Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070410
ISSN 2523-2576 Vol. 7 Num. 4

Research on Personal Credit Risk Assessment Model Based on Proportional Bootstrap and Stacking

Shenghao Deng[1,*], Shengran Fu[2]

[1]School of Mathematics and Systems Science, Guangdong Polytechnic Normal University, Guangzhou, 510450, China
[2]School of Finance, Jilin University of Finance and Economics, Changchun, 130117, China
[*]Corresponding author

Keywords: Credit Risk, Imbalanced Sample, Bagging, Stacking

Abstract: The personal credit business is an integral component of the modern financial framework. The automation of personal credit evaluation and approval processes, facilitated by machine learning technologies, has the potential to significantly enhance the operational efficiency of financial services. Nevertheless, a critical challenge that must be confronted is the imbalance in sample proportions between defaulting and non-defaulting client categories. In light of this issue, the present study introduces a Bagging algorithm that incorporates proportional sampling and employs the stacking approach to reintegrate the Bagging model, aiming to augment the predictive capabilities of the model. This methodology serves a dual purpose: it mitigates the overfitting induced by imbalanced samples through strategic resampling, while also enhancing the accuracy and robustness of the prediction model via the application of model fusion techniques. Empirical data analysis corroborates that the proposed method outperforms several classical prediction models in terms of Area Under the Curve (AUC) scores and demonstrates superior robustness. Furthermore, since the base model within the Bagging algorithm is agnostic to the specific model, it allows for the fitting of flexible and varied connection functions. When the base model utilizes interpretable machine learning methods, it additionally enables the extraction of the significance of each credit feature in relation to the probability of default.

1. Introduction

In the throes of the national economy's rapid ascent, we observe a persistent expansion in the scope of Internet financial services. Concurrently, the personal credit market is undergoing accelerated growth, which has been accompanied by a steep rise in the demand for personal credit services. This surge has become an essential barometer for the enduring development of a broad spectrum of financial institutions. The paramount challenge that arises amidst this landscape is the objective and scientific appraisal of an individual's borrowing capacity while concurrently diminishing personal credit risk. Consequently, personal credit risk assessment has emerged as an imperative facet of the loan application decision-making process for diverse financial entities. The conventional personal credit evaluation processes often necessitate substantial human resource involvement and time investment, substantially impeding transactional efficiency and impairing the overall experience for all parties involved. Against the backdrop of progressive advancements in computing and data storage capabilities, a novel paradigm integrating machine learning techniques with credit risk assessment models has materialized. This approach has received widespread endorsement and is increasingly being implemented in practice.

In recent scholarship, the application of machine learning models to personal credit assessment has garnered considerable research attention. For instance, Zhou Yuping and Chen Guanyu (2018) [1] scrutinized the theoretical feasibility of machine learning applications in personal credit risk assessment and constructed a corresponding framework predicated on such methodologies. Tao Zhourong (2023) [2] employed a linear weighting strategy to amalgamate individual models into a composite model that augments classification performance. Bai Pengfei et al. (2017) [3] utilized support vector machines, random forests, and XGBoost to independently construct credit prediction models, subsequently fusing them through a voting-weighted mechanism. Their empirical analyses indicated that the voting ensemble model outperformed singular models in terms of comprehensive performance. Liao Wenxiong et al. (2020) [4] introduced an XGBoost feature selection method, XGBFS, grounded in the Embedded concept, aimed at reducing user credit data dimensions to achieve effective user credit risk assessment. Zhang Liying and Yang Ruojin (2022) [5] incorporated the Stacking model to amalgamate various models, enhancing credit risk assessment outcomes by integrating machine learning models into a two-tier learning system. Wang Yufei (2024) [6] proposed a hybrid model based on the Stacking model's fusion methodology, demonstrating through comparative analysis that the ensemble model significantly improved predictive efficacy regarding credit default over singular models.

However, preceding personal credit risk assessment systems have seldom addressed data imbalance prior to modeling. Some classical classifiers are premised on the assumption of fundamental balance within the data sample space; however, they struggle to attain optimal accuracy rates for minority class classification when confronted with imbalanced samples. To address this imbalance, this paper adopts a Bagging algorithm [7] complemented by proportional sampling and synergizes it with the stacking concept [8] to reconstitute the Bagging model, thereby optimizing its predictive capabilities. This methodology aims to alleviate overfitting induced by unbalanced samples through proportional sampling and enhances the model's accuracy and robustness via model fusion techniques. Empirical results indicate that our proposed method surpasses traditional prediction models in terms of AUC scores and model stability. Moreover, the base model within the Bagging algorithm offers flexibility in adapting to various connection functions, and when utilizing interpretable machine learning techniques for the foundational model, it becomes feasible to discern the influential weights attributed to each credit feature in relation to default likelihood, offering enhanced clarity and more precise guidance for business decisions that require meticulous examination.

2. Personal Credit Risk Assessment Model Based on Ratio Bootstrap and Stacking

2.1 Establishment of the Assessment Index System

The construction of a personal credit assessment index system is pivotal for the development of an effective assessment model. The index system utilized in this study primarily encompasses the customer's fundamental personal information and loan details. Each information entry comprises 20 characteristic indicators, which are significantly correlated with the assessment of personal credit levels and can be utilized as input variables for training and analysis. These indicators include age, occupation, marital status, education, default history, housing situation, loan status, contact information, call duration, loan timing, days since last contact, number of marketing campaigns, contacts from previous marketing efforts, outcomes of previous marketing activities, employment change rate, Consumer Price Index, Consumer Confidence Index, loan interest rate, number of employed individuals, and subscription status (i.e., whether a purchase was made or not), among others. Table 1 presents the specifics of these indicators.

Table 1: Assessment Indicators and Their Descriptions

	Indicator	Description
Basic personal information	age	Age
	job	Occupation
	marital	Marital status
	education	Education
	default	Default
	housing	Housing
Loan information	loan	Loan status
	contact	Contact information
	duration	Call length
	month day_of_week	Loan time
	pdays	Days since the last contact
	campaign	Number of marketing activities
	previous	Number of contacts from previous marketing activities
	poutcome	Results from previous marketing activities
	emp_var_rate	Rate of change in employment
	cons_price_index	Consumer Price Index
	cons_conf_index	Consumer Confidence Index
	lending_rate3m	The interest rate of the loan
	nr_employed	The number of people employed
	subscribe	Whether or not to buy

Through comprehensive examination and analysis of these indicators, it is feasible to assess an individual's credit capacity more thoroughly. This aids financial institutions in identifying potentially risky customers, thus providing a more objective and scientific foundation for making precise and stable credit risk decisions.

2.2 Modeling Principles

2.2.1 Bagging and Stacking

The Bagging algorithm, initially introduced by Leo Breiman in 1996, is a parallel ensemble learning technique. It employs a resampling strategy for the training set, where random sampling with replacement is conducted to generate multiple distinct subsamples as subsets of the training data. Subsequently, numerous base learners are trained on these subsets independently, and then they are aggregated. When combining the predictions, a simple majority voting approach is typically utilized for classification tasks, while a simple averaging method is applied for regression problems. The Bagging algorithm reduces the variance of the base classifier models and enhances the overall generalization capability of the model through averaging or voting mechanisms. Its fundamental workflow is depicted in Figure 1.

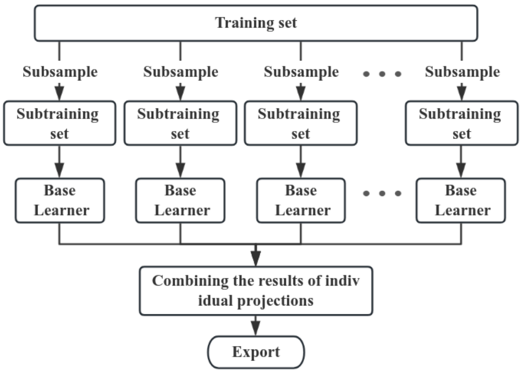

Figure 1: Flowchart of Bagging Algorithm

Stacking represents a more sophisticated ensemble learning concept that involves fusion techniques. An initial training set is used to train a collection of base learners, after which the outputs from several base learners are treated as a novel training set. This is then input into a meta-learner for further training, with the final output being the resultant prediction. In other words, the secondary training set is derived from the primary learners; however, direct use of the primary learners' outputs to form the secondary training set often leads to model overfitting. To mitigate this, methods such as K-fold cross-validation or the leave-one-out approach are commonly implemented. These strategies utilize the samples unused in the primary learner's training as the training samples for generating the secondary learner within the primary training set. The underlying idea flow is illustrated in Figure 2.

2.2.2 Personal Credit Risk Assessment Model Based on Proportional Bootstrap and Stacking

Building upon the Bagging algorithm and stacking concept, and integrating proportional sampling of the dataset $\{X_i, Y_i\}_{i=1}^{N}$, a model fusion technique is employed to merge stacking with a weighted Bagging model. The construction process is depicted in Figure 3, with the following sequential steps:

Step 1: Randomly sample the training set data and utilize the subsamples to fit and generate multiple base classifier models $M_1, M_2, M_3, \dots, M_k$;

Step 2: Assess and test the performance of these models on the validation set, and compute the AUC score AUC_{M_i} for each classifier model, which serves as an index for weighting the model;

Step 3: Introduce the test set data into the model for weighted prediction to obtain the weights $W_{M_1}, W_{M_2}, W_{M_3}, \dots, W_{M_k}$ for each sub-classifier model.

Step 4: On the validation set, input the sub-samples into the weighted classifier model for prediction, and derive its predicted outcome Y_{weight}.

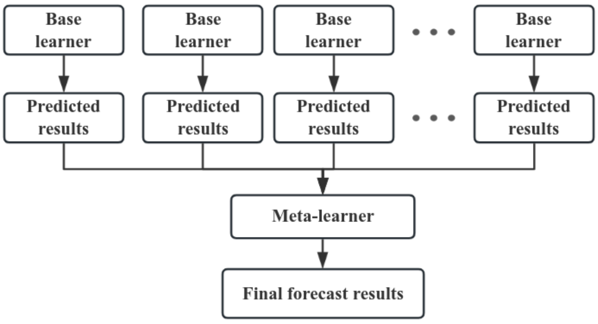

Figure 2: Flowchart of Stacking Idea

Step 5: Utilize the probability $P(Y = 1|X)$ predicted as 1 in the previous step as a novel training set, subsequently construct a new classifier model for fitting training, output, and complete the final model M_{final} construction.

Step 6: Employ the ultimately constructed model M_{final} to predict and analyze the test set data.

This model synergistically combines proportional sampling, the stacking concept, and a weighted Bagging model through model fusion techniques, effectively enhancing the generalization ability and predictive accuracy of the model. By generating multiple base classifier models on the training set and calculating weighting indexes based on their performance on the validation set, the strengths of each sub-classifier model are better harnessed to improve the overall model's performance. This adaptability to diverse data types and problems enhances the model's stability and precision.

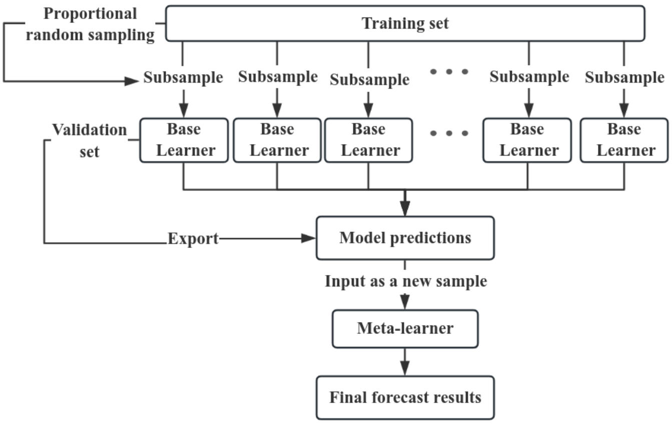

Figure 3: Flowchart of Credit Risk Assessment Model

3. Data Analysis

3.1 Data Source and Preprocessing

The dataset employed in this study is sourced from the AliCloud Tianchi platform, which contains a banking dataset that holds essential customer personal information as well as loan status data. This information facilitates predictions regarding a customer's eligibility for a loan. The dataset includes fields such as age, occupation, marital status, education, default history, housing situation, loan status, contact details, loan duration, number of marketing campaigns, employment change rate, Consumer Price Index, Consumer Confidence Index, loan interest rate, and purchase intent. This particular dataset is frequently utilized for credit scoring, risk management, and marketing analysis purposes. By analyzing a customer's personal details and loan history, it is feasible to assess their credit score and forecast their future repayment behavior and default risk. Additionally, the data can be leveraged for targeted bank marketing initiatives, including the identification of potential customer groups and the development of marketing strategies. The total dataset comprises 22,500 entries, with 19,548 non-defaulting customers and 2,952 defaulting customers, resulting in an imbalanced sample ratio of approximately 6.6:1 non-default to default. Table 2 and Figure 4 provide further details on these imbalanced sample categories.

Table 2: Imbalanced sample categories

Sample	Number	Percentage (%)
Non default sample	19548	86.88
Default sample	2952	13.12

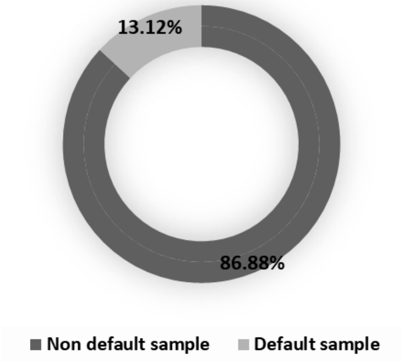

Figure 4: Imbalanced samples

In terms of credit risk indicator values, several variables are implicated, each with distinct characteristics such as nature, scale, and order of magnitude. To address these disparities, this paper will subject the characteristic indicators to standardization. For a given indicator with an eigenvector $X_i = \left[X_{i1}, X_{i2}, X_{i3}, \ldots, X_{in}\right]$, the standardization process is expressed as:

$$x_{ij} = \frac{x_{ij} - \overline{x_i}}{s}, \; j = 1, 2, 3, \ldots \tag{1}$$

Where $\overline{x_i} = \frac{\sum x_{ij}}{n}$ and $\mathrm{s} = \sqrt{\frac{1}{\mathrm{n}-1}\sum\left(x_{ij} - \overline{x_i}\right)^2}$.

3.2 Experimental Results

In this study, we propose a personal credit risk assessment model that harnesses the power of proportional bootstrap and stacking. The model employs a decision tree as the primary base classifier and logistic regression as the secondary classifier within the fusion model framework. By engaging in a proportionate sampling of imbalanced data, various ratios of $K_1 : K_2$(non-default samples to default samples) and $n_1 : n_2$ are considered. Consequently, six distinct models — W_stacking, Decision Tree (DT), K Nearest Neighbors (KNN), XGBoost, Random Forest (RF), and AdaBoost — are constructed. These models are evaluated based on the mean and standard deviation (SD) of the AUC scores obtained from different data proportions. The results are presented in Tables.3 through 7.

Table 3: Performance of each model at different ratios of $n_1 : n_2$ with $K_1 : K_2 = 1:1$

Model	$K_1 : K_2 = 1:1$									
	$n_1 : n_2 = 40:1$		$n_1 : n_2 = 20:1$		$n_1 : n_2 = 10:1$		$n_1 : n_2 = 4:1$		$n_1 : n_2 = 2:1$	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
W_stacking	0.9065	0.0485	0.9100	0.0592	0.8492	0.0319	0.8896	0.0159	0.9439	0.0099
DT	0.5793	0.0570	0.5865	0.0782	0.6398	0.0426	0.7485	0.0227	0.8448	0.0145
KNN	0.7422	0.0802	0.7238	0.0829	0.6458	0.0365	0.7072	0.0235	0.8528	0.0153
XGBoost	0.8804	0.0684	0.8883	0.0875	0.8090	0.0509	0.8569	0.0265	0.9249	0.0153
RF	0.8410	0.0809	0.8449	0.0873	0.8229	0.0303	0.8757	0.0193	0.9388	0.0097
AdaBoost	0.7918	0.1353	0.8298	0.0456	0.5684	0.0686	0.6840	0.0177	0.8393	0.0134

By putting the same set of data into different models for prediction output, the performance of each model at different scale values is obtained, and in general, the higher the mean and the lower the standard deviation, the better the performance of the model. It can be concluded from the above table:

Table 4: Performance of each model at different ratios of $n_1{:}n_2$ with $K_1{:}K_2 = 1{:}2$

Model	$K_1{:}K_2 = 1{:}2$									
	$n_1{:}n_2 = 40{:}1$		$n_1{:}n_2 = 20{:}1$		$n_1{:}n_2 = 10{:}1$		$n_1{:}n_2 = 4{:}1$		$n_1{:}n_2 = 2{:}1$	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
W_stacking	0.9237	0.0433	0.8927	0.0390	0.8630	0.0303	0.8932	0.0179	0.9454	0.0094
DT	0.6011	0.0704	0.6135	0.0550	0.6428	0.0432	0.7470	0.0245	0.8459	0.0142
KNN	0.7323	0.0725	0.6687	0.0459	0.6503	0.0316	0.7041	0.0228	0.8536	0.0163
XGBoost	0.8839	0.0742	0.8789	0.0475	0.8122	0.0489	0.8530	0.0251	0.9222	0.0173
RF	0.8564	0.0741	0.8295	0.0545	0.8182	0.0346	0.8747	0.0187	0.9382	0.0102
AdaBoost	0.8129	0.1227	0.8155	0.0739	0.5612	0.0566	0.6791	0.0186	0.8408	0.0130

In W_stacking model, the mean value is higher and the standard deviation is lower, the performance is stable and excellent, and the performance is better than the other models under different scale values; Decision Tree model has relatively weak performance under each index, the mean value is lower and the standard deviation is higher, and the fluctuation is larger; K Nearest Neighbor model has medium performance under most of the indexes, and the mean and the standard deviation are in the middle of the other models, and the performance is medium; The XGBoost and Random Forest models performed well in most cases, with high mean and low standard deviation, and good stability; the AdaBoost model performed well in terms of mean, but had a high standard deviation and high volatility. Considering the mean and standard deviation of the model's AUC score, it can be concluded that the W_stacking model performs more stable and better compared to the other models.

Table 5: Performance of each model at different ratios of $n_1{:}n_2$ with $K_1{:}K_2 = 1{:}4$

Model	$K_1{:}K_2 = 1{:}4$									
	$n_1{:}n_2 = 40{:}1$		$n_1{:}n_2 = 20{:}1$		$n_1{:}n_2 = 10{:}1$		$n_1{:}n_2 = 4{:}1$		$n_1{:}n_2 = 2{:}1$	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
W_stacking	0.9275	0.0403	0.9009	0.0273	0.8765	0.0239	0.8908	0.0166	0.9444	0.0095
DT	0.5984	0.0677	0.6133	0.0471	0.6366	0.0401	0.7482	0.0262	0.8449	0.0140
KNN	0.7281	0.0756	0.6639	0.0499	0.6432	0.0367	0.7067	0.0253	0.8539	0.0141
XGBoost	0.8817	0.0600	0.8735	0.0537	0.8094	0.0520	0.8576	0.0228	0.9254	0.0168
RF	0.8494	0.0697	0.8322	0.0499	0.8194	0.0322	0.8743	0.0187	0.9398	0.0100
AdaBoost	0.7960	0.1381	0.8102	0.0839	0.5711	0.0650	0.6809	0.0178	0.8420	0.0128

Table 6: Performance of each model at different ratios of $n_1{:}n_2$ with $K_1{:}K_2 = 2{:}1$

Model	$K_1{:}K_2 = 2{:}1$									
	$n_1{:}n_2 = 40{:}1$		$n_1{:}n_2 = 20{:}1$		$n_1{:}n_2 = 10{:}1$		$n_1{:}n_2 = 4{:}1$		$n_1{:}n_2 = 2{:}1$	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
W_stacking	0.8851	0.0669	0.8664	0.0423	0.8282	0.0334	0.8833	0.0172	0.9439	0.0098
DT	0.5972	0.0637	0.6255	0.0549	0.6352	0.0370	0.7455	0.0251	0.8472	0.0159
KNN	0.7359	0.0725	0.6748	0.0536	0.6442	0.0350	0.7078	0.0253	0.8552	0.0155
XGBoost	0.8780	0.0820	0.8786	0.0374	0.7976	0.0548	0.8555	0.0260	0.9249	0.0172
RF	0.8449	0.0784	0.8341	0.0470	0.8165	0.0313	0.8758	0.0181	0.9416	0.0109
AdaBoost	0.8095	0.1191	0.8122	0.0827	0.5632	0.0601	0.6809	0.0199	0.8428	0.0139

Table 7: Performance of each model at different ratios of $n_1 : n_2$ with $K_1 : K_2 = 4 : 1$

Model	$K_1 : K_2 = 4 : 1$									
	$n_1 : n_2 = 40 : 1$		$n_1 : n_2 = 20 : 1$		$n_1 : n_2 = 10 : 1$		$n_1 : n_2 = 4 : 1$		$n_1 : n_2 = 2 : 1$	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
W_stacking	0.8903	0.0527	0.8570	0.0452	0.8243	0.0320	0.8798	0.0185	0.9403	0.0102
DT	0.6019	0.0669	0.6216	0.0478	0.6424	0.0437	0.7497	0.0250	0.8468	0.0147
KNN	0.7336	0.0857	0.6722	0.0523	0.6487	0.0314	0.7102	0.0252	0.8527	0.0143
XGBoost	0.8848	0.0647	0.8828	0.0406	0.8060	0.0478	0.8577	0.0242	0.9246	0.0167
RF	0.8461	0.0714	0.8361	0.0498	0.8207	0.0295	0.8743	0.0196	0.9402	0.0092
AdaBoost	0.8204	0.1180	0.8258	0.0610	0.5678	0.0662	0.6852	0.0181	0.8405	0.0115

4. Conclusion and Outlook

In the context of the rapid development of big data and computers and other related technologies, machine learning and other related credit risk assessment modeling methods have gradually emerged and have been widely recognized and applied. In this paper, taking the bank data set containing customers' personal basic information and loan situation as a sample, a personal credit risk assessment model based on ratio bootstrap and stacking is proposed, and the same data are used to predict the results of the decision tree, K-nearest neighbor, XGBoost, Random Forest, AdaBoost, which are the classic and the more advanced models at present, and carry out a comparative analysis, which improves the accuracy and stability of the model prediction results and effectively enhances the prediction performance of the model.

For the future research direction, firstly, the classifier model in the Bagging model adopted in this paper is relatively free, so whether it is possible to automate the selection of a suitable sub-model is an issue that still needs further research. Secondly, when this paper carries out model reintegration based on stacking idea, the selection of a new classifier model is very important, how to select a more appropriate classifier is a problem that still needs further research. Finally, this paper adopts proportional sampling when dealing with imbalanced samples, so how to choose a suitable proportion for sampling needs further research.

References

[1] Zhou Yuping, Chen Guanyu. Research on the Individual Credit Evaluation Based on the Machine Learning Method [J]. Financial Theory & Practice, 2019(12): 1-8.

[2] Tao Zhourong. Research on Combined Model of Personal Credit Risk Assessment [D]. East China Normal University, 2023.

[3] BAI Pengfei, AN Qi, Nicolaas Fransde ROOIJ, et al. Internet Credit Personal Credit Assessing Method Based on Multi —Model Ensemble [J]. Journal of South China Normal University (Natural Science Edition), 2017, 49(06): 119-123.

[4] Liao Wenxiong, Zeng Bi, Liang Tiankai, et al. Personal Credit Assessment Method for High-Dimensional Data [J]. Computer Engineering and Applications, 2020, 56(04): 219-224.

[5] Zhang Liying, Yang Ruojin. The Application Research of Huoseholds' Loan Default Prediction Model Based on Machine Learning [J]. Financial Regulation Research, 2022, (06): 46-59.

[6] Wang Yufei. Research on Credit Default Prediction Based On Stacking [D]. Lanzhou University, 2024.

[7] Cao Jie, Shao Xiaoxiao. Analysis of Personal Credit Evaluation Method Based on Information Gain and Bagging Integration Learning Algorithm [J]. Mathematics in Practice and Theory, 2016, 46(08): 90-98.

[8] Wang Ruijie, Bao Tengfei, Li Yangtao, Song Baogang, Xiang Zhenyang. Combined prediction model of dam deformation based on multi —factor fusion and Stacking ensemble learning [J]. Journal of Hydraulic Engineering, 2023, 54(04): 497-506.