Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 35 No. 1s (2025)

Customer Churn Prediction Model Based on Adaptive Clustering Mixed-Sampling

MA Wen-bin[1],Chen Shuofeng [2,*]

[1]Assistant Researcher, Department of Academic Affairs, Guangxi University of Finance and Economics, Nanning Guangxi 530007, China;

[2].PhD Researcher, Binary University of Management & Entrepreneurship, Malaysia；Senior economist, Guangxi University of Finance and Economics, Nanning, Guangxi, China.;

*.Corresponding author: 2004110399@gxufe.edu.cn

Article History:

Received: *25-08-2024*

Revised: *03-10-2024*

Accepted: *19-10-2024*

Abstract:

Predicting the probability of customer churn is an important reference for formulating and implementing customer retention strategies. Compared with single classification method, ensemble learning method can obtain better generalization ability and ensemble learning-based customer churn prediction method has gradually become a research focus. However, the prediction data of customer churn is usually imbalanced in class, and there are problems of overlapping and class-imbalance, so the prediction effect of ensemble learning model is ineffective. Therefore, an ensemble learning method based on adaptive clustering mixed-sampling (CUS-Ensemble) is proposed. This method is based on Bagging and regarded the gradient boosting decision tree as classifier, Firstly, cleanse the date of the whole training set, Secondly, the clustering under-sampling method is used to sample majority class instances, so that the sampled majority class instances are slightly more than the minority class instances, then, the sampled majority class instances and all minority class instances form a new training subset and provide a data cleansing again, next, using Borderline-SMOTE to balance the cleaned subset, and input it into the decision tree. Finally, the output of several gradient boosting decision trees is integrated as the final prediction result. Experimental results on six imbalanced customer datasets show that, compared with EasyEnsemble and other comparison methods, the AUC and Recall value of this method are increased by 4% and 13.6% on average, which is helpful to reduce the losses caused by customer churn.

Keywords: Imbalanced data; Adaptive clustering mixed-sampling; Ensemble learning; Customer churn prediction

1 Introduction

With the increasingly fierce market competition, there are few potential customer groups, so how to improve customer retention rate has become an important issue faced by enterprises. Customer churn prediction technology can identify customers with churn tendency in the coming period by mining their historical data, which is an important link of customer retention.

The key to predict customer churn lies in the high accuracy of the model and the ability of data processing. At the early part of the study, logistic regression, decision tree, neural network, support vector machine and other methods have played an important role in the construction of the churn prediction model. However, this kind of model is relatively simple, and its generalization ability on data with complex characteristics is ineffective and the prediction effect of customer churn needs to be improved.

Compared with the above-mentioned single model, ensemble learning integrates the results of multiple base classifiers with differences by using appropriate combination strategies, which has stronger generalization ability. Therefore, the researchers will introduce random forest[1,2], Adaboost[3,4], GBDT[5], XGBoost[6] and other classical ensemble learning methods into customer churn prediction, which further improves the prediction performance of the model. According to the weight allocated by boosting algorithm, Lu et al[7] have divided customers into two clusters, and established churn prediction model on each cluster. Li Weikang and others[8] have proposed a two-layer integrated forecasting model including Stacking and Voting layer, which avoids the dimension problem and data sparseness of customer churn data. In order to model the time series data in customer data. Zhou Jie et al[9] have proposed an ensemble learning method based on long-term and short-term memory network (LSTM), which significantly improves the churn prediction effect.

The introduction of ensemble learning method promotes the development of customer churn prediction, but the imbalanced customer churn data poses challenges to ensemble learning. For example, in the music streaming media customer data provided by WSDM Cup 2018, the number of non-churn customers is 14.6 times that of churn customers. The results of ensemble learning on such data will be biased towards majority class instances, resulting in a high mis-judgment rate of a few lost customer samples, which may bring great losses to enterprises. How to improve the prediction effect of customer churn under imbalanced data has become a hotspot for researchers.

At present, the research on imbalanced data classification can be divided into two categories: resampling and re-weighting. Resampling balances the class distribution by adjusting the quantity of instances in the training set, mainly including Under-sampling[10], Over-sampling [11], Mixed-sampling[12] and other methods. However, under-sampling will lose majority class instances which containing important information; Oversampling method will introduce too much data, which

will lead to over-fitting and inefficient in the training of large-scale and highly imbalanced data.

The re-weighting method assigns different weights to different classes or even different samples to reduce the deviation of classifier on imbalanced data. For example, cost-sensitive learning[13], Focus Loss for deep learning[14], et al. Re-weighting method usually only modifies the loss function on the original data, which makes effective use of all sample information and maintains the computational complexity of the algorithm. However, the cost matrix of cost-sensitive learning needs to be set according to the prior knowledge of experts, which often lacks the guidance of prior knowledge in reality.

Furthermore, the ensemble learning method combined with resampling technique or re-weighting method shows superior performance in the task of imbalanced data classification[15-17]. Zhu et al[18] compares the performance of techniques dealing with category imbalance in customer churn prediction. The experimental results show that the adopted evaluation indexes have a greater impact on technical performance, and the combination of Bagging and random under-sampling has more advantages when using AUC.

However, the customer churn prediction data is imbalanced among class, as well as the distribution overlapping of data and intra-class imbalance. In the above-mentioned studies, few of them study and solve these problems at the same time. Therefore, based on the existing research, this paper incorporates the mixed-sampling method based on adaptive clustering, and uses the characteristics of clustering method "high similarity within clusters and low similarity between clusters", in order to reduce the influence of data imbalance on the model, reduce the overlapping of data distribution by using the data cleansing method, balance data categories by using the oversampling method, and improve the utilization rate of majority class instances by multiple clustering mixed with under-sampling. Experiments on imbalanced customer data show that the proposed method performs well on AUC, Recall, Precision and other indicators and improves the classification ability of ensemble learning on imbalanced customer churn data.

1 Related research

1.1 Sampling method based on clustering

Clustering is a common unsupervised learning method, which divides the dataset into several disjoint clusters by using the intrinsic properties of data through certain learning strategies, so that the similarity of instance within the same cluster is high, but the similarity of instances between clusters is low. According to different learning strategies, it can be divided into prototype clustering, density clustering, hierarchical clustering and so on.

K-means clustering is the most common prototype clustering method. Rayhan et al[19] have integrated K-means into the training process of Boosting method, and proposed a cluster-based boosting method CUSBoost. In each iteration, majority class instances are clustered into k clusters,

then random under-sampling is performed on each cluster, 50% samples are randomly selected, and then these representative samples are combined with a few samples to obtain a balanced dataset. In the data pre-processing stage, Lin et al[20] have proposed two clustering techniques to replace the random under-sampling technique. The cluster center and the nearest neighbour of the cluster center are used to represent the majority class respectively, and the number of clusters in the majority class is set to be equal to the number of data points in the minority class. In order to find the optimal clustering of different datasets, Zhou Qian et al[21] have proposed *k*A distance weighted sampling algorithm based on adaptive k-means clustering. Chloe Wang, et al[22] consider that K-means clustering cannot distinguish the importance of different features in the clustering process, so Euclidean distance is introduced into the clustering under-sampling process, and sample weights are assigned according to the distance.

To some extent, the above research results overcome the limitations of under-sampling, but there are still some shortcomings:

First, most studies set the cluster number of majority class instances as the number of minority class instances. When there are minority class instances, the clustering process is inefficient and not best; Second, clustering method is mainly used as data pre-processing method, and less combined with ensemble method. Although the samples sampled from this method are representative, the problem of information loss has not been solved. Third, the clustering under-sampling method and oversampling method are not combined, so that the advantages of the two methods cannot be comprehensively utilized.

1.2 Ensemble learning method

Ensemble learning can obtain better generalization ability than a single classifier by combining multiple classifiers. At present, ensemble learning methods can be divided into two categories: serial serialization method and parallelization method. Among the serial serialization methods represented by Boosting class methods, the most famous ones are AdaBoost, gradient boosting tree and so on. The representative methods of parallelization methods are Bagging, Random Forest and so on.

The advantage of Bagging method is to reduce the variance of the model, while Boosting method is mainly to reduce the deviation. The combination of different integration strategies can gain stronger generalization ability.

2 Ensemble model of adaptive clustering mixed-sampling

Ensemble learning method has good generalization ability and can improve the accuracy of customer churn prediction. However, the classification ability of ensemble learning method is reduced by the imbalance among class, inter-class imbalance and overlapping distribution of customer churn data. In order to improve the classification ability of churn prediction model based on ensemble learning under imbalanced data, this paper combines clustering under-sampling

method, data cleansing method, oversampling method and ensemble learning method to deal with imbalanced data, and proposes an ensemble learning method based on cluster-based mixed-sampling (CUS-Ensemble), to alleviate imbalance among class, inter-class imbalance and overlapping distribution of customer churn data. Based on Bagging, CUS-Ensemble uses gradient boosting tree as the classifier. When training gradient boosting tree, the mixed-sampling method based on clustering is used to balance data categories and improve its learning ability for majority class instances. The framework of customer churn integrated prediction method based on cluster-based mixed-sampling is shown in Figure 2.

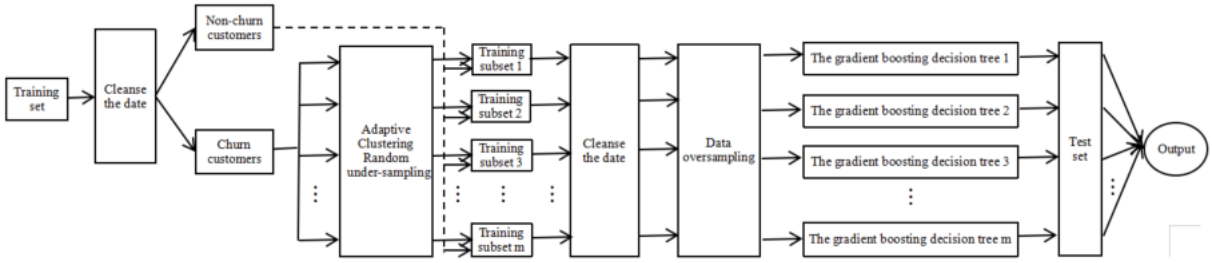

Figure 2 CUS-Ensemble Flow chart

2.1 Data cleansing

Data cleansing refers to clearing samples near the decision boundary, so as to maintain the differentiation of different types of samples. In the process of analysing customer churn data, we find that there are overlapping areas between churn customers and non-churn customers, which aggravates the difficulty of classifying imbalanced customer churn data. Therefore, the data cleansing module is added to CUS-Ensemble, and the non-churned customers near the churned customers are removed by using Edited Nearest Neighbours (ENN) method.

As you can see in Figure 2, CUS-Ensemble contains two data cleansing modules. The first is to cleanse the whole training set from the whole level, and the second time is to cleanse the balanced dataset after adaptive clustering and under-random and under-sampling from the local level. Among them, the first one is greater, and the second one is smaller.

2.2 Adaptive clustering random under-sampling

Cluster-based under-sampling method has been proved to maintain the distribution of initial data after sampling and improve the quality of under-sampled samples. However, setting the number of clusters as the number of minority class instances cannot guarantee the best number of clusters for different datasets, especially when the number of minority class instances is large, it is easy to increase the complexity of the algorithm, which is not suitable for large-scale customer churn data. Therefore, in the clustering random under-sampling method of this paper, firstly, the contour coefficient method is used to adaptively determine the appropriate number of clusters, and then the clustering is carried out under the guidance of this method. Hypothetical training set $D=\{(x_1,y_1)$,

$(x_2,y_2),...,(x_n,y_n)\}$, among them, the large class sample set is *N*. The minority class instances set is *P*. The random under-sampling process based on clustering is as follows:

Input: Imbalanced training set *D*.

Output: Balanced training subset *T*.

(1) Identify majority class instances set *N* and majority class instances set *P in D* set.

(2) Calculate the number of clusters *k*. Using contour coefficient method to search the best number of clusters in a specified range *k*.

(3) Initialize the cluster center. Order t=0, from *N* Random selection in *k* Sample points are used as the initial clustering center $m^{(0)}$= (m(0) 1,m(0)I, m(0)k).

(4) Cluster majority class instances. Alignment of class center $m^{(t)}$= (m(t) 1, ..., m(t)I, m(t)k), in which m(t)I is the center of class C(t)i. Calculate the distance from each majority class instances to the center of the class, and divide it into the nearest class center to form the clustering result $C^{(t)}$.

(5) Calculate the new class center. For clustering results $C^{(t)}$. Calculate the mean of samples in each current class as the new class center $m^{(t+1)}$=(m(t+1) 1, m(t+1)I, m(t+1)k).

(6) If the iteration converges or meets the stop condition, the clustering result C=$C^{(t)}$. Otherwise, make t=t+1 and return to step (3).

(7) Randomly select samples. According to *C* Sampling part of samples from each class in proportion to form a new majority class instances, and then combine the sampled majority class instances with the minority class instances *P, c*ompose a balanced training subset and then output.

2.3 Data oversampling

Data oversampling balances the categories of data by synthesizing new minority class instances. In CUS-Ensemble, the customer data after the second data cleansing will show two states: one is that non-churned customers are less than churned customers, and the other is that non-churned customers are more than churned customers. Therefore, it is necessary to oversample one kind of customer data and rebalance the categories of customer data. In order to avoid the overlapping between the generated data and the surrounding majority class instances, we choose to use Border-line SMOTE to synthesize samples.

2.4 Ensemble learning method based on adaptive clustering mixed-sampling

Many studies have shown that ensemble learning can improve the accuracy of customer churn prediction model, but the classification ability of imbalanced customer data is weak. In order to improve the classification ability of ensemble learning model on imbalanced data, this paper integrates the adaptive clustering mixed-sampling method into the training process of ensemble learning, and proposes an ensemble learning method CUS-Ensemble for customer churn prediction

under imbalanced data. The main steps are as follows:

Input: Imbalanced training set D; Base classifier G; Number of training rounds L.

Output: Forecast results $H(x)$.

(1) Extract majority class instances in N and majority class instances P *in* D.

(2) Calculate the optimal number of clusters k. Using contour coefficient method to search the best number of clusters in a specified range k.

(3) Iterative training Tree L, l= 1, 2, …, L

(3.1) Cluster majority class instances. Using K-Means algorithm, the majority class instances are divided into k class.

(3.2) Randomly under-sample the clustering results. Assume that the number of majority class instances is $|N|$. The number of minority class instances is $|P|$, The number of samples in class C_i is $|C_i|$, Number of majority class instances num_i sampled in C_i is: $num_i = \alpha \times |C_i| \times \frac{|P|}{|N|}, \alpha \geq 1.0$. α is the adjustment coefficient of instances, and the quantity of majority class instances after sampling can be controlled by adjusting α.

(3.3) Construct training subset. After sampling, the majority class instances and minority class instances form a training subset T_l.

(3.4) Oversampling. Pair training subset T_l Oversampling is performed to generate a balanced training subset B_l.

(3.5) Base classifier Training. Balanced training subset B_l Train one on Gradient boosting tree $G_i(x,y)$.

(4) Model integration. According to the given integration strategy, combine the generated L Gradient boosting tree. Output prediction results:

$$H(x) = \frac{1}{L}\sum_{l=1}^{L} G_i(x)$$

.

The idea of CUS-Ensemble is similar to that of EasyEnsemble. EasyEnsemble uses random under-sampling method to balance the class distribution of training instances, and takes Adaboost as the base learner. The main improvements of CUS-Ensemble are as follows: first, mixed-sampling based on K-means clustering method avoids the lack of representativeness and overlapping of majority class instances when the internal distribution of majority class instances is imbalanced; Second, the more advanced gradient boosting tree is used as the base learner to further improve the model performance.

3 Experiments and results analysis

In order to evaluate the classification ability of CUS-Ensemble on imbalanced customer churn data, this paper selects six imbalanced customer datasets, and takes AUC value, Recall value and Precision value as evaluation indexes, and compares them with five imbalanced data classification methods based on ensemble learning: EasyEnsemble[15], RUSBoost[16], CUSBoost[19]Equilibrium random forest[23](Balanced Random Forest, BRF), SPE[24](Self-paced Ensemble).

3.1 Experimental data

The basic information of the six datasets used in the experiment is shown in Table 1. Among them, the dataset WSDM is the customer data of music streaming media, DUKE1 and KDD and other five datasets contain the customer data of telecom industry.

Table 1 Basic information of data

Dataset	Source	instances quantity	Characteristics number	Degree of imbalance
CELL1	Kaggle	31407	67	50.0
DUKE1		51306	171	54.5
DUKE2	DUKE	100462	171	54.6
DUKE3		151768	171	54.6
WSDM	WSDM2018	992931	16	14.6
KDD	KDDCup2009	50000	230	12.6

3.2 Evaluation Indicators

In the classification prediction of imbalanced data, because the majority class instances occupy a high proportion, even if the correctly predicted number of minority class instances is small, the high rate of correct classification can also be ensured, so the correct rate cannot be used as a performance measure of classification algorithms on imbalanced data. In this paper, AUC, Recall and Precision are used as the performance evaluation indexes of churn prediction model under imbalanced customer data. AUC is the area under the Receiver Operating Character (ROC) curve. The ROC curve takes False Positive Rate (FPR) as the horizontal axis and True Positive Rate (TPR) as the vertical axis. The larger the area under the ROC curve is, the better the prediction performance of the model will be. Precision means "the number of real lost customers in the prediction ", while Recall means "the predicted number of lost customers". According to the confusion matrix of customer churn prediction (Table 2), the calculation formula of Recall and Precision is as follows:

Table 2 Confusion Matrix

Actual status of customers	Predicted loss	Predicted non-loss
Drain	A	B
Non-drain	C	D

$$\text{Precision} = \frac{A}{A+C} \quad (1)$$

$$\text{Recall} = \text{True Positive Rate(tpr)} = \frac{A}{A+B} \quad (2)$$

3.3 Experimental design

In order to reduce the influence of randomness of dataset partition on the experimental results, all experiments in this paper adopt the way of 50-fold hierarchical cross-validation twice, and take the mean of ten experiments as the final experimental results. CUS-Ensemble and the five methods compared are all ensemble learning methods, and the number of base classifiers is an important factor affecting the prediction results. To be fair, the number of base classifiers of each method is set to 200.

3.4 Results analysis

Table 3 to 5 are the comparison of evaluation indexes of each algorithm on different datasets. Among them, the last column shows the average results of the corresponding algorithm on 6 datasets.

Table 3 AUC values of different methods

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
CUS-Ensemble	**0.6348**	**0.6468**	**0.6729**	**0.6702**	**0.7347**	0.9466	**0.7177**
EasyEnsemble	0.6173	0.6350	0.6580	0.6610	0.7313	0.9452	0.7079
BRF	0.6267	0.6195	0.6532	0.6478	0.7166	**0.9524**	0.7027
SPE	0.6139	0.6035	0.6437	0.6394	0.7189	0.9505	0.6950
RUSBoost	0.5673	0.5741	0.6092	0.6219	0.6879	0.8057	0.6443
CUSBoost	0.5693	0.5545	0.5746	0.5805	0.6552	0.8955	0.6383

Table 4 Recall values for different methods

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
CUS-Ensemble	**0.7554**	**0.8090**	**0.8056**	**0.8047**	0.7428	**0.9047**	**0.8037**
SPE	0.7423	0.7765	0.7871	0.7912	**0.7686**	0.7782	0.7740
CUSBoost	0.6619	0.6315	0.6811	0.6616	0.7097	0.8996	0.7076
EasyEnsemble	0.6446	0.6288	0.6457	0.6389	0.6861	0.8500	0.6824
BRF	0.5928	0.5660	0.6018	0.6040	0.6536	0.8918	0.6517
RUSBoost	0.4368	0.4064	0.4994	0.5395	0.6009	0.6552	0.5230

Table 5 Precision values for different methods

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
SPE	0.0244	0.0219	0.0234	0.0230	0.1163	**0.4463**	**0.1092**
EasyEnsemble	0.0274	**0.0264**	**0.0275**	**0.0281**	**0.1349**	0.3288	0.0955
BRF	**0.0286**	0.0256	0.0274	0.0276	0.1321	0.3173	0.0931
CUS-Ensemble	0.0254	0.0235	0.0247	0.0244	0.1255	0.2823	0.0843
RUSBoost	0.0266	0.0243	0.0261	0.0267	0.1263	0.2619	0.0820
CUSBoost	0.0229	0.0203	0.0211	0.0212	0.1159	0.2405	0.0737

From the data in the above table, it can be seen that CUS-Ensemble has a good performance on six datasets:

In terms of AUC, CUS-Ensemble has achieved the best results on five datasets except WSDM, and the mean of AUC on six datasets is higher than that of five methods such as EasyEnsemble, which is 4% higher on average.

In terms of Recall, although the value of CUS-Ensemble on KDD is 2.58% lower than that of SPE, it has achieved the best results on the other five datasets and has the highest mean on six datasets. Compared with the mean of EasyEnsemble and other five methods, the mean of CUS-Ensemble is 2.97%, 9.61%, 12.13%, 15.2% and 28.07% higher respectively.

In Precision, CUS-Ensemble is not good enough, SPE, EasyEnsemble and BRF have their own advantages, among which EasyEnsemble has achieved the highest value in four datasets. The main reason for the relatively low Precision value of CUS-Ensemble is its high Recall value, and Precision and Recall are a pair of evaluation indexes. In addition, SPE has the highest mean because its value on WSDM is much higher than that of other methods, but the performance of SPE on other five data is slightly lower than that of CUS-Ensemble.

3.5 Significance testing

Statistical hypothesis testing provides an important basis for comparing the performance of classification algorithms. For further comparison of the performance of CUS-Ensemble and Balanced Random Forest algorithms on six datasets, the performance of all algorithms is assumed to be the same, and then Friedman based on algorithm ranking is used for testing. The AUC comparison values of each algorithm are shown in Table 6, and the Recall comparison values are shown in Table 7. Order N represents the number of datasets, k represents the number of algorithms, γ_i indicates the mean of i algorithm, γ_i obey normal distribution, then the variable

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \quad (3)$$

Among them,

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}\gamma_i^{\,2} - \frac{k(k+1)^2}{4}\right) \quad (4)$$

If the hypothesis is rejected, Nemenyi is used to further test each algorithm. The critical range of Nemenyi test is:

$$q_\alpha\sqrt{\frac{k(k+1)}{6N}} \quad (5)$$

Table 6 AUC comparison ridit of algorithm

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
CUS-Ensemble	**1**	**1**	**1**	**1**	**1**	3	1.333
EasyEnsemble	3	2	2	2	2	4	2.500
BRF	2	3	3	3	4	**1**	2.667
SPE	4	4	4	4	3	2	3.500
RUSBoost	6	5	5	5	5	6	5.333
CUSBoost	5	6	6	6	6	5	5.667

Table 7 Recall comparison ridit of algorithm

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
CUS-Ensemble	**1**	**1**	**1**	**1**	2	**1**	**1.167**
SPE	2	2	2	2	**1**	5	2.333
CUSBoost	3	3	3	3	3	2	2.833
EasyEnsemble	4	4	4	4	4	4	4.000
BRF	5	5	5	5	5	3	4.667
RUSBoost	6	6	6	6	6	6	6.000

Table 8 Precision comparison ridit of algorithm

Method	Dataset						Average
	CELL	DUKE1	DUKE2	DUKE3	KDD	WSDM	
SPE	5	5	5	5	5	**1**	4.333
EasyEnsemble	2	**1**	**1**	**1**	**1**	2	**1.333**
BRF	**1**	2	2	2	2	3	2.000
CUS-Ensemble	4	4	4	4	4	4	4.000
RUSBoost	3	3	3	3	3	5	3.333
CUSBoost	6	6	6	6	6	6	6.000

From the data in Table 6, it can be calculated that τ_F= 23.647, then search the critical value table of F : its critical value is 2.603 when outweigh α = 0.05, so the hypothesis is rejected and subsequent test is carried out. According to equation (8) and table searching, the critical value range can be calculated to be equal to 3.078. Combined with the average ridit in table 8, it can be seen that the difference between CUS-Ensemble, EasyEnsemble, BRF and SPE does not exceed 3.078, but the difference between RUSBoost and CUSBoost exceeds 3.078. Therefore, according to the test results, it can be considered that when AUC is used as the evaluation index, the performance of CUS-Ensemble is not significantly different from that of BRF, EasyEnsemble and SPE, but significantly different from that of RUSBoost and CUSBoost.

Similarly, according to the data in Table 7 and 8 and the above formula, it can be seen that:

In terms of Recall, the performance of CUS-Ensemble is not significantly different from SPE, CUSBoost and EasyEnsemble, but significantly different from BRF and RUSBoost.

In terms of Precision, there is no significant difference in performance between CUS-Ensemble and SPE.

Generally speaking, CUS-Ensemble can improve the prediction ability of the model in imbalanced customer churn prediction, and the overall prediction performance is the best, which has important reference value for enterprises to formulate customer retention strategies.

4 Conclusion

The dataset of customer churn prediction belongs to typical imbalanced data. On such data, especially in the case of class-imbalance and overlapping distribution, the classification ability of ensemble learning method is insufficient. In this paper, an ensemble learning algorithm based on adaptive clustering mixed-sampling is proposed to improve the generalization ability of ensemble learning method in imbalanced customer churn prediction. The algorithm integrates clustering random under-sampling, data cleansing and oversampling into ensemble learning, which reduces the imbalance of training data, ensures the representativeness of sampled majority class instances, and alleviates the information loss of majority class instances caused by single random under-sampling. The experimental results on six imbalanced customer churn data show that, compared with five ensemble learning methods such as EasyEnsemble, this method has better performance in AUC, Recall and other evaluation indexes, which is helpful to improve the performance of customer churn prediction model based on ensemble learning under imbalanced data, and improve the customer retention rate for entrepreneurs. And it also has certain reference significance for related problems in other fields. Future study will focus on how to reduce the time complexity of the algorithm and further improve the recognition rate of churned customers.

Acknowledgements：

Project fund:Guangxi First-class Discipline Applied Economics Construction Project Fund(2022GSXKA01，2022GSXKA05); 2024 Guangxi Higher Education Institutions Young and Middle-aged Teachers' Research Basic Capability Improvement Project Fund（2024KY0667.

References:

[1] ULLAH I, RAZA B, MALIK A K, et al. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector[J]. IEEE access, 2019,7:60134-60149.

[2] KIGUCHI M, SAEED W, MEDI I. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest[J]. Applied Soft Computing, 2022,118:108491.

[3] IDRIS A, IFTIKHAR A, UR REHMAN Z. Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling[J]. Cluster Computing, 2019,22(3):7241-7255.

[4] WU Z, JING L, WU B, et al. A PCA-AdaBoost model for E-commerce customer churn prediction[J]. Annals of Operations Research, 2022:1-18.

[5] WANG Q, XU M, HUSSAIN A. Large-scale ensemble model for customer churn prediction in search ads[J]. Cognitive Computation, 2019,11(2):262-270.

[6] AHMAD A K, JAFAR A, ALJOUMAA K. Customer churn prediction in telecom using machine learning in big data platform[J]. Journal of Big Data, 2019,6(1):1-24.

[7] LU N, LIN H, LU J, et al. A Customer Churn Prediction Model in Telecom Industry Using Boosting[J]. IEEE Transactions on Industrial Informatics, 2014,10(2):1659-1665.

[8] LI W K, YANG X B. Customer Churn Prediction Model Based on Double Layer Fusion Structure.[J]. Journal of Chinese Computer Systems, 2020,41(08):1634-1640.

[9] ZHOU J, YAN J F, YANG L et al. Application of LSTM Ensemble Method in Customer Churn Prediction [J]. Computer Applications and Software, 2019,36(11):39-46.

[10] TSAI C, LIN W, HU Y, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019,477:47-54.

[11] AMIN A, ANWAR S, ADNAN A, et al. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study[J]. IEEE Access, 2016,4:7940-7957.

[12] ZHU B, PAN X, VANDEN BROUCKE S, et al. [J]. Information Sciences, 2022,609:1397-1411.

[13] XIAO J, ZHOU X, ZHONG Y, et al. Cost-sensitive semi-supervised selective ensemble model for customer credit scoring[J]. Knowledge-Based Systems, 2020,189:105118.

[14] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017:2380-7504.

[15] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008,39(2):539-550.

[16] SEIFFERT C, KHOSHGOFTAAR T M, Van HULSE J, et al. RUSBoost: A hybrid approach to alleviating class

imbalance[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2009,40(1):185-197.

[17] WONG M L, SENG K, WONG P K. Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain[J]. Expert Systems with Applications, 2020,141:112918.

[18] ZHU B, BAESENS B, VANDEN BROUCKE S K. An empirical comparison of techniques for the class imbalance problem in churn prediction[J]. Information sciences, 2017,408:84-99.

[19] RAYHAN F, AHMED S, MAHBUB A, et al. CUSBoost: cluster-based mixed-sampling with boosting for imbalanced classification[C]// 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). Bengaluru, India: IEEE, 2017:1-5.

[20] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017,409:17-26.

[21] ZHOU Q, YAO Z, SUN B. Under-sampling Algorithm with Weighted Distance Based on Adaptive K-Means Clustering[J]. Data Analysis and Knowledge Discovery, 2022,6(05):127-136.

[22] WANG L, LIU Y, LIU Z, et al. Clustering under-sampling weighted random forest for processing unbalanced data[J]. Application Research of Computers, 2021,38(05):1398-1402.[23] CHEN C, LIAW A, BREIMAN L. Using random forest to learn imbalanced data[EB/OL]. (2004-07-07)[2022-03-05]. https://digitalassets.lib.berkeley.edu/sdtr/ucb/text/666.pdf.

[24] LIU Z, CAO W, GAO Z, et al. Self-paced ensemble for highly imbalanced massive data classification[C]// 2020 IEEE 36th International Conference on Data Engineering (ICDE). Dallas, TX, USA: IEEE, 2020:841-852.