Neural Computing and Applications (2021) 33:4065–4075
https://doi.org/10.1007/s00521-020-05489-z

S.I. : SPIOT 2020

# Two-stage adaptive integration of multi-source heterogeneous data based on an improved random subspace and prediction of default risk of microcredit

Anzhong Huang[1] · Fei Wu[2]

Received: 16 August 2020 / Accepted: 27 October 2020 / Published online: 11 November 2020


**Abstract**
Some scholars have shown that the machine learning methods based on a single-source data can successfully monitor the risks of formal financial activities, but not those of informal financial activities. This is because the data generated by formal financial activities, whether it is the structured or unstructured data, are of high quality and quantity, while the data generated by informal financial activities are not. Therefore, multi-source data are the key to monitor the risks of informal financial activities through machine learning. Although a few studies attempted to use multi-source data for financial risk prediction, they simply stack the obtained multi-source data, but ignore the original sources, heterogeneity, mutual redundancy and other characteristics of the data, so that the improvement of the prediction effect is not obvious. Therefore, TSAIB_RS method based on the two-stage adaptive integration of multi-source heterogeneous data was constructed in the paper, in which the data with different sources and different distributions were adaptively integrated. In order to test the reliability of TSAIB_RS method, the paper takes the default risk of microcredit in China as the test target and compares the prediction results of various test methods. It concludes that TSAIB_RS method can significantly improve the prediction effects.

**Keywords** Multi-source heterogeneous data · Adaptive integration · Microcredit risk

## 1 Introduction

Information asymmetry is the root cause of financial risks, and obtaining as much information as possible is the key to predict financial risks. As a result, some scholars put forward the problem of multi-source information in financial risk monitoring earlier [1, 2], which means that banks should not only use hard information (financial statement information), but also use soft information (financial statement information) to reduce credit risk.

However, soft information is often unstructured data, which cannot be used by statistics and econometrics, which are the traditional financial risk prediction methods. It greatly limits the improvement of financial risk prediction accuracy, because a large amount of information in the Internet era is unstructured data. Therefore, machine learning is an excellent supplement to the traditional methods of financial risk prediction.

As for the relationship between the data used in machine learning and the prediction effect, Tsai found that the algorithm of risk identification model showed different advantages in different sample data [3]. This is an earlier study that focused on the correlation between data sources and algorithms. Liu believed that unbalanced data had a serious impact on the performance of classifiers [4], and the operation would cause the big classes to be "valued" and the small classes to be "neglected." According to the work of West and Bhattacharya, the algorithm-level approach is to compensate for unbalanced data to a certain extent [5]. It can handle unbalanced data well by using cost-sensitive learning methods or the skewed distribution methods adopted by the learners themselves [6]. Ghatasheh believes

✉ Fei Wu
2013310135@live.sufe.edu.cn

[1] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212003, China

[2] School of Law, Shanghai University of Finance and Economics, Shanghai 200433, China

that data processing is more effective than algorithm optimization in unbalanced data, so most fraud risk prediction systems choose to balance data at the data level [7].

Such studies are successful because the financial activities they researched are formal, such as credit between companies and banks, and credit card banking. The structured and unstructured data generated by formal financial activities are of high quality and quantity. Therefore, the traditional method, which only uses structured data, and machine learning, which only uses unstructured data can both obtain satisfactory prediction results. For example, Fanning found that BP neural network was no inferior to discriminant analysis and logistic regression [8]. Some scholars used the bank's revenue and expenditure data (structured data) to monitor credit card risks [9, 10]. Anzhong processes network text data (unstructured data) through affective computing when monitoring the fraud risk of e-commerce [11].

However, the quality and quantity of the data, no matter structured data and unstructured data, produced by some informal financial activities are low, which cannot meet the needs of risk prediction, such as microcredit, P2P. For example, many scholars found that the basic information of P2P borrowers, such as gender, age, education level and income, has limited interpretation on their default risk, and it is even impossible to reveal the risk status of risky borrowers directly [12, 13].

Recently, some studies have tried to use the data got from social media and other sources to comprehensively assess the risk status of individual borrowers so as to improve the prediction effect of risk prediction [14, 15]. Meanwhile, some researchers found that the application of information related to social networks can reduce the information asymmetry between lenders and borrowers in P2P market [16]. Ge used microblog data to extract the number of microblog friends, fans, followers and microblog of borrowers, and on the basis of this, he predicted the default risk [17]. Ma revealed the consumption habits and essential characteristics of characters through the data of mobile phone usage [18].

These studies broadened the data sources used in machine learning, but the data sources they used are still single, which can also lead to the homogenization of the information, low accuracy and poor of prediction model. From the perspective of comprehensive utilization of information, if the relationships between some specific factors, such as the different data sources and the different prediction effects, the different prediction scenarios and the different information sources, as well as the problems could be handled properly, the effect of risk prediction through machine learning will be improved from the source.

In view of this, the paper tries to propose TSAIB_RS method, which bases on adaptive fusion of multi-source heterogeneous data and starts from two-stage integration, namely feature integration and result integration to achieve a deep information integration, to improve the efficiency of machine learning in predicting financial risk. In order to test the effectiveness of TSAIB_RS method, the paper takes the default risk in Chinese microcredit market as the test target and draws a conclusion through comparing the prediction effects of various methods

## 2 TSAIB_RS method: two-step adaptive integration based on random subspace

In order to realize the efficient utilization of multi-source heterogeneous data and improve the prediction effect of microcredit default risk, from a two-stage integration perspective, namely features integration and result integration, the paper proposes the TSAIB_RS method, which can achieve adaptive integration of multi-source heterogeneous data.

### 2.1 Formal definition of problem

We assume that there are $n$ samples in the given data set $D = [X, y]$, where $X = [x_1, x_2, \ldots, x_p]$, $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$, $y = (y_1, y_2, \ldots, y_n)^Y$ and $y_i \in \{-1, +1\}$. Let $W^{(1)} = (w_1^{(1)}, w_2^{(1)}, \ldots, w_p^{(1)})^T \in R_+^p$ to be the weight vector of features, $W^{(2)} = (w_1^{(2)}, w_2^{(2)}, \ldots, w_M^{(2)})^T \in R_+^p$ to be the weight vector of base classifiers, $R = (R_1, R_2, \ldots, R_M)^T \in R_+^M$ to be the degree of confidence vector of base classifiers, $X = [x_1^{(1)}, x_2^{(1)}, \ldots, x_{p_1}^{(1)}, x_1^{(2)}, x_2^{(2)}, \ldots, x_{p_2}^{(2)}, \ldots, x_1^{(J)}, x_2^{(J)}, \ldots, x_{p_J}^{(J)}]$ to be the feature space after grouping. In order to build sparse model, we assume that the linear model is $y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + e_i$, where $\beta_j = (\beta_1^{(j)}, \ldots, \beta_{p_j}^{(j)}) \in R^{p_j}$ is the vector of regression coefficients, $e_i$ is a random variable with a normal distribution whose mean is 0 and variance is $\sigma^2$. $|.|$ represents the norm of $L_1$ and $||.||_2^2$ the norm of $L_2$. Data are normalized and centralized before using, namely $\sum_{i=1}^{n} x_{ij} = 0$, $||x_j||^2 = 1$.

The framework structure of TSAIB_RS method is shown in Fig. 1, which mainly includes modules such as data acquisition, feature extraction and model building.

### 2.2 Predicting model: two-step adaptive integration based on random subspace

Most of the existing researches are based on the basic information of borrowers, which is a single data source.

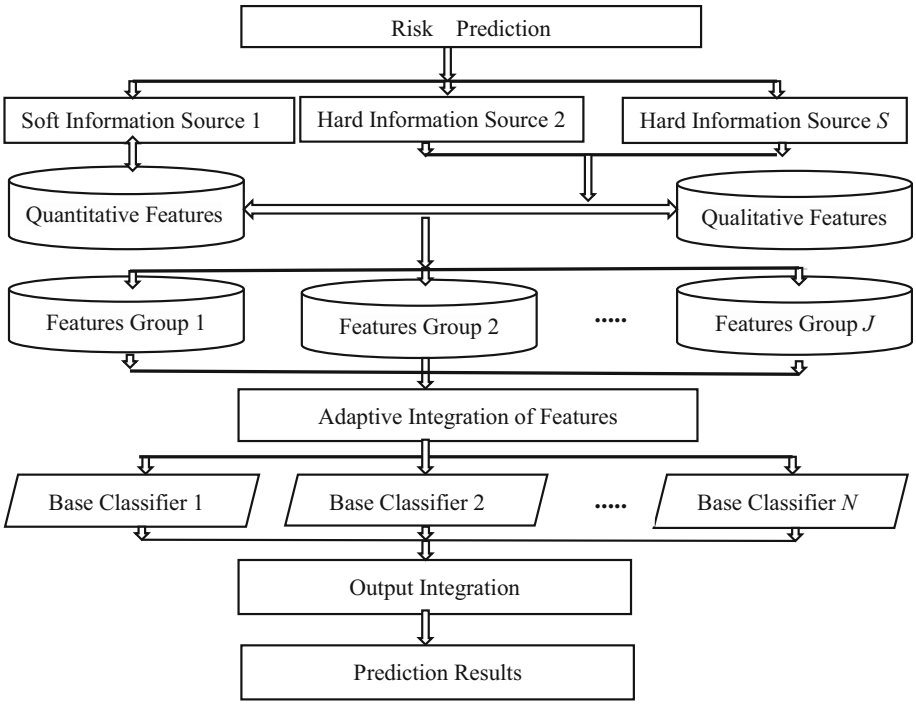


**Fig. 1** Framework of TSAIB_RS method

From the perspective of methodology, it is common to apply a single model to predict default risk. So, in the process of building the prediction model, this paper focuses on discovering and extracting more new useful features from multiple data sources and designs a mechanism that can effectively utilize multi-source features, so as to improve the accuracy and stability of the model.

In general, the ensemble learning methods can achieve more accurate and stable learning results than a single model, because ensemble learning methods take both accuracy and diversity of learning into consideration. If a good balance can be struck between the diversity of classifiers and quasi-certainty, the error and generalization ability of model learning will be improved. Therefore, we introduce ensemble learning methods as the basic prediction model. Among the three kinds of commonly used ensemble learning methods, the random subspace is more suitable for predicting microcredit default risk, because multi-source hard features, soft features and the integration of these features are from the perspective of features. So ensemble learning method RS, which is based on feature division, is the best choice for the paper to build a risk prediction model based on multi-source heterogeneous data integration. Therefore, the random subspace method TSAIB_RS based on two-stage adaptive fusion proposed in the paper is formed. The flow of this method is shown in Fig. 2.

In the first stage of the model, the adaptive integration of multi-source heterogeneous features is carried out to improve the feature quality, reduce the complexity of the model while comprehensively utilizing a variety of prediction information and achieve a good balance between accuracy and generalization ability of the model.

First of all, during indenting the important features, the researchers found that feature weighting method is more efficient than the method to directly select features, because the feature subset which is obtained by feature-weighting-sifting has the capacity to identify and explain the features. That is to say, even if some features are deleted, we can know their degree of importance because of the weighting process.

Secondly, under the special requirements of adaptive integration of multi-source heterogeneous features proposed by the paper, feature integration is facing some new challenges, for example, the heterogeneous features naturally reflect the different aspects of the target, and there are some differences in the feature space distribution. These factors lead to their different prediction capacity, so it is important to deal with characteristics of heterogeneous structure.

Based on the above two points, we consider to introduce the regularized sparse model into the traditional RS method.

Recently, the regularized sparse model has been successfully applied in many important fields such as image identification, target tracking, bioinformatics. Among the regularized sparse models, there is a special method based on group concept, which can effectively handle variables with group structure, for example the typical representative, GL (Group Lasso) method. In order to identify the

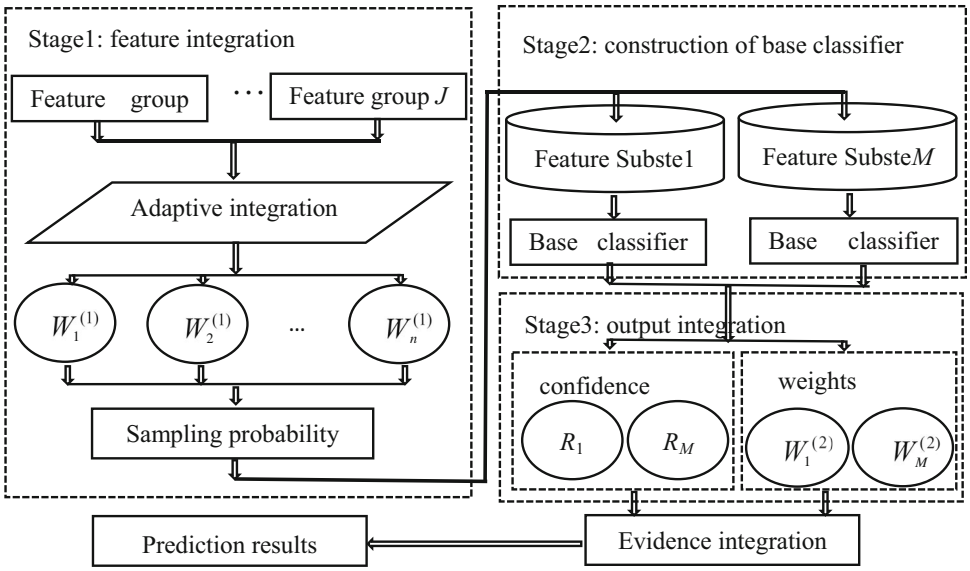


**Fig. 2** Process of the TSAIB_RS method

different role of natural group structure variables in the model learning, some put forward an improved regular norm to punish the regression coefficients of variables in group, which achieved the goal of variable selection in groups. Subsequently, many scholars pointed out the shortcoming of GL method [19, 20], that is, it could only sift the group form of variables and ignored the different effects of intra-group variables on the model. At the same time, they proposed SGL (sparse group Lasso) method to achieve the purpose of simultaneous filtration of variables within and between groups.

In summary, by introducing the SGL model, the multi-source heterogeneous features that have been grouped can be effectively integrated. The SGL model used by the paper is as follows:

$$\beta^*_{SGL} = \arg\min_{\beta}\left\{\frac{1}{2}||y - \sum_{i=1}^{J} x_j\beta^{(j)}||_2^2 + \lambda(1-\alpha)\sum_{j=1}^{J}||\beta^{(j)}||_2 + \lambda\alpha|\beta|\right\} \tag{1}$$

Obviously, the regular term of SGL model includes both the norms of $L_1$ and $L_2$, where the role of norm of $L_1$ is to sift the features in group, while role of norm of $L_2$ is to sift the features between groups. These two kinds of sifting mechanisms will be adjusted through the parameters: when the value is bigger, the punishment, inflicted by SGL model, on the regression coefficients of variables in group becomes greatly, and the compression on those among groups becomes little. Additionally, regularized parameters will adjust the compression scope of features in group: when the value is bigger, more coefficients of feature weights will be compressed to 0; on the contrary, when it is small, the number of reserved features will become more. The feature coefficients obtained by regressing SGL model will be used as the weights, so we will obtain the weight vector $W^{(1)} = (w_1^{(1)}, w_2^{(1)}, \ldots, w_p^{(1)})^T \in R_+^p$, which will be used for feature integration. Then, according to the weight of the probability sampling, the feature subset will be obtained. During the second stage of TSAIB_RS method, we still choose SVM as the base classifier.

There are many methods to integrate the outputs of base classifier, such as average method, voting method, but these methods have some defects, because they will lose a certain amount of prediction information during process of integrating results of base classifier and result in the decline of prediction accuracy. For example, during the process of voting method, it only considers the opinions of most of the base classifiers and ignores role of those classifiers which are minority. However, each classifier holds the different knowledge for a given learning task, therefore, during the integration of the results, not only to consider the importance of the classifier itself but also to fully consider the overall effect. Therefore, the paper introduces the evidential reasoning rule (ER) to realize the full integration of the results. When Yang et al. proposed ER rule method, they pointed out that the integrated evidences should be independent of each other [21], and the reliability and weight of each evidence should be defined before fusion integration. Zhou et al. applied ER rule in ensemble learning and pointed out that the output of each base classifier is independent of each other [22]. So ER rule is available in output integration. We introduce the ER rule method as follows:

$$E_i = \left\{(\theta, p_{\theta,i}), \forall\theta \subseteq \Theta, \sum_{\theta\subseteq\Theta} p_{\theta,i} = 1\right\} \tag{2}$$

where $E_i$ is evidence; $\Theta = \{h_1, h_2, \ldots, h_H\}$ is a set of complete and mutually exclusive assumptions. In a

dichotomous scenario, you can consider $h_H$ as a category distribution. $(\theta, p_{\theta,i})$ is evidence factor, which indicates the support degree $p_{\theta,i}$, to which the evidence $E_i$ supports the proposition $\theta$. When considering the weight and reliability of evidence, the support degree $m_{\theta,i}$ is

$$m_{\theta,i} = \{(\theta, m_{\theta,i}), \forall\theta \subseteq \Theta; (P(\Theta), m_{p(\Theta),i})\} \tag{3}$$

$$m_{\theta,i} = \begin{cases} 0, \theta = \varphi \\ c_{R_{w,i}} m_{\theta,i}, \theta \subseteq \Theta, \theta \neq \varphi \\ c_{R_{w,i}}(1 - R_i), \theta = P(\Theta) \end{cases} \tag{4}$$

where $c_{R_{w,i}} = \frac{1}{1+w_i-R_i}$. When local ignorance is not considered, the total support degree of proposition $\theta$ is:

$$m_{\theta,E(N_c)} = [(1 - R_i)m_{\theta,E(i-1} + m_{p(\Theta),E(i-1)}m_{\theta,i}], \forall\theta \subseteq \Theta \tag{5}$$

The remaining support degree is:

$$m_{p(\Theta),E(N_c)} = (1 - R_i)m_{p(\Theta),E(i-1)} \tag{6}$$

The probability of the support degree of all the final base classifiers (evidence) on the category (proposition) is obtained after the alignment:

$$p_\theta = \frac{m_{\theta,E(L)}}{1 - m_{p(\Theta),E(L)}}, \forall\theta \subseteq \Theta \tag{7}$$

According to the above results, the final prediction results of TSAIB_RS method can be obtained. Considering that the default risk prediction is an unbalanced classification problem, the AUC index can more fully reflect the advantages and disadvantages of the learning effect of the classifier. Therefore, the AUC obtained by testing the base classifiers is regarded as its reliability vector $R = (R_1, R_2, \ldots, R_M)^T \in R_+^M$. At the same time, considering that the risk samples have a higher concern value than the normal samples, the recall rate index on the risk sample is used as the weight of the base classifier $W^{(1)} = (w_1^{(1)}, w_2^{(1)}, \ldots, w_p^{(1)})^T \in R_+^p$.

## 3 Algorithm of TSAIB_RS

For a given data set $X = [x_1, x_2, \ldots, x_p]$, the first step of algorithm is to group the data according to the heterogeneous structure of the data source and the data to obtain the features vector $X = [x_1^{(1)}, x_2^{(1)}, \ldots, x_{p_1}^{(1)}, x_1^{(2)}, x_2^{(2)}, \ldots, x_{p_2}^{(2)}, \ldots, x_1^{(J)}, x_2^{(J)}, \ldots, x_{p_J}^{(J)}]$ after grouping. In the second step, formula (1) is used to obtain the grouping feature weight $W^{(2)} = (w_1^{(2)}, w_2^{(2)}, \ldots, w_M^{(2)})^T \in R_+^M$. The third step is to conduct probability sampling based on features, obtain feature subset and train the base classifiers. Then, based on the reliability and weight of the pre-determined base classifier, the total support on each category is calculated by using Eqs. (4)–(7). Step 4: get the prediction. The specific flowchart of the algorithm is shown in Fig. 3.

## 4 Feature extraction and construction

In predicting financial default risk, there are two types usually used information, namely hard and soft information. Hard information refers to the basic information directly related to the credit and financial statuses of the borrower, which is published on the lending platform, such as balance of payments information, credit score information and historical loan information. In contrast, soft information is difficult to obtain and process. For example, social media information related to borrowers, mobile phone usage data and so on are all soft information. Although soft information is not directly related to the default risk of the borrower, it can help to reflect the borrower's personality characteristics, consumption habits and social status and so on. This important external information cannot be reflected in the basic information. In order to fully and thoroughly reveal the default risk, the model constructed in this research not only takes into account the basic information, but also collects the social information, online consumption information and mobile phone use information of borrowers and predicts their risk status from indirect, direct, internal and external perspectives. The TSAIB_R method constructed in this paper uses both basic hard and soft information to predict the default risk of individual borrowers. Therefore, the corresponding feature extraction also includes hard and soft features.

In the process of feature extraction of basic hard information, we collect the fields based on the platform disclosure information that have been widely used by existing researches as basic features, such as education level, income and expenditure status, credit status. For some specific soft information sources, which contain both qualitative and quantitative data, it is also important to reveal the default risk of borrower. Ge confirmed that the quantitative fields, such as the number of focus extracted from Weibo [17], the number of fans and friends, can be effectively applied to predict the default risk. At the same time, according to the existing research, many researchers proposed that the text information of microblog may reflect the risk preference and credit-related personality and emotional characteristics of borrowers to some extent [23, 24], so it is also useful for the predicting default risk. For this text information, the paper uses word embedding model to transform them into learnable structured fields and finally obtains some emotional features and a series of text features.

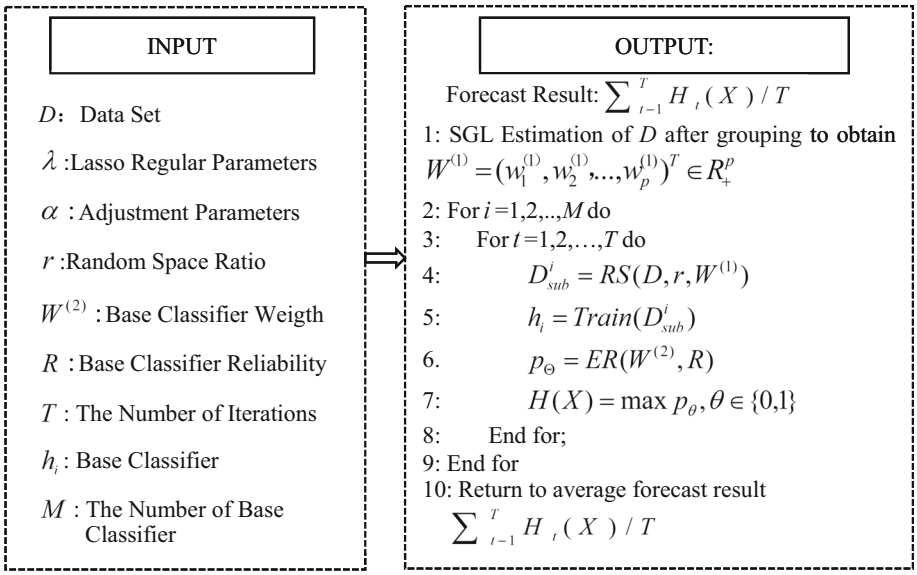


**Fig. 3** The TSAIB_RS algorithm

For hard and soft features from different sources, their effects on default risk prediction may be different, because their distributions and meanings are different. For example, although a few studies [24, 25] had used a variety of information sources to predict risks, the improvement of prediction effects is not significant. Obviously, in order to give full play to the prediction effect of multi-source features, the multi-source heterogeneous characteristics of features need to be processed. Therefore, in the construction of soft features, we clearly group the features of different sources and different data structures. Firstly, features are divided into multiple feature groups according to different data sources. Secondly, the features based on microblog data are divided according to their different qualitative and quantitative data formats. For example, the quantified microblog fields are divided into a feature group, and the text features a feature group. In addition, the extracted emotional features are grouped according to the different emotion polarity. Among them, positive and strong modal emotional features are classified as positive emotional feature group, negative, weak modal and legal-related emotional features are classified as negative emotional features group, and the uncertain emotional features are neutral emotional feature group.

According to the above method, we will finally obtain multi-source and heterogeneous feature set $F = \{F^{(1)}, F^{(1)}, \ldots, F^{(s)}\}$, where $s$ is the number of information sources and $F^{(i)} = \{F_1^{(i)}, F_2^{(i)}\}$(subscripts 1 and 2 represent quantitative characteristics and qualitative data-based characteristics, respectively). So the feature set after being divided in $J$ groups will be $X = [x_1^{(1)}, \ldots, x_{p_1}^{(1)}, \ldots, x_1^{(j)}, \ldots, x_{p_j}^{(j)}, \ldots, x_1^{(J)}, \ldots, x_{p_J}^{(J)}]$.

## 5 Experimental design

In order to compare and verify the improvement effect of TSAIB_RS method in two stages, the paper will apply comparative method among base classifiers and the usually used three method of ensemble learning. Besides SVM, Bagging, Boosting, RS (Random Subspace) methods, we also introduce Lasso_RS method, in which Lasso method is introduced in the feature sampling process of RS method, and ER_RS method, in which ER rule method is applied in the result integration stage of RS method, where the reliability of base classifier is also set as AUC and the weight is set as equal. In addition, PMB_RS, a new and effective method in the field of financial risk prediction, which was verified [26], is also used as a comparison method to verify the effectiveness of TSAIB_RS method in predicting financial risk.

### 5.1 Model evaluation indexes

In the experiment, the AUC, the recall rate and precision of the model in positive and negative categories were selected as evaluation indexes. The value of AUC is equal to the area under the ROC curve, which is usually between 0 and 1. ROC is a curve in two-dimensional coordinate department, with the horizontal axis as the false-positive rate (FPR) and the vertical axis as the true-positive rate (TPR). If the possible classification results are represented as true positive (TP), true negative (TN), false positive (FP) and false negative (FN) and are, respectively, represented as the indicators on the risk category and the risk-free category; the calculation method of various indicators is as follows:

$$R\_Recall = TP/(TP + FN) \quad (8)$$

$$R\_Precision = TP/(TP + FP) \quad (9)$$

$$NR\_Recall = TN/(TN + FP) \quad (10)$$

$$NR\_Precision = TN/(TN + FN) \quad (11)$$

### 5.2 Data set

There are two sources of experimental data used by the paper; the hard information of borrowers comes from CHFS (China Household Finance Survey); the soft information of the borrower comes from web crawler data. The data include three kinds of features: the basic information features, the payment transaction features and the consumption data features.

The basic information features include: gender, age, region, nationality, education level, historical loan amount and historical loan default times in 7 fields. Specifically, the characteristics of mobile phone use include whether the borrower has a real name, and the number and length of calls between the borrower and his/her parents, relatives, spouse, colleagues, friends and classmates during the loan period, totaling 13 fields. The payment transaction features include: whether real name, number of bank cards bound, payment balance, sesame credit rating, total expenditure, total income, total loanable amount, total remaining loan amount, 8 fields. The consumption data characteristics based on the online shopping platform include: the total number of online purchases, the total number of orders, the average price of purchased items, the average price of each purchase and the maximum amount of consumption per purchase, in a total of 5 fields. The characteristics based on the microblog platform include: the number of fans, the number of followers, the number of microblog posts, the number of friends (the situation of mutual attention is defined as the relationship of friends) and a total of 4 fields, as well as 12 emotional characteristics and text characteristics based on the microblog.

The data involved 86 microcredit institutions in 24 provinces and 183 counties (districts and cities) in China. The sample period was from 2011 to 2018, including data of 9688 microcredit borrowers. All data are desensitized before use, in line with the borrower's personal privacy protection requirements.

### 5.3 Experimental process

The paper will take the 10-times 10-fold cross-validation, that is to say the 10-times 10-fold cross-validation will be repeated ten times. The random subspace ratio is set to 0.1 to 0.9 (increase by 0.1). The regular parameters are, respectively, 0.0001, 0.001, 0.01, 0.1 and 1, and the SGL parameters are 0.1 to 0.9 (the increase rate is 0.1).

## 6 Analysis and discussion of experimental results

### 6.1 Experimental results

The experimental results are shown in Tables 1, 2, 3 and 4, where the bold value is the maximum value of AUC index column.

It can be seen from the experimental results of the TSAIB_RS method (two-stage adaptive integration) proposed in the paper achieving the best performance in soft feature set, hard feature set and integrated feature set, especially in the case of integrated feature, the highest AUC value reaches 0.9620. From the perspective of precision index, the classified precision values of risk are generally lower than those of risk-free, while the values of the classified recall rate of risk are generally higher than those of risk-free. The disequilibrium of the sample distribution of default risk is the possible cause of the above phenomenon. Its essence is that the training on minority samples is not enough, so it is easy to be misclassified.

It is worth noting that, even though the sample data distribution is unbalanced, the prediction effect of the TSAIB_RS method on both types of samples has reached a good level, which fully demonstrates the effectiveness and stability of TSAIB_RS method in default risk prediction.

According to the AUC index, under the hard feature set, the average improvement rate of TSAIB_RS method to the comparison method reached 2.7%, and under the soft feature set and the integrated feature set, the average improvement rate was 3.2% and 4.1%, respectively. When compared with the new PMB_RS method in the domain, the AUC improvement rate of TSAIB_RS method is still 1% on average.

### 6.2 Analysis and discussion

Starting from the feature dimension and the method dimension, the paper carries on the comparative analysis. In the feature analysis, the paper explores the effects of traditional basic features, multi-source soft features and integrated features. In order to verify the rationality of the multi-source heterogeneous feature integration proposed in the paper, we also make a comparative analysis of the prediction effects of different methods under the features.

**(1) The influence of different feature sets on the predicted results**

In order to fully verify the fusion effect of TSAIB_RS method on multi-source features, we selected the AUC

**Table 1** Experimental results on hard features

| Hard feature | AUC | Risk | | Risk-free | |
|---|---|---|---|---|---|
| Method | | *R_Precision* | *R_Recall* | *NR_Precision* | *NR_Recall* |
| SVM | 0.8544 | 0.4013 | 0.7955 | 0.9244) | 0.6716 |
| Bagging | 0.8933 | 0.4127 | 0.8134 | 0.9305) | 0.6801 |
| Boosting | 0.8944 | 0.4107 | 0.8510 | 0.9289) | 0.6459 |
| RS | 0.8923 | 0.5456 | 0.8127 | 0.9177) | 0.7533 |
| ER_RS | 0.8921 | 0.5863 | 0.8958 | 0.9301) | 0.7218 |
| Lasso_RS | 0.9059 | 0.6011 | 0.8876 | 0.9300 | 0.7573 |
| TSAIB_RS | 0.9135 | 0.5832 | 0.9321 | 0.9281 | 0.8127 |

**Table 2** Experimental results on soft features

| Soft feature method | AUC | Risk | | Risk-free | |
|---|---|---|---|---|---|
| | | *R_Precision* | *R_Recall* | *NR_Precision* | *NR_Recall* |
| SVM | 0.8637 | 0.5537 | 0.8172 | 0.9127 | 0.7850 |
| Bagging | 0.9133 | 0.5721 | 0.8054 | 0.9254 | 0.8107 |
| Boosting | 0.9127 | 0.6238 | 0.8327 | 0.9261 | 0.8231 |
| RS | 0.9043 | 0.6709 | 0.7958 | 0.9389 | 0.8729 |
| ER_RS | 0.8993 | 0.7153 | 0.8301 | 0.9347 | 0.8914 |
| Lasso_RS | 0.9209 | 0.6971 | 0.8733 | 0.9563 | 0.8876 |
| TSAIB_RS | 0.9449 | 0.7316 | 0.9215 | 0.9712 | 0.8719 |

**Table 3** Experimental results on integrated features

| Integrated features method | AUC | Risk | | Risk-free | |
|---|---|---|---|---|---|
| | | *R_Precision* | *R_Recall* | *NR_Precision* | *NR_Recall* |
| SVM | 0.9271 | 0.5627 | 0.8523 | 0.9587 | 0.8033 |
| Bagging | 0.9104 | 0.5871 | 0.8601 | 0.9613 | 0.8217 |
| Boosting | 0.9276 | 0.6120 | 0.9017 | 0.9687 | 0.8229 |
| RS | 0.9115 | 0.7023 | 0.9124 | 0.9745 | 0.8730 |
| ER_RS | 0.9422 | 0.6522 | 0.9233 | 0.9769 | 0.8409 |
| Lasso_RS | 0.9354 | 0.7216 | 0.9557 | 0.9836 | 0.8564 |
| TSAIB_RS | 0.9620 | 0.7753 | 0.9608 | 0.9875 | 0.8675 |

**Table 4** Experimental results on compared methods

| Method | Features | AUC | Risk | | Risk-free | |
|---|---|---|---|---|---|---|
| | | | *NR_ Precision* | *R_ Recall* | *NR_ Precision* | *NR_ Recall* |
| PBM_RS | Hard features | 0.9063 | 0.6013 | 0.9202 | 0.9569 | 0.8367 |
| | Soft Features | 0.9212 | 0.6241 | 0.8835 | 0.9513 | 0.9124 |
| | Integrated features | 0.9135 | 0.6874 | 0.9439 | 0.9704 | 0.8876 |
| TSAIB_RS | Hard features | 0.9218 | 0.5942 | 0.9387 | 0.9472 | 0.8256 |
| | Soft features | 0.9449 | 0.6911 | 0.9451 | 0.9721 | 0.8409 |
| | Integrated features | **0.9620** | 00.7126 | 0.9571 | 0.9853 | 0.8627 |

index to compare and analyze the prediction effects of hard features, soft features and integrated features under different methods. The comparison results are shown in Fig. 1.

As shown in Fig. 1, the AUC values of integrated feature are significantly higher than that of those of basic hard features and soft features, which fully demonstrates that it is reasonable and feasible to use multi-source information

to improve the prediction effect of default risk. Taking the SVM method as an example, which is the worst among all methods, the AUC obtained by the integrated feature is increased by 7.4% compared with the hard features (from 0.8544 to 0.9271) and about 7.3% compared with the soft feature (from 0.8637 to 0.9271). Under the proposed TSAIB_RS method, the improvement of the integrated feature is up to 5.3% compared with the hard feature (from 0.9135 to 0.9620) and 1.8% compared with the soft feature (from 0.9449 to 0.09602) (Fig. 4).

These results support the use of multi-source information integration for predicting default risk prediction is better than the existing methods. It also suggests that if we can design a reasonable mechanism or a method, which can effectively integrate the multi-source data based on the online platform with the traditional basic data, the prediction accuracy and stability can be improved significantly. It also continue to dig deep for subsequent research and explore more of the available prediction information provides a method for reference.

**(1) The influence of different methods on the prediction results**

In order to fully verify the effectiveness of the proposed method in default risk prediction, the improvement effect of the existing method and the fusion effect of multi-source features, this section makes a comparative analysis of the predicted AUC results of different methods, as shown in Fig. 2.

**(2)The influence of different methods on results**

Obviously, in experiments involving all kinds of methods, TSAIB_RS method of prediction effect is the most prominent and stable performance, the hard feature set, soft features and integration feature sets is to reach the best prediction level; this shows fully convincingly that based on two-phase fusion strategy of success, but also reflects the presented method for different default risk prediction scenarios to have strong adaptive capacity and high-level prediction ability, to verify the proposed method to forecast model accuracy, stability and adaptability of improvement. In addition, it can be seen that the Lasso_RS method and ER_RS method achieve a higher prediction level than the RS method, which indicates the contribution to feature fusion based on the regularized sparse model and the rationality of using evidential reasoning rules to synthesize the prediction results of the base classifier. The TSAIB_RS method has two fusion strategies, so it has the best prediction effect. For the existing method PBM_RS, the increase value of TSAIB_RS method is 1.3% under the hard feature and 1.4% under the soft feature. With the increase of the number of features, the enhancement effect of TSAIB_RS method is more obvious. For example, under the integrated features, TSAIB_RS method improves the PBM_RS method by 1.3% (Fig. 5).

**(3) Analysis of Parameters Sensitivity of TSAIB_RS Method**

According to the model construction above, we can see that TSAIB_RS method has subspace ratio $r$, regularized parameters $\lambda$ and group adjustment parameters $\alpha$, whose values have an important impact on the predicted effects of the TSAIB_RS method. Therefore, we analyze the sensitivity of these three parameters of TSAIB_RS method.

According to the experimental results, TSAIB_RS method performs better when the subspace ratio $r$ is in the middle. Therefore, we will, respectively, analyze the common sensitivity of $\lambda$ and $\alpha$ under $r = 0.3$,$r = 0.5$ and $r = 0.7$. The results are shown in Fig. 3, with $\alpha$ to be the horizontal axis, $\lambda$ to be the vertical axis (Fig. 6).

The results shown in Fig. 3 indicate that when $r = 0.3$, the value of AUC of TSAIB_RS method achieves the highest value (0.9623) under the condition of $\alpha = 0.5 \cup \lambda = 0.01$. Similarly, when $r = 0.5$, the highest value of AUC is 0.9487 under the condition of $\alpha = 0.1 \cup \lambda = 0.001$, and the highest value is 0.9401 under the condition of $\alpha = 0.7 \cup \lambda = 1$.

Generally speaking, the AUC of TSAIB_RS method decreases gradually with the increase in $\alpha$, which implicates the increase in the sparse efforts in the group. It means that the change in features in group impacts the predicted effects significantly. Meanwhile, the performance of TSAIB_RS method on parameters $\lambda$ is $V$-shaped, which indicates that when the intermediate value of $\lambda$ is taken, the features in and between groups are compressed to a certain extent, and the prediction effect is not ideal. On the contrary, when the TSAIB_RS method tends to compress only

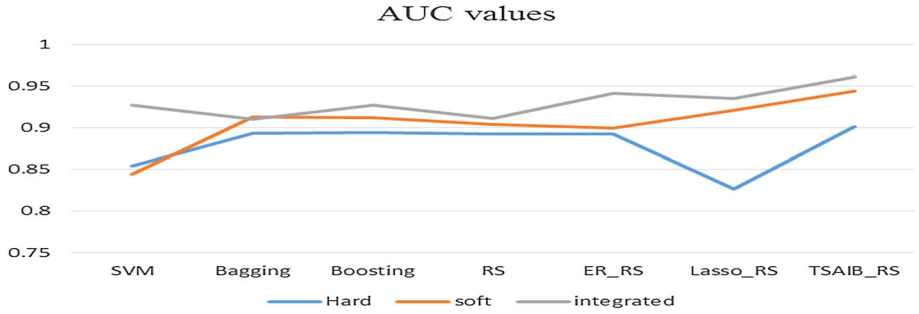


**Fig. 4** Comparisons among Different Feature Sets

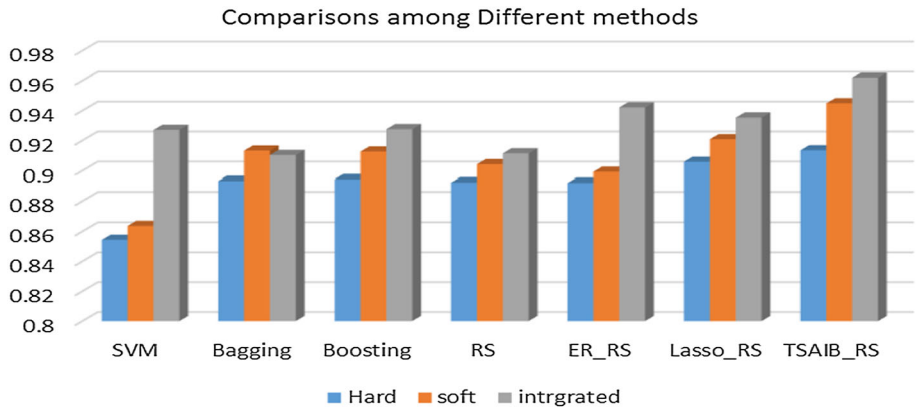


**Fig. 5** Comparisons among different methods

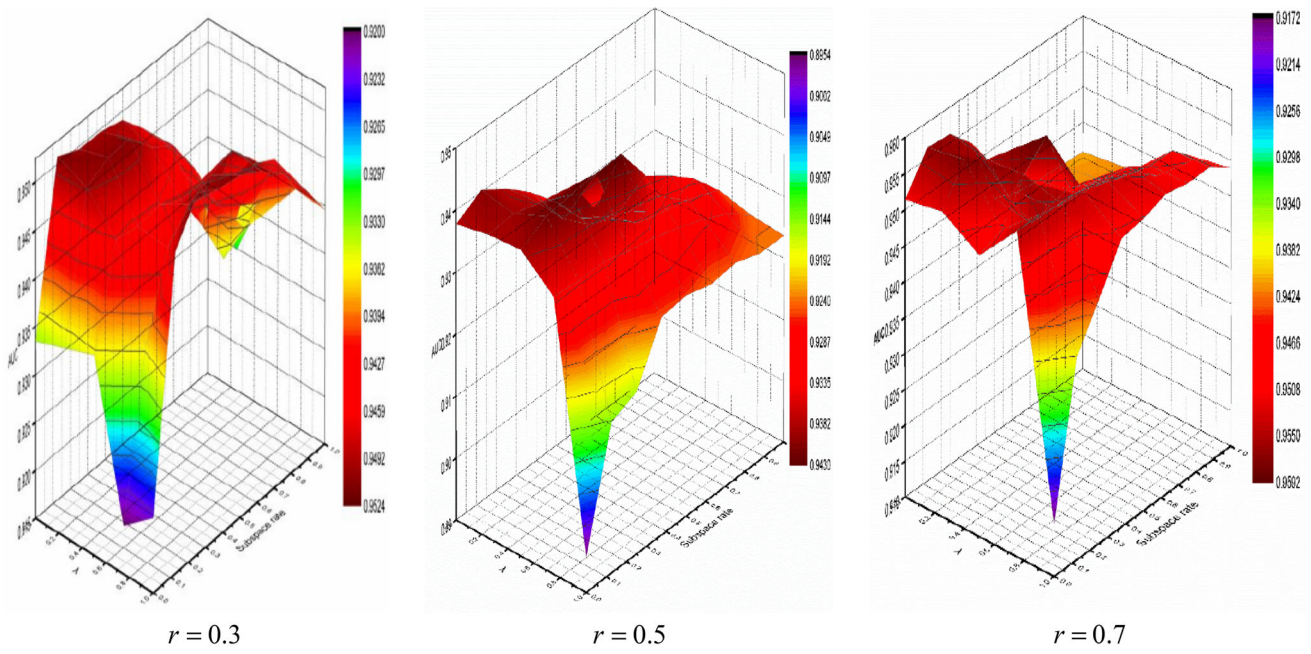


**Fig. 6** Sensitivity analysis of the TSAIB _RS method

the features of a group or compress only the features in groups, it can achieve a more ideal prediction effect.

## 7 Conclusions and future research prospect

The TSAIB_RS, a method proposed by the paper, is to predict default risk of microcredit which is an informal financial activities that cannot generate enough structured or unstructured date for conventional methods to use. Through experiments on collected multi-source data sets, it is found that the proposed two-stage fusion strategy plays a significant role in improving the prediction effect. At the same time, the stable and accurate experimental results also prove that it is feasible and reasonable to use multi-source data to improve the prediction effect of default risk.

Although the paper takes the fusion and efficient utilization of multi-source heterogeneous data as the starting point and constructs an adaptive fusion model based on multi-source heterogeneous characteristics and obtains good prediction results and high prediction stability, there are still some problems to be further solved in the future research.

From the perspective of features, first, due to the limitation of data acquisition channels, this study has not considered the prediction effect of other feasible and useful data sources (such as news, online comments). Therefore, the meaning of "multi-source" in multi-source information needs to be further expanded. Second, with the extensive and successful application of deep learning model in text feature extraction, the text feature extraction method adopted in this study needs to be further improved to give full play to the prediction effect of unstructured

information. Third, in view of some new multi-source data currently used in this study, we need to conduct a comprehensive comparison and deeper analysis of its differentiated prediction effect in future research and provide a reasonable (theoretical) explanation for its usefulness.

From the perspective of methodology, firstly, the prediction effect of the new method proposed in this study needs to be compared with more mature methods in this field, so as to further verify the vitality of multi-source data fusion strategy in the field of financial risk prediction. Secondly, the effectiveness and stability of the proposed method in this study need to be verified on more data sets. Finally, one of the key research directions in the future is to build a more efficient automatic iterative algorithm to realize the new method proposed in this study, which lays a foundation for its popularization in practical application.

**Acknowledgements** The paper is one of mid-term results of the humanities and social science planning project funded by the ministry of education of PRC, named "Researches of the Formation Mechanism of Low Efficiency of Poverty Alleviation of Microcredit and Innovation Practice Model in Jiangsu Province" (20YJA790028), a major project of philosophy and social science research in Jiangsu universities, "Researches on the Optimization of Fintech Innovation Supervision Path (2019SJZDA060)" as well as Anhui Province Social Science Association Project (2019CX079).

## Compliance with ethical standard

**Conflict of interest** The authors declare that they have no conflict of interests.

## References

1. Rajan RG (1992) Insiders and Outsiders. The choice between informed and Arm's-length debt. J Finance 47(4):1367–1400
2. Boot AWA, Thakor AV (1994) Moral Hazard and secured lending in an infinitely repeated credit market game. Int Econ Rev 35(4):899–920
3. Tsai CF, Hsu Y-F, Yen DC (2014) A comparative study of classifier ensembles for bankruptcy prediction. Appl Soft Comput 24:977–984
4. Liu X, Xu Z, Yu R (2012) Spatiotemporal variability of drought and the potential climatological driving factors in the Liao River. Hydrol Process 26(1):1–14
5. West J, Bhattacharya M (2016) Intelligent financial fraud detection: a comprehensive review. Comput Secur 57(47):66
6. Nazari M, Alidadi M (2013) Measuring credit risk of bank customers using artificial neural network. J Manag Res 5(5):17
7. Ghatasheh N (2014) Business analytics using random forest trees for credit risk prediction: a comparison study. Int J Adv Sci Technol 72:19–30
8. Fanning KM, Cogger KO (1998) Neural network detection of management fraud using published financial data. Int J Intell Syst Account Finance Manag 7(1):21–41
9. Bhattacharyya S, Jha S, Tharakunnel K (2011) Data mining for credit card fraud: a comparative study. Decis Support Syst 50(3):602–613
10. Sahin Y, Bulkan S, Duman E (2013) A cost-sensitive decision tree approach for fraud detection. Expert Syst Appl 40(15):5916–5923
11. Huang Anzhong (2018) A risk detection system of e-commerce: researches based on soft information extracted by affective computing web texts. Electronic Commerce Res 18:143–157
12. Guo Y, Zhou W, Luo C, Liu C, Xiong H (2016) Instance-based credit risk assessment for investment decisions in P2P lending. Eur J Oper Res 249(2):417–426
13. Serrano-Cinca C, Gutiérrez-Nieto B (2016) The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decis Support Syst 89:113–122
14. Estrada F (2011) Theory of financial risk. University Library of Munich, Munich
15. Chen D, Han C (2012) A comparative study of online P2P lending in the USA and China. J Internet Bank Commerce 17(2):1–15
16. Chen N, Ribeiro B, Chen A (2016) Financial credit risk assessment: a recent review. Artif Intell Rev 45(1):1–23
17. Ge R, Feng J, Gu B, Zhang P (2017) Predicting and deterring default with social media information in peer-to-peer lending. J Manag Inf Syst 34(2):401–424
18. Ma L, Zhao X, Zhou Z, Liu Y (2018) A new aspect on P2P online lending default prediction using meta-level phone usage data in China. Decis Support Syst 111:60–71
19. Meier L, Van De Geer S, Bühlmann P (2008) The group lasso for logistic regression. J R Statist Soc Ser B (Statist Methodol) 70(1):53–71
20. Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. J Comput Graph Statist 22(2):231–245
21. Yang J-B, Xu D-L (2013) Evidential reasoning rule for evidence combination. Artif Intell 205:1–29
22. Zhou L, Tam KP, Fujita H (2016) Predicting the listing status of Chinese listed companies with multi-class classification models. Inf Sci 328:222–236
23. Loughran T, Mc Donald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10 Ks. J Finance 66(1):35–65
24. Simian D, Stoica F, Bărbulescu A (2020) Automatic optimized support vector regression for financial data prediction. Neural Comput Appl 32:2383–2396
25. Xu Z, Cheng C, Sugumaran V (2020) Big data analytics of crime prevention and control based on image processing upon cloud computing. J Surveill Secur Saf 1:16–33
26. du Jardin P (2016) A two-stage classification technique for bankruptcy prediction. Eur J Oper Res 254(1):236–252