Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 1
AIEA 2023

Study on User Satisfaction Prediction and Influencing Factors based on GBDT Modeling

Shuirou Zeng [1], Wenqi Yang [1], Zihao Wang [2], Qiang Zeng [3], Yuncheng Wang [3], Xuan Liu [1]

[1] School of Mathematics and Statistics, Sichuan University of Science & Engineering, Zigong 643000, China

[2] School of Information and Computing Science, Sichuan University of Science & Engineering, Zigong 643000, China

[3] School of Brewing Engineering, Sichuan University of Science & Engineering, Zigong 643000, China

Abstract. A deep understanding of the factors affecting the characteristics of user experience helps to promote the benign development of China's mobile social network services, while how to improve the satisfaction of China's user experience has become a major key to the development of China's network commercialization applications. This paper draws on the application of machine learning algorithms in other areas of classification and identification, replicated to the classification and identification of user experience satisfaction to promote the high-quality sustainable development of mobile networks, and then through the analysis of the various factors affecting user satisfaction, to provide a basis for decision-making, so as to achieve an earlier and more comprehensive enhancement of user satisfaction. This paper adopts SMOTE oversampling for the unbalanced data of mobile user scoring dataset, and then adopts factor analysis for the main factors affecting customer satisfaction, and carries out Bartlett's spherical test. After that, GBDT model is established for user satisfaction prediction, and robustness test is used to optimize the model parameters, and the results show that the AUC value of GBDT model prediction is about 0.99. The time complexity of the algorithm can be greatly reduced and is more acceptable to industry.

Keywords: User Experience; SMOTE Oversampling; Factor Analysis; GBDT Modeling.

1. Introduction

1.1. Background and Significance of the Study

The rapid development of mobile communication technology provides people with great convenience, and more and more people can not live without the convenience brought by mobile communication technology, at the same time, along with the continuous construction of the network, the network coverage is also becoming more and more perfect, especially in today's information transparency and product homogenization, the performance of customer satisfaction has become an important reflection of the operation of the market of major operators. In the digital transformation of customer satisfaction evaluation, the customer experience is made to empower business decisions, so that business decisions can truly serve customers and jointly promote the high-quality sustainable development of mobile networks.

Based on customer complaints, point-by-point resolution of issues affecting user experience is the traditional method of improving customer satisfaction. However, as the number of subscribers increases dramatically, the variety of mobile products becomes richer, and the needs of customers become higher and higher, it is difficult for the traditional method to effectively improve customer satisfaction. This study intends to provide a decision-making basis by analyzing the factors affecting user satisfaction, so as to achieve earlier and more comprehensive enhancement of user satisfaction.

1.2. Literature Review

In terms of user satisfaction, in 1965, Cardozo [1] and others first put forward the concept of customer satisfaction, and confirmed through experiments that the factors affecting customer satisfaction include the customer's effort to obtain the target goods or services, as well as the customer's willingness to obtain the goods or services; Yu Zemin (1993) believes that in the fierce competition in the market, insisting on the service first is the way to win customer satisfaction is the key, through the product tracking service, can be very good to improve the quality of service and product quality, improve user satisfaction. [2] Gong Daning et al. in the global 5G commercial development trend pointed out that the global 5G commercial network development is rapid, as of the end of the third quarter of 2020, China's 5G users were 104 million, occupying 83.9% of the global 5G users, but the number of 5G users compared to the size of China's population base is still not large, and there is still a lot of room for development in the new era of 5G-enabled economic and social transformation. [3]

How to improve the satisfaction of China's user experience has become a major key to the development of China's network commercialization applications. Drawing on the application of machine learning algorithms for classification and identification in other fields, replicating the classification and identification of user experience satisfaction to promote the high-quality and sustainable development of mobile networks may be a proven method.

2. Structural Optimization of Data

In this paper, a total of 12,452 mobile subscriber usage data, including voice services and Internet data services, were provided from China Mobile Communications Group Beijing. A large number of missing values are found in the original dataset during extraction, and a certain imbalance in the amount of positive and negative class sample data is also found. Therefore, the mobile user dataset is first normalized to improve the speed of the subsequent algorithm solution and the generalization ability of the model. For the unbalanced data, SMOTE oversampling and other treatments are used to optimize the sample data structure, and logistic regression algorithms are used to classify and identify the original data and the processed data such as SMOTE oversampling and evaluate their effects, respectively, to select the final sample data in this paper.

After that, in order to make the algorithm can treat all features equally, so before using the algorithm modeling needs to be normalized, feature variable values are scaled using formula (1).[4]

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Where, x the original feature value, x_{norm} is the value after feature scaling, $x_{\max}$ is the maximum value of the original feature value, $x_{\min}$ is the minimum value of the original feature value.

Without changing the characteristics of the data distribution, feature scaling compresses all the feature values in the data set to the same scale ([0 , 1]), which makes the machine learning algorithm able to take into account all the feature variables, simplifies the calculation, and greatly improves the convergence speed of the algorithm and the accuracy of model identification.

3. User Satisfaction Prediction based on GBDT

3.1. Model Evaluation Indicator System

The performance evaluation of data mining models can be finalized by confusion matrix to derive the model evaluation metrics. [5] T he classification confusion matrix, which evaluates the recognition effectiveness of the algorithmic model by calculating the precision (P) and recall (R) of the algorithmic model for classification and recognition is given by the following formula:

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN} \tag{2}$$

This paper identifies the potential user satisfaction prediction as a starting point, for users who do not have the conditions of the user experience network, should also be accurately identified so as to avoid disturbing this type of users, so this paper considers the recall rate and precision rate as equally important indicators, take the β value of 1, and use the F1 value scores to evaluate the classification and identification effect of the model [6-7]. the F1 expression is as follows:

$$F_1 = \frac{2 \cdot P \cdot R}{P+R} \tag{3}$$

3.2. GBDT Classification Prediction Model for User Satisfaction

The GBDT algorithm belongs to the classical gradient boosting algorithm, which can solve classification and prediction problems, and the computational process of the GBDT classification algorithm will be described below.

In the case of continuous variables in the dependent variable, you can not directly use each class to fit the residuals r_{mi}, this is due to the residuals are continuous values caused by the residuals, so the residuals numerical can solve this problem, that is, the GBDT algorithm's loss function transformed into an exponential loss function or a log-likelihood loss function, in fact, in the case of the loss function is a log-likelihood loss function Logistic regression log-likelihood loss function and the GBDT algorithm is similar. [8] In this paper, the GBDT model is applied to classify the prediction of voice service and Internet service, and the GBDT classification algorithm [9] is as follows:

Table 1. Algorithm 2: GBDT Classification Algorithm

$F_0(x) = f_0(x)$;

for $i = 1, 2, \ldots, M$ do;

    $M_r(i) \leftarrow 0 \text{for} r = 1..s, i : \sigma_r(i) \le 2^{j+1}$;

    //Calculating gradient g_i;

    for $k = 1, 2, \ldots, m$ do

        $M_r(i) \leftarrow 0 \text{for} r = 1..s, i : \sigma_r(i) \le 2^{j+1}$;

        $g_i(x^k) = -\frac{\partial E[L(y.F(x^k))]}{\partial F(x^k)}\Big|_{F(x^k)=F_{m-1}(x^k)}$;

    end for

    Calculating step length α_i· $\alpha_i = \arg min\ E[L(y, F_{i-1}(x) + \alpha_i g_i(x))]$;

    Calculating replacement amount. $f_i(x) = \alpha_i g_i(x)$;

    Renewal Function. $F_i \mid (x) = F_{i-1}(x) + f_i(x) = F_{i-1}(x) + \alpha_i g_i(x)$;

end for

Output: $F = F_M(x) = f_0 + \sum_{i=1}^{M} \alpha_i f_i(x)$

4. Analysis of Results

4.1. SMOTE Oversampling Optimization

Firstly, the data set is randomly divided according to 70% of the training set and 30% of the test set, in order to ensure that the test can be close to the real data situation, so only the structure of the training set is optimized, and the test set does not have any structural changes. For the unbalanced distribution of samples in the dataset, this paper uses SMOTE oversampling, SMOTEENN hybrid sampling and SMOTETomek comprehensive sampling to optimize the original data structure, respectively. [10] Figure 1 shows the number of training sets after optimization of the three sampling methods.

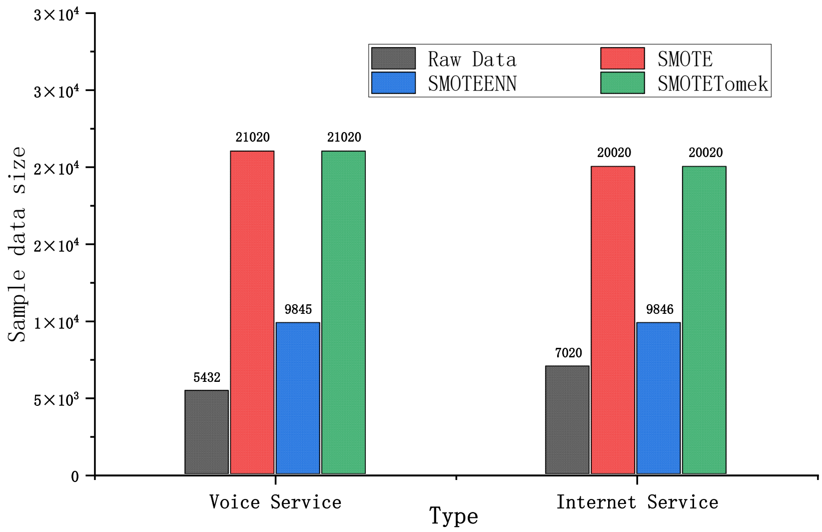

Figure 1. Number of training sets after optimization of sampling method

Next, the optimized classification effect of the oversampling approach is evaluated and also compared with the classification effect of the original data.

Table 2. SMOTE Classification effect of optimized logistic regression algorithm

Data Type	Typology	Accuracy	Recall Rate	F1 score	Weighted F1 score
Original	**voice service**	0.873	0.902	0.887	**0.817**
	surfing service	0.664	0.462	0.545	
SMOTE	**voice service**	0.951	0.794	0.865	**0.818**
	surfing service	0.509	0.841	0.634	
	surfing service	0.508	0.840	0.633	

As seen in Table 2, under the logistic regression classification algorithm, the recognition weighted F1 score of the SMOTE sampling optimized dataset is not comparable with the original dataset, but in the recognition of voice business class, its recall and F1 score are greatly improved compared to the original dataset.The optimized dataset with the SMOTE sampling method has a recognition recall of 0.545 in the classification of the Internet business class, and at the same time the F1 score is 0.634 compared to the original dataset, which is the best performance. the original dataset to 0.634, and the overall weighted F1 score is 0.818, which is the best performance.

It can be seen that the SMOTE sampling method can well improve the recognition performance of the logistic regression algorithm, which substantially improves the recognition effect of the voice business class while ensuring the overall recognition effect of the model as much as possible.

4.2. User Experience Influencing Factors

4.2.1. Extraction of User Satisfaction Factors

In the analysis of factors affecting user experience satisfaction, factor analysis [11] was used for feature extraction, and KMO and Bartlett's test are shown in Table 3.

Table 3. KMO and Bartlett's test

Name of test indicator		voice service	surfing service
KMO sampling and testing		0.764	0.835
Bartlett's test of sphericity	approximate chi-square	277172.191	162238.305
	degrees of freedom	741	1830
	significance	0.000***	0.000***
Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.			

In the above test results, it can be seen that the KMO value in the test of voice service and Internet service is 0.764 and 0.835 respectively, while the value of Bartlett's spherical test is equal to 0.000, which is less than 0.05, indicating that we are in the 95% confidence water, rejecting the original hypothesis, that is, we consider the data suitable for factor analysis. To this end, the main factors affecting user experience satisfaction are established as shown in Figure 2.

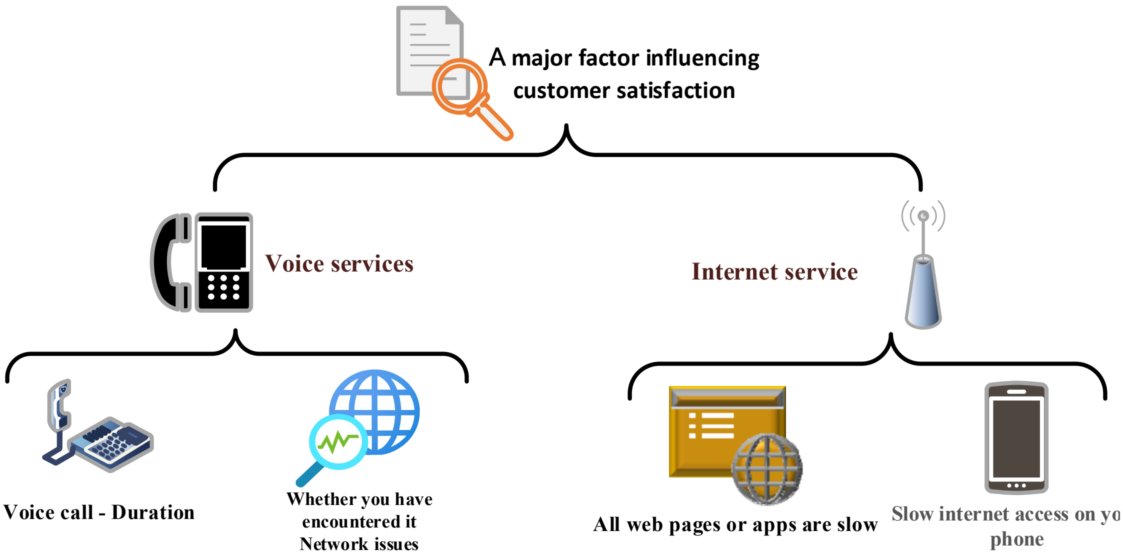

Figure 2. Key Factors Affecting Customer Satisfaction

4.2.2. User Experience Relevance Analysis

In the SMOTE sampling method to optimize the data, on the basis of which a total of 51 characteristic indicators in the voice business is considered, and a total of 125 characteristic indicators in the voice business, for this reason, the correlation analysis is used to rank the top eight items of the correlation coefficient of the factors affecting the customer's scoring. The heat map of Spearman's correlation coefficient and Kendall's correlation coefficient of the degree of influence of user satisfaction is shown in Figure 3.

As can be seen from Figure 3, the Spearman correlation and Kendall correlation of whether encountering network problems among the influencing factors of overall satisfaction in voice service present the highest positive correlation, while the correlation of cell phone Internet speed among the influencing factors of Internet service presents the highest negative correlation, then it is considered that whether encountering network problems and cell phone Internet speed have the highest degree of influence on user satisfaction.

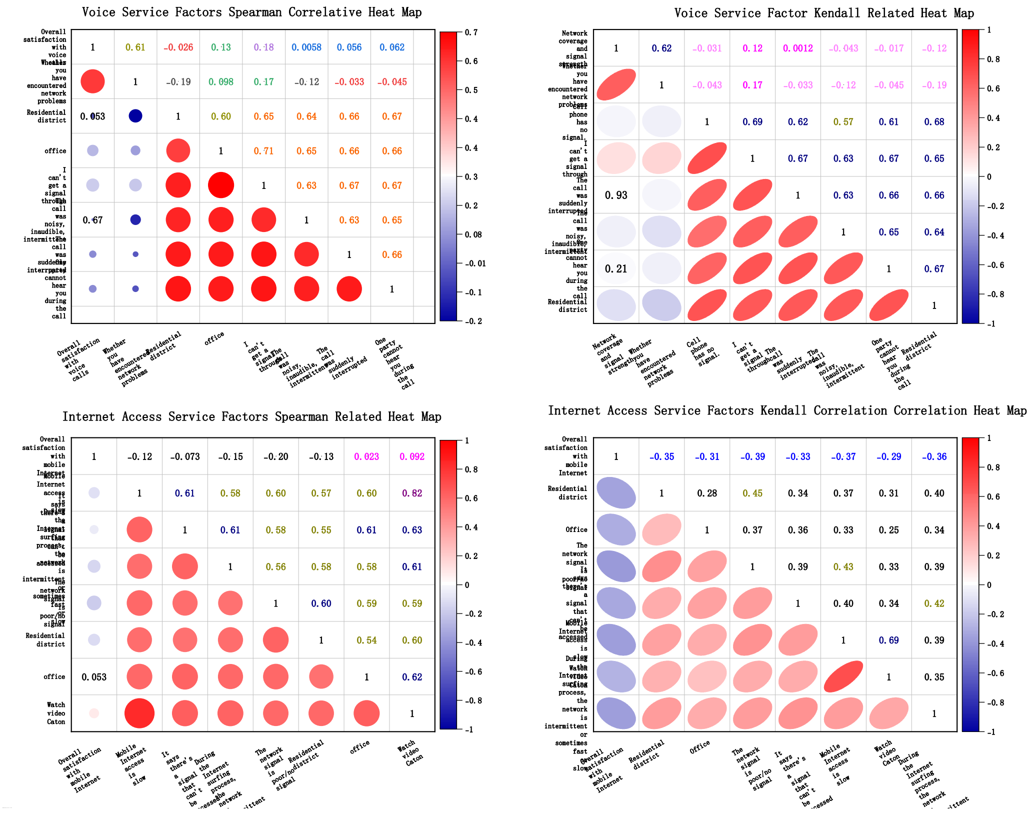

Figure 3. Heat map of the degree of influence of user satisfaction

In summary, in the process of analyzing the degree of influence on user experience satisfaction, the factor analysis method is used to extract the importance characteristic indexes, and then the correlation coefficients of the extracted characteristic variables are analyzed, and the degree of identification of the model presents a high degree of correlation and consistency, which lays the foundation for the construction of the identification model of user experience.

4.3. The Extent to Which the Model Recognizes the Data

Based on the above model, this paper solves the calculation of decision tree, GBDT, XGBoost and CatBoost fitting effects, and the scores between the models are shown in Table 4.

Table 4. The extent to which different models recognize the data

norm / learning device	**voice service**			**surfing service**		
	accuracy	recall rate	F1 value	accuracy	recall rate	F1 value
Decision Tree	0.839	0.717	0.815	0.839	0.817	0.861
GBDT	0.898	0.96	0.97	0.898	0.988	0.998
XGBoost	0.753	0.72	0.72	0.782	0.671	0.705
CaBoost	0.864	0.75	0.72	0.877	0.866	0.896

From the information in Table 4, it can be concluded that the F1 value of the GBDT model is the highest among the four classifiers, all of which reach more than 90%, which indicates that the GBDT model has a high generalization ability in terms of user satisfaction prediction. From the training effect it is found that the fused model works better than the strong learner alone.

4.4. Robustness Test of User Experience Prediction Results

In this paper, a linear combination prediction model with absolute error sum as the objective function is used. The decision tree, GBDT, XGBoost, CatBoost prediction results are optimized to improve the accuracy of the model.

Let the observation sequence of the object be $\{x_1, t = 1, 2, 3, \cdots, N\}$, and now use m single unbiased prediction methods for prediction. Let the prediction value of the i single prediction method at the moment of t be $\hat{x}_t^{(i)}$, then $e_t^{(i)} = x_t - \hat{x}_t^{(i)}$ is the prediction error of the i single prediction method at the moment of t , and let $w_1, w_2, \cdots, w_m$ be the weight coefficients of the m single prediction methods, respectively, and the weight coefficients are determined by the entropy weighting method [12]. Then the optimization problem is

$$\min Q = \sum_{t=1}^{N} |e_t| = \sum_{t=1}^{N} \left| \sum_{i=1}^{m} w_i e_t^{(i)} \right|$$

$$s.t. \begin{cases} \sum_{i=1}^{m} w_i = 1 \\ w_1, w_2, \cdots, w_m \geq 0 \\ e_t \in N_+ \\ e_t = \{1, 2, 3, \cdots, 10\} \end{cases} \tag{4}$$

The linear combination prediction model is used to optimize the prediction results, and Lingo is used to solve the calculations, and the optimization of its prediction results for each user satisfaction is shown in Fig. 4.

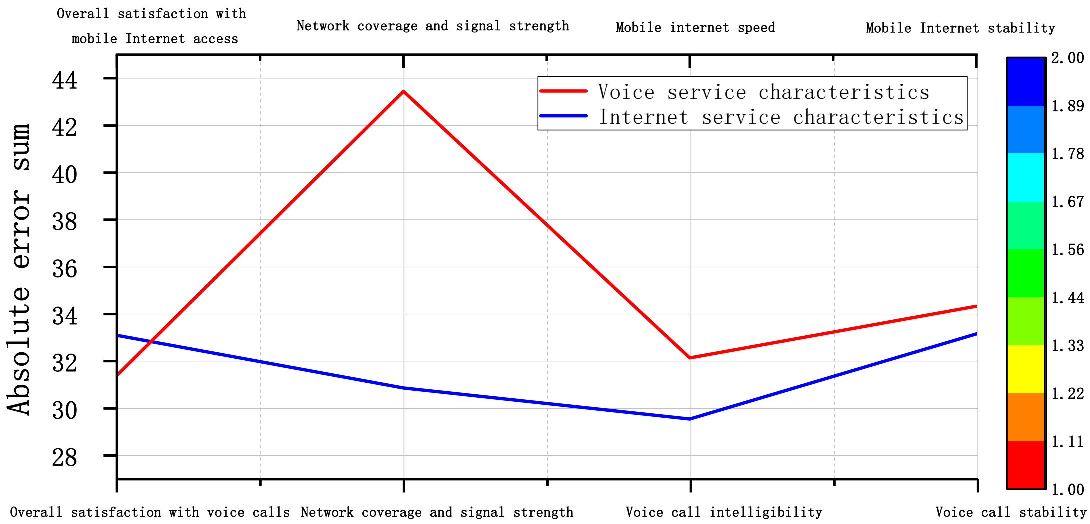

Figure 4. Optimization of user satisfaction prediction results

5. Conclusion

In this paper, when exploring the reasonableness and accuracy of the "user experience satisfaction prediction model", we optimize the data structure and training method according to its principle, adopt feature deflation for the sample data of user experience satisfaction, and optimize the transformed data with SMOTE oversampling. In the marginal process of dividing the threshold of high and low user satisfaction, factor analysis was used to determine the threshold of high and low user satisfaction on SMOTE oversampled data, and then based on the correlation analysis to extract the characteristic attributes of users with high and low satisfaction, the data showed that the impact of cell phone Internet speed on user satisfaction showed significance.

Finally, this paper constructs a user experience prediction model using multiple models (decision tree, GBDT, XGBoost, CatBoost), and optimizes the parameters using robustness test on this basis, while reducing the riskiness of model identification, improving the inadequacy of a single algorithmic model for user experience prediction, which is of some practical significance.

References

[1] Cardozo R N. An Experimental Study of Customer Effort, Expectation, and Satisfaction [J]. Journal of Marketing Research,1965,2(3):244-249.

[2] Yu Zemin. Maximizing user satisfaction[J]. Water conservancy and electric power machinery,1993(03):1-3.

[3] GONG Daning, WANG Xuemei, CAO Lei. Global 5G commercial development and trend outlook[J]. Information and Communication Technology and Policy,2020(12):7-10.

[4] S. Bernstein, R. Bernstein. Principles of Statistics. Upper volume, descriptive statistics and probability [M]. Science Press, 2002.

[5] Wang, X. M. Applied multivariate statistical analysis [M]. Shanghai: Shanghai University of Finance and Economics Press, 2021.

[6] Piantadosi, J.; Howlett, P.; Boland, J. (2007) "Matching the grade correlation coefficient using a copula with maximum disorder", Journal of Industrial and Management Optimization, 3 (2), 305–312.

[7] Maurice G. Kendall, “The treatment of ties in ranking problems”, Biometrika Vol. 33, No. 3, pp. 239-251. 1945.

[8] Quantum and classical machine learning for the classification of non-small-cell lung cancer patients[J]. SN Applied Sciences,2020.

[9] Liu L. Research on 5G potential user identification based on XGBoost optimization algorithm [D]. Jiangxi University of Finance and Economics,2022.

[10] Wang W. Research on the prediction of personal overdue risk of credit card of An Agricultural and Commercial Bank based on CatBoost model [D]. Southwest University of Finance and Economics,2020.

[11] WEN Xu, WANG Hao, HUANG Gang, et al. Identification method of abnormal bus load data based on factor analysis[J]. Journal of Chongqing University,2021,44(8):91-102.

[12] MA Key, HU Yi, LIU Jinquan. A statistical test method for robustness combining sample partitioning and system equations[J]. Systems Engineering Theory and Practice,2020,40(06):1371-1381.