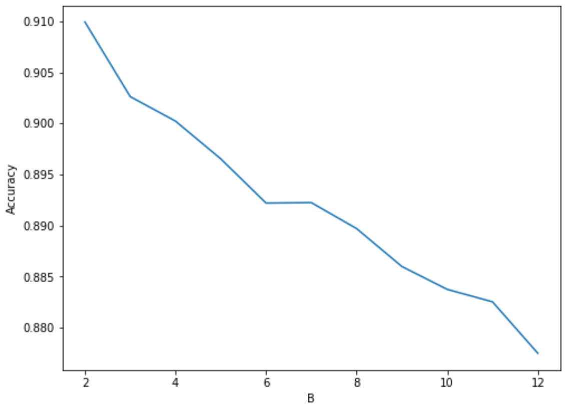


Figure 4: Comparison of Model Accuracy Rates for Different B

4 Conclusion

In this paper, a method combined Rotation Forest and LightGBM is proposed to improve the classification accuracy and solve the problem of low efficiency of Rotation Forest algorithm. Experiments show that the ROF-LGB method with an accuracy of nearly 92%, not only improves the classification accuracy effectively, but also greatly reduces the complexity of the algorithm and speeds up the prediction.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 91746111, Grant No.71702143), Ministry of Education & China Mobile Joint Research Fund Program (No. MCM20160302), Shaanxi provincial development and Reform Commission (No. SFG2016789), the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Huang B, Kechadi M T and Buckley B. 2012. Customer churn prediction in telecommunications. *Expert Systems with Application*. 39, 1, 1414-1425.

[2] Vafeiadis T, Diamantaras K I and Sarigiannidis G, et al. 2015. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*. 55: 1-9.

[3] Aheleroff S. 2011. Customer segmentation for a mobile telecommunications company based on service usage behavior. In *International Conference on Data Mining & Intelligent Information Technology Applications*. Coloane, Macao, 308-313.

[4] Kolarovszki P, Tengler J and Majerčáková M. 2016. The new model of customer segmentation in postal enterprises. *Procedia-Social and Behavioral Sciences*. 230: 121-127.

[5] Basaran A A, Cetinkaya M and Bagdadioglu N. 2014. Operator choice in the mobile telecommunications market: Evidence from Turkish urban population. *Telecommunications Policy*. 38, 1, 1-13.

[6] Su Z X. 2015. Study on the evaluation of telecom operator's mobile service package based on AHP method. *Telecom Engineering Technics and Standardization*. 28, 4, 50-53.

[7] Gao X, Cao Z and Zhang X. 2017. Research on the recommendation of mobile phone tariff package based on time series analysis. In *2017 6th International Conference on Computer Science and Network Technology (ICCSNT)*. Dalian, China, 213-217.

[8] Shuochen X, Lianju N and Wenying Z. 2017. Study of matching model between tariff package and user behavior. *The Journal of China Universities of Posts and Telecommunications*. 24, 3, 91-96.

[9] Miao Y H, Tang J F and Zhang T J. 2013. Behavior prediction of telecom consumers choice with packages based on improved MNL model considering reference dependent. *Systems Engineering*. 31, 6, 78-82.

[10] Ke G, Meng Q and Finley T, et al. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*. Long Beach, California, United States, 3146-3154