Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3062

CREDIT MEASUREMENT OF RURAL INTERNET CONSUMER FINANCE BASED ON BLOCKCHAIN CLUSTERING AND FUSION MODELS

HaoYang Tan[1*], Lei Hu[2]
[1]*College of Economics, Hunan Agricultural University, Changsha 41000, Hunan, China.*
[2]*College of Finance and Statistics, Hunan University, Changsha 41000, Hunan, China.*
Corresponding Author: HaoYang Tan, Email: tlltwork@163.com

Abstract: This paper focuses on the empirical analysis of personal credit assessment of online lending platform from the perspective of personal credit, and the security of credit privacy data can be guaranteed by blockchain classification model. This paper is mainly based on the chain security encryption operation and decentralized data classifier training model, blockchain storage credit data between the ecological nodes through the transmission of transaction decision data return beacons, to achieve the data retrieval, use, confirm the rights and rewards, and at the same time the use of clustering learning algorithms combined with the decentralized training model to build a unique algorithmic training system, through the machine learning to backtrack all the transaction records, the sharing of the After data processing of credit data information, the fiducial correction fitting model using feedback from data samples, thus opening the modeling method of blockchain and clustering algorithm combined application in the field of credit. In the final analysis, the research on the application of blockchain technology in the credit collection industry should not stop at guaranteeing the security and traceability of data, but rather apply the "pre-credit review", "credit monitoring" and "post-credit management" to the entire credit collection industry. "Instead, it should be applied to the entire credit collection process, and used to guide Internet credit bureaus in their daily credit collection activities. Blockchain technology mainly solves the problem of credit trust and security, for this reason, it is necessary to construct a complete set of methods for analyzing, verifying and measuring Internet credit data. This paper combines the blockchain and the clustering algorithm in machine learning, and empirically analyzes the credit data of Internet consumer financial institutions under this framework.

Keywords: Blockchain; Clustering and Fusion Models; Credit Measurement; Rural Internet Consumer Finance

1 INTRODUCTION

Traditional classification models have their own advantages for data processing, and their applicability has been continuously proved by relevant research. However, each model is more or less insufficient for specific problems, and no one model can solve all problems well. Applying integrated learning to the analysis and processing of credit data before uploading, using integrated algorithms to integrate multi-party data processing, and uploading the model parameters of the model training phase to the blockchain and synchronizing them quickly, the blockchain-oriented clustering and fusion algorithms improve the security of the training phase and storage phase of the credit data, reduce the cost of the credit data storage and the transmission of the model parameters, and improve the security of the credit data. It reduces the cost of credit data storage and model parameter transmission, and improves the application system chain of credit data screening, analysis, storage and transaction synchronization. It also reduces the risk of credit data leakage due to model gradient update before uploading. The empirical analysis results in the later section show that it can improve the security of data and model while ensuring the accuracy of the model. Compared with the traditional data classification model, the blockchain clustering fusion model will improve the effect of Internet credit data analysis.

The blockchain clustering and fusion model utilized in this paper mainly starts from three aspects: credit data feature fusion, data segmentation and integrated learning. At the feature level, a new feature set is formed by extracting the social relationship information of credit users and quantifying the text description content information to supplement the basic features. The use of user soft and hard information at the same time promotes the richness and accuracy of the feature types of the dataset, which improves the quality of the dataset and makes the final assessment more realistic and reliable.

2 LITERATURE REVIEW

2.1 Blockchain Clustering Fusion Algorithm

K-means is one of the most classical clustering algorithms and is widely used in blockchain data analysis. The clustering center keeps changing with each iteration and eventually the center of the new round coincides with the previous round or is less than a certain threshold, then the clustering is complete. Numerous scholars have conducted studies to try to determine the optimal K value. Hongzhi Liu proposed an automatic fusion algorithm SCDP-MI based on credit data to automatically find the optimal number of fusions, which calculates the distance between the samples and stores them in a matrix, and then calculates the local density P of each data point as well as finds all the other points whose local density is greater than the point, and finds the minimum distance from these data points that satisfy the conditions to the point is set as &. After normalizing P and &, the 2D vector is mapped onto a 1D straight line, and finally the points on the straight line are sorted, and then the proper threshold is found to split the centroid of each cluster to get the optimal number of fusions, which is the K value[1].

The fusion algorithm clusters the data and divides the samples under different sample spaces. This improves the similarity of data within the same sample space, and classification models under the same sample space tend to get better classification results. The use of blockchain fusion analysis models and parameters to segment credit samples is the beginning of the problem, and the subsequent way of processing data under each sample space needs to be improved by other data processing techniques.

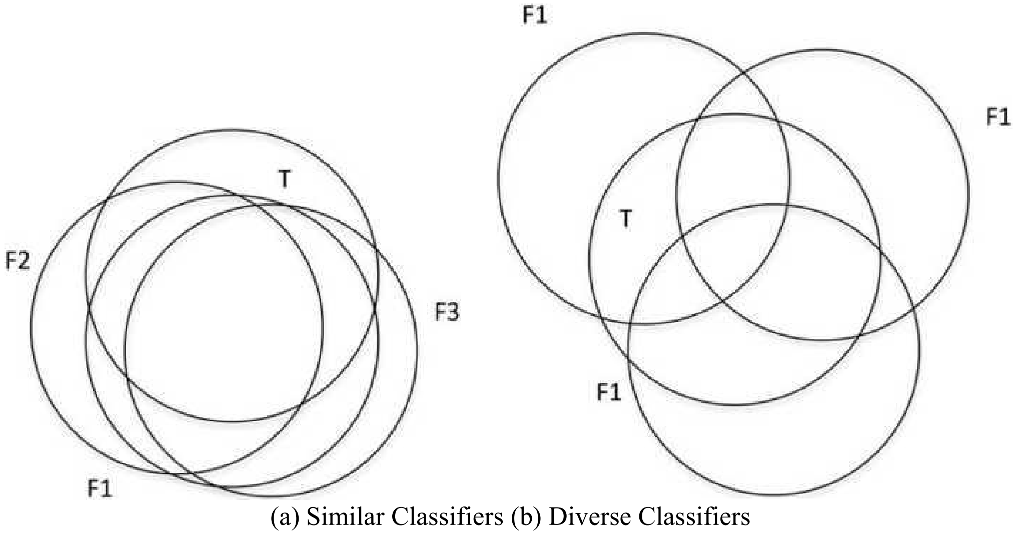
2.2 Application of Integrated Learning to Consumer Finance Data

Integration learning is a research hotspot in the field of machine learning, and the combination with blockchain technology can better handle massive credit data. Unlike ordinary machine learning algorithms that use only one classifier for decision determination, integrated learning generates and uses a set of classifiers, and then outputs the final decision result through a certain decision fusion method, thus obtaining better data generalization ability and classification accuracy.

The generation of base classifiers is the first part of integrated learning, the key is the overproduction of base classifiers that have diversity among each other. According to Fu Bin, in order to maximize the diversity among base classifiers and improve the classification performance of classifier combinations, researchers proposed the strategy of over product and select to assist in the design of multi-classifier combinations[2]. Obtaining different base classifiers can be done either by changing the data used during model training or by changing the parameters of the model itself.

Chekauer divided a dataset with 119 attributes into several feature subsets and manually trained the neural network model on different feature subsets while also varying the size of the neural network, obtaining 32 different combinations of neural network classifiers, with improved final classification results[3]. In addition, Turner conducted experiments using the same approach on a training set with 25 features and exceeded expectations by decreasing the classification effectiveness of the final classifier combinations even when only a small number of features were removed at a time, saving the vast majority of the features to be used to train the classifiers[4]. These two experiments illustrate that the method of manipulating feature values to train base classifiers achieves better results only on datasets with more feature values.

In order to get a better classification effect, many scholars have carried out research on how to select different feature values, such as through the random subspace method or the use of more complex genetic algorithms, as well as the diversity value of the combination of base classifiers as the objective function as a more direct method to select a subset of attributes. Among them, Pairwise Measures is a classical way to measure the diversity among classifiers, which is based on the principle of calculating the diversity values between any two classifiers in a classifier set, and then averaging the diversity values of all the combinations, and taking the average value as the overall diversity value of the set; whereas unpaired measures calculate the overall diversity of a number of base classifiers at one time. In comparison, the unpaired metric is slightly more efficient and not inferior to the paired metric in terms of performance [5].



(a) Similar Classifiers (b) Diverse Classifiers

Figure 1 Venn Diagram of the Concept of Classifier Diversity.

Kuncheva experimented with and summarized a variety of diversity measures, exploring the diversity of base classifier combinations under different datasets and the diversity of base classifier combinations generated using different strategies under the same dataset, recording the diversity values as well as the model accuracy. The results of this study conclude that the diversity of classifier combinations generated by Boosting method is greater than the diversity of classifier combinations generated by Bagging method, which is also related to the base classifier training strategy of Boosting method, which makes the differences between classifiers larger and complementary by changing the training sample set[6]. Experiments have confirmed that there is a correlation between diversity and accuracy, but there are also some empirical results that show that model accuracy is decreasing as diversity increases. It also shows that the current diversity metric is not perfect enough. When selecting the optimal subset of base classifiers, it is necessary to comprehensively consider various metrics and not only rely on the size of the diversity of the classifier combination.

3 SAMPLE SELECTION AND DESCRIPTIVE STATISTICS

3.1 Data Sources

In recent years, China's consumer finance market has transformed from the former traditional offline credit card model to the current Rural Internet Consumer Finance credit products, and credit consumption behaviors such as buying a house, education and training, and medical and aesthetic care have promoted the development of the Internet credit system. Therefore, the Internet credit data object used in this paper is Rural Internet Consumer Finance companies, and the analysis software is Weka, an open-source software based on JAVA environment, which is used to assist the research in the field of consensus algorithm and data fusion analysis.

There are many Rural Internet Consumer Finance platforms, but the degree of information disclosure varies from platform to platform, and this paper obtains the small amount of borrowing and lending data from Rural Internet Consumer Finance platforms from Jetson Consumer Finance Company (Homecredit). In addition, taking into account that borrowing and repayment is a cyclical process, it needs to go through a longer period of time in order to make a more realistic evaluation of the customer's credit situation, and it is not possible to discern the customer's goodness or badness immediately after the loan is issued to the applicant, so it is selected to pay back customers with a length of time of more than 12 months.

There are about millions of borrowing data in the whole Rural Internet Consumer Finance platform. According to the empirical data of sampling in mathematical statistics on the total number of samples and the number of samples sampled in random sampling, with a confidence level of 95%, when the overall sample is larger than 20,000 inches, it is sufficient to take 400 samples. Considering that the data of Rural Internet Consumer Finance is highly variable, as much data as possible is needed for modeling and validation, and, the larger the data sample size, the better the accuracy of modeling is supposed to be. Therefore, this paper obtains 24014 customers with repayment length of 12 months or more from the third-party platform as samples.

The information contained in the sample is: customer code, history of repayment, gender, education, age, marital status, whether he/she owns a house, whether he/she has a mortgage, whether he/she owns a car, whether he/she has a car loan, whether he/she has a credit card, whether he/she has a child, nature of the organization, where he/she resides, salary, years of working experience, type of occupation (whether he/she has a job or not), total amount of the loan, amount repaid in each installment, type of the borrowing, term of the loan, whether he/she has had a history of borrowing, etc.

The output of the model is to determine whether a customer is a "good customer" or a "bad customer". The data obtained in this paper is the history of the customer's repayment for each installment, so based on the customer's history of repayment, the customer is categorized as a "good customer" or a "bad customer".

Category 1, Bad Customers: Borrowers who are seriously delinquent within 12 months of the arrival of the loan will be determined to be bad customers. Serious delinquency is defined as having one installment of a loan that is more than 90 days delinquent or having two or more installments of a loan that are more than 60 days delinquent.

Category 2, good customers: those without the above are considered good customers.

After determining the discriminatory conditions of "bad customers" and "good customers", this paper categorizes the sample data according to the borrowers' historical repayment records. In this paper, the original sample data are screened and an equal number of good and bad customers are extracted. Because the number of bad customer samples is limited, so get all the 1594 bad customer samples, and then randomly selected 1594 good customer samples, the chaotic combination of these samples is a complete sample for analysis and modeling. The results of screening the samples are shown in Table 1.

Table 1 Percentage of Modeled Data Classifications

	good customer	proportions	bad customer	proportions
pre-sampling	22419	93.40%	1594	6.60%
post-sampling	1594	50%	1594	50%

3.2 Analysis and Selection of Characteristic Variables

Each sample contains the following information: gender, education, age group, marital status, whether or not they have a house, whether or not they have a mortgage, whether or not they have a car, whether or not they have a car loan, whether or not they have a credit card, whether or not they have children, nature of the organization, location of residence, salary, years of experience, type of occupation, total amount of the loan, amount of the loan repayment per installment, type of the borrowing, loan tenure range, and whether or not they have had any historical borrowing.

The following is an explanation of the terms used in the modeling process of this paper: "characteristic variable" is equivalent to the "independent variable" in the credit scoring model; "whether the customer is a bad customer/probability of bad customer" is the "dependent variable" in the credit scoring model. The "probability that the customer is a bad customer" is the "dependent variable" in the credit scoring model. Each piece of information in the sample may become a "characteristic variable" for modeling, so each piece of information is called an "alternative indicator". Alternative indicators are analyzed, and if an alternative indicator is effective in predicting the outcome, then that "alternative indicator" becomes a "characteristic variable", and vice versa.

3.2.1 Selection of alternative indicators

The selection of alternative indicators is crucial to the construction of the model, and the alternative indicators for personal loans of commercial banks with more mature development generally contain basic personal information, occupational information, assets and liabilities information, and so on.

If there is a significant difference in the credit behavior of customers between the values taken in the same alternative indicator, it can be filtered as a characteristic variable. For example, in the indicator of stable occupation or not, if the credit situation of people with stable occupation is better than that of people without stable occupation, and they can accurately distinguish good and bad customers with a high probability, this alternative indicator is considered to have a high degree of differentiation, and it can be a characteristic variable. The alternative indicators after grouping were analyzed using the indicator Information Value (III), which is commonly used in statistics. The IV values of the alternative indicators for each variable are shown in Table 2:

Table 2 Alternative Indicator IV Values

variable name	IV value
ivloanlife	1.371645554
ivpayamount	0.280030079
ivloanamount	0.265127812
ivlendhis	0.024939962
ivloantype	1.211117758
ivjob	0.001154599
ivworkyears	0.022754313

3.2.2 Modeling training set and validation set establishment

After defining the characteristic variables as well as the dependent variables of a credit scoring model, modeling and analysis can be performed. After the modeling is completed, it is necessary to objectively determine whether a model is valid or not, or to compare the advantages and disadvantages of different models. A common method is to sample the original samples as the "training set" and the remaining samples as the "validation set". The training set is used to solve the parameters of the model, and the validation set is used for simulation to check the correctness of the predictions. Since the training set for solving the parameters does not contain the information of the validation set, it is fair and reasonable that all models are tested with the validation set. Considering the sample size, a larger training set is needed for adequate training, so random sampling is used to draw 25% to be used as the test set and the remaining 75% as the training set. The sampling results are shown in Table 3:

Table 3 Number and Percentage of Samples in the Training Set and Validator

	No. of good clients	No. of bad customers	add up the total
training set	1182	1194	2376
Percentage within sample	49.75%	50.25%	100%
validation set	402	390	792
Percentage within sample	50.75%	49.24%	100%

The data used in this paper to empirically study the credit assessment model of Rural Internet Consumer Finance network credit comes from the open dataset of Gitzo (homecredit). This open data set in the form of a database consists of seven tables, and the relevant data can be directly imported into the database for the researcher to analyze and organize.

This dataset contains user borrowing records between 2019 and 2021, and all loan records during this time period have now been confirmed as repaid or in default, with a relatively complete record for analytical research. This database contains a wealth of personal information, including basic credit information, bidding information, loan application information and other hard information, but also includes relationship information with other users on the platform, guarantee relationship and friend relationship information and other soft information, in addition to the current market information and some macro-economic information, etc., the data cleaning and pre-processing of this table to obtain the original data used in this experiment. The original data set used in this experiment is obtained by data cleaning and preprocessing of this table. After the statistics, there are 20,106 records available for the experiment during the three-year period, of which 10,666 are default records and 18,440 are trustworthy records, and the ratio of default and trustworthiness is 1:1.73.

4 EXPERIMENTAL RESULTS

4.1 Cross-Validation Assessment

The processed Gitzo (homecredit) dataset is randomized into 5 copies, and 1 copy is selected in sequential order as the test dataset 2 and the remaining 4 copies are used as the training dataset. Then 5 training test dataset pairs are generated and used to perform 5-fold cross-validation analysis. All accuracy comparisons of model classification performance in the subsequent experimental results were performed using the mean value of the 5-fold cross-validation.

Table 4 Single Classifier 5-fold Cross-Validation Classification Accuracy

classification model	model parameter	accuracy (essential features)	accuracy (after feature fusion)	Categorization effect
Na Bayes	default (setting)	68.24%	63.26%	1
Random Forest	default (setting)	69.57%	69.74%	T
logistic	default (setting)	69.69%	71.21%	t
J48	C0.01	69.38%	69.82%	T
	c0.10	69.24%	67.70%	1
	C0.20	68.36%	64.84%	1
	default (setting)	67.86%	65.48%	1
	C0.30	67.01%	64.67%	1
	u	64.82%	63.14%	1
	R	68.73%	68.34%	1
BP	default (setting)	69.51%	69.61%	T
	H1	68.53%	67.47%	1
	H2	68.84%	70.23%	T
	H3	68.82%	69.58%	T
SVM	T0	69.42%	63.88%	1
	T2	69.67%	64.93%	1
	T3	60.52%	52.32%	1

As shown in Table 4, some classical single classification models provided by the Weka platform are used in the experimental process to first evaluate the credit assessment effect under the 15 basic features, and then use the more comprehensive credit data set with more comprehensive information obtained from feature fusion to evaluate the classification effect.

Analyzing the results of the experiment, it can be seen that with the dataset having 15 basic features, the accuracy reached 69.69%. The other classification results are slightly worse, but most of them reach more than 68%, which is overall favorable. Through blockchain feature layer fusion, a dataset with 115 features was obtained, including both hard and soft information. This more informative dataset was then used to evaluate the classification effect of the classical classification models mentioned above, and the experimental data showed that the Logistic model exhibited the highest classification accuracy of 71.21%.

Comparing the data from the two rounds of experiments before and after, it can be seen that the classification effect of SVM and NaiveBayes model decreases significantly, the classification effect of logistic regression model and random forest model rises, the classification effect of BP neural network model under some parameters is improved, and the accuracy of decision tree model is improved only when the confidence factor is very small.The performance of the NaiveBayes model and the support vector machine is very dependent on the quality of the features of the data set, so they do not incorporate some noise features in the analysis process. The performance of NaiveBayes model and Support Vector Machine is very much dependent on the quality of the features in the dataset, and they do not do any filtering of the features used in the analysis process, so the incorporation of some noisy features reduces the classification effect. In contrast, the logistic regression model and the decision tree model filter the noisy features by setting feature weights and pruning, respectively, which can differentiate between features of different quality. Combining the classification principles of each model and further analyzing the experimental results, it can be found that after the feature layer fusion, some of the newly incorporated 100 features are effective and can significantly improve the data quality, while some are useless features that will increase the misclassification rate of the model. Therefore, in order to achieve the best classification results, these noisy features need to be screened.

The best performing classification model in this round of experiments was the logistic regression model, which achieved the highest classification accuracy of 71.21% for a single classifier, and the improvement over the previous round of experiments (15 features) is shown in Table 5.

Table 5 Classification Accuracy of Logistic Regression Models

Different sets of features	Maximum accuracy
Basic features (15)	69.69%
Post-fusion features (115)	71.21%
Experimental gains	+1.52%
Different sets of features	Maximum accuracy

Credit soft information such as descriptive text and social relationships and hard information mined based on credit information expand the number of feature vectors in the credit dataset can complement the amount of credit data. The experimental results strongly suggest that if this new dataset is slightly filtered with features, then the classification effect of the model can be improved. In addition, the best single classifier for both rounds of experiments is the logistic regression model, which further validates the reason why logistic regression models are widely used in the field of credit assessment, which can better deal with the credit assessment problem and provide a certain degree of interpretability for the decision-making results by feeding back the model parameters. The above data analysis illustrates that in the field of Rural Internet Consumer Finance network lending, the soft information of borrowers will improve the information asymmetry phenomenon, and the feature fusion part of the model in this paper can improve the accuracy of credit assessment is clearly verified.

4.2 Assessment of the Effectiveness of Data Segmentation

The K-Means clustering algorithm is selected for data segmentation, and the K-value, an important parameter of the clustering algorithm, is assisted by the SCDP-MI algorithm. Then, several candidate K-values are further tested to find the optimal K-value that is suitable for the dataset. The following method is used here to evaluate the appropriateness of the K-value:

In the first step, a certain value of K is set to train the K-Means classifier, and then the training dataset is clustered in order to get K different sample spaces.

In the second step, a number of classifiers are obtained using different single classification models trained using data under each sample space.

In the third step, the test data is inputted, and the test data is divided onto different feature spaces using the above trained K-Means classifier, and then the sample is classified using the classifier that performs the best in that feature space, and the classification accuracy is finally counted.

4.3 Overall Effectiveness Evaluation

From the above experimental results, it can be learned that the credit assessment model based on credit data fusion proposed in this paper step by step improves the credit assessment accuracy rate of Rural Internet Consumer Finance online credit through the three major links of feature fusion, data segmentation and fusion learning. The overall classification accuracy of the assessment model is cross-validated with 5 folds, and the results are shown in Figure 2:

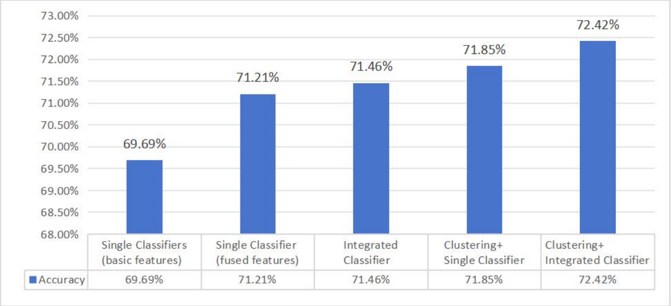


Figure 2 Multi-Classifier Clustering Fusion Effect

From the figure, it can be learned that after the feature layer fusion, the amount of data information has been greatly improved, and the evaluation effect of the optimal single classifier has been improved from 69.69% to 71.21%. After using the clustering algorithm to subdivide the data and then using the single classification model, the classification effect has been improved to some extent, and a classification accuracy of 71.85% has been achieved. On the basis of data segmentation, after completing the credit assessment of the samples under different sample spaces using the fusion learning method, the model classification effect reaches the best, and the accuracy is improved to 72.42%. After feature layer fusion, data segmentation, and decision layer fusion, the assessment effect of the information fusion-based assessment model proposed in this paper is improved from 69.69% to 72.42%, and the assessment accuracy is improved by 2.73%, which obtains a sizable improvement in classification effect.

Through the above analysis of the experimental data, not only a preliminary description of the classification effect of the model, but also concluded that: first of all, after clustering the training data set, a number of disjoint sub-training sets are obtained. It is worth noting that not all sub-training sets are suitable for fusion, for example, the effect of the fusion model made based on the differences in user features used in this paper is not as effective as the single classification model under the third type of feature space. Again, the selection of base classifiers should not only consider the degree of diversity of the combination, nor should it only consider the performance of the base classifiers, and the two need to be considered for screening to obtain the optimal set of base classifiers.

5 CONCLUSION

In this paper, we design a block fusion-based credit data evaluation model and empirically test it.According to the results, we recommend the following actions.

First, the "soft information" should be more fully utilized to improve the information asymmetry phenomenon at and to increase the amount of information in the information assessment dataset. In the context of the big data era, the explosive growth in the amount of information brings many opportunities for credit assessment. As people's use of the Internet and social platforms increases, personal information becomes more and more transparent, and it is increasingly important to make full use of the data from third-party websites, as well as to mine the user's social relationship information and analyze the text description information. Because this kind of "soft information" can alleviate the information asymmetry between lenders and lending platforms, and between lenders and investors, thus improving the accuracy of credit assessment, accelerating the speed and volume of transaction, completing the economic deployment more rationally, and improving the speed of social and economic development. The blockchain fusion model proposed in this paper realizes the extraction of various types of "soft information" in the decision-making layer fusion, including soft information based on text description and social relationship information based on social networks. This kind of "soft information" is rich in variety and comes from a wide range of sources, which can improve the amount of data information and mitigate the negative impact of information asymmetry in Rural Internet Consumer Finance lending.

Second, based on the fact that different credit users have different characteristics and preferences, borrowers are clustered and analyzed and divided into different categories. "Things are grouped together and people are divided into groups", and the same group of people tends to have certain commonalities. The process of cluster analysis further extracts such commonalities within groups and differences between groups, which is a complement to traditional population division indicators such as income level and education level. The design of different credit assessment models for different groups can improve the classification effect of the models, and the diversified use of models for different groups also makes up for the shortcomings of the traditional single model, which is too "middle-of-the-road".

This paper starts with the description of the data set, introduces the characteristics of the data set as well as the data processing methods, and explains how the data set required for the experiment was obtained from the original data set. Then, the experimental process as well as the experimental results are described in detail, explaining the reasons for the design of each step of the model and its benefits, and verifying the significance of the model for solving the personal credit assessment of Rural Internet Consumer Finance network credits.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research was funded by Scientific Research Fund of Hunan Provincial Education Department, grant number 21B0210; and project supported by Hunan Provincial Natural Science Foundation of China, grant number 2024JJ6255.

REFERENCES

[1] Zhang H, Li M, Wang P. Research on credit assessment model of Rural Internet Consumer Finance with blockchain technology support. Finance and Economic Research, 2023, 45(3): 120-135.

[2] Zhao Tingting, Liu Yang. Credit Risk Clustering Analysis of Rural Internet Consumer Finance Market Based on Blockchain. Journal of Management Science, 2024, 37(2): 34-51.

[3] Chen Chen, Ma Chao. Evaluation of the effectiveness of blockchain clustering fusion model in credit scoring. Electronics and Information Engineering, 2012, 44(6): 105-116.

[4] Zhou Y, Yang L, Sun L. Blockchain and machine learning fusion for credit prediction in Rural Internet Consumer Finance, Computer Applications Research, 2013, 40(4): 56-68.

[5] Sun Xian, Zhou S. Credit risk management in decentralized finance: the role of blockchain technology. Modern Economic Information, 2023, 31(1): 22-28.

[6] Zhou Shengdi, Wu Diqi. Convergence application of blockchain and big data technology in credit assessment. Financial Electronic, 2023, 6: 44-49.