Study of Personal Credit Risk Assessment Based on SVM

Xin LI[1], Han XIA[2]

Abstract

Purpose: Support vector machines (SVMs) ensemble has been proposed to improve classification performance of Credit risk recently. However, currently used fusion strategies do not evaluate the importance degree of the output of individual component SVM classifier when combining the component predictions to the final decision. To deal with this problem, this paper designs a support vector machines (SVMs) ensemble method based on fuzzy integral, which aggregates the outputs of separate component SVMs with importance of each component SVM. **Research design, data, and methodology**: This paper designs a personal credit risk evaluation index system including 16 indicators and discusses a support vector machines (SVMs) ensemble method based on fuzzy integral for designing a credit risk assessment system to discriminate good creditors from bad ones. This paper randomly selects 1500 sample data of personal loan customers of a commercial bank in China 2015-2020 for simulation experiments. **Results**: By comparing the experimental result SVMs ensemble with the single SVM, the neural network ensemble, the proposed method outperforms the single SVM, and neural network ensemble in terms of classification accuracy. **Conclusions**: The results show that the method proposed in this paper has higher classification accuracy than other classification methods, which confirms the feasibility and effectiveness of this method.

Keywords: Credit index, Credit evaluation, Support vector machines ensemble, customer to customer, bagging, Fuzzy integral

JEL Classification Code :C15, G21, G28

1. Introduction

Credit risk assessment is essentially a classification problem. It plays an important role in banks credit risk management. Credit risk assessment is the set of decision models and techniques that aid lenders in granting consumer credit by assessing the risk of lending to different consumers. It is an important area of research that enables financial institutions to develop lending strategies to optimize profit. Additionally, bad debt is a growing social problem that could be tackled partly by better informed lending enabled by more accurate risk assessment models. An accurate assessment of risk could be translated into a more efficient use of resources or a less loan loss for a bank.

In recent years, many statistical models like logistic regression and discriminant analysis have been used to make credit risk assessment models (Zhao & Li, 2022; Jadwal et al, 2022). They are mostly used as kinds of binary classifiers in the traditional credit risk studies. Their purpose is to classify customers into good or bad credit accurately. Discriminant analysis and logistic regression have been the most widely used techniques for building credit risk models. Both have the merits of being conceptually straightforward and widely available in statistical software packages. But discriminant analysis is statistically valid only if the independent variables are normally distributed. However, this assumption is often violated (Khemakhem & Younés Boujelbène, 2015).

---

1 First Author. Master student, College of Business Administration, Research Assistant, Henan University of Science and Technology, China. Email:276204899@qq.com
2 Corresponding Author. PhD student, College of Business Administration, Research Assistant, Assistant Professor, Henan Finance University, China, Email: xiahan0823@163.com

Moreover, a priori probability of failure is needed for the prediction of a failure. It is not always easy to find any sensible estimate for the prior probability of failure although discriminant analysis gives the posterior probability.

These shortcomings have led to the use of the logistic regression model which does not assume multinormality and also gives an estimate for the probability of bad credit (Abid, 2022). This methodology assumes that the predicted values of the probability are cumulatively constrained by the logistic function. Srinivasan and Kim (1987) included logistic regression in a comparative study with other methods for a corporate credit granting problem.

In recently literatures, artificial neural network (Hu & Su, 2022; Mahbobi et al, 2021) and fuzzy theory (Wójcicka-Wójtowicz, 2021; Xie et al, 2022) are usually used to evaluate credit ranks. Applications of decision tree methods in credit scoring are described by Hand and Henley (1997). Neural networks are well suited to situations, where we have a poor understanding of the data structure. In fact, neural networks can be regarded as systems which combine automatic feature extraction with the classification process, i.e. they decide how to combine and transform the raw characteristics in the data, as well as yielding estimates of the parameters of the decision surface. This means that such methods can be used immediately, without a deep grasp of the problem. In general, however, methods, which utilize a good understanding of the data and the problem, might be expected to perform better. The type of neural network that is normally applied to credit scoring problems can be viewed as a statistical model involving linear combinations of nested sequences of non-linear transformations of linear combinations of variables. Tsai (2008) described neural networks to corporate credit decisions and fraud detection. However, some inherent drawbacks neural networks possessed can't be surmounted. The drawbacks include: 1. Gradient algorithm to train the optimal solution is a local optimal problem often induces that the optimal solution is a local optimal, not global optimal, resulting in the trained network useless. 2. The theory of training a neural network is based on empirical risk minimization (ERM), provided the objective function minimum errors to the training samples, but lack of generalization to unknown samples. 3. The construction designs of neural networks, such as number choose of hidden layer node, strongly depend on users' experiences, lack of strict design proceeding in theory. 4. Whether the training algorithms of neural networks converge and how speed to converge is uncontrolled.

Fortunately, Vapink (1995) first present a novel pattern identification method, based on statistic learning theory, called support vector machine (SVM). Comparing with artificial neural networks, SVMs have better generalization and more precise categorization, and give rid of 'dimension calamity' and 'overfitting' (Zhao & Li, 1995). And SVMs are used to forecast time series and evaluate credits, and general SVMs treat all training samples equivalently. Recently, support vector machines ensemble has been proposed and applied in many areas in order to improve classification performance (Anil Kumar et al, 2022; Dou et al, 2020). The experimental results in these applications show that support vector machines ensemble can achieve equal or better classification accuracy with respect to a single support vector machine. However, a common used majority voting aggregation method is adopted in these papers, and this method does not consider the degree of importance of the output of component support vector machines classifier when combining several independently trained SVMs into a final decision. In order to resolve this problem, a support vector machines ensemble strategy based on fuzzy integral is proposed in this paper. And then evaluate credit in C2C mode business based on the proposed method. The experimental result shows that this method is stable, highly accurate, strong robust and feasible. It demonstrates SVMs ensemble is effective for website credit evaluation.

2. Credit Evaluation Indexes

Figure 2 illustrates a credit evaluation system model of customer. The model consists of three parts: credit index system construction, appreciation model, and evaluation results (credit ranks).

The appreciation model part of Figure 1 shows a general scheme of the proposed SVMs ensemble. SVMs ensemble process is divided into following steps:

Step 1: Generate m training data sets via bagging from original training set according to 4.1 and train a component SVM using each of those training sets.

Step 2: Obtain probabilistic outputs model of each component SVM according (Platt, 1999).

Step 3: Assign the fuzzy densities { $g_k(\{SVM_i\})$, $k=1,...,c$ }, the degree of importance of each component SVM_i, i=1,...,m, based on how good these SVMs performed on their own training data.

Step 4: Obtain probabilistic outputs of each component SVM when given a new test example.

Step 5: Compute the fuzzy integral e_k for w_k, $k=1,...,c$ according to 4.2.

Step 6: Get the final decision $k^* = \arg\max_k e_k$.

In this paper, the credit indexes are filtrated and listed as Table 1. There are 16 credit indexes.
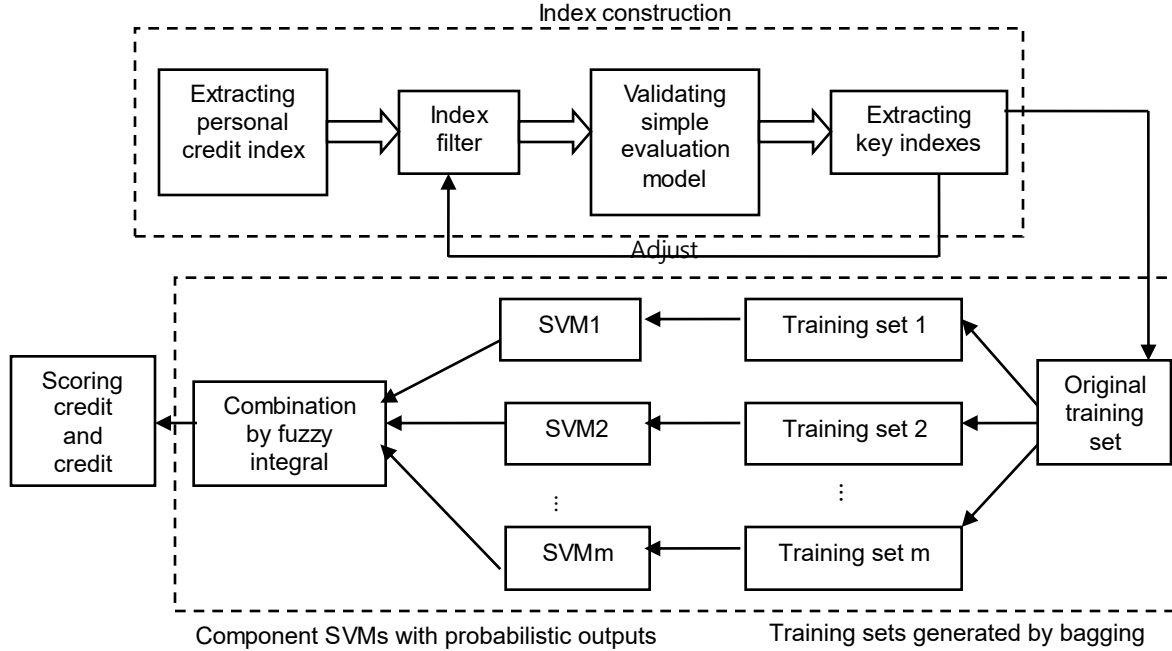

Figure 1: Credit Evaluation Model

Table 1: Indexes of Credit Evaluation

No.	Indexes	Weight	No.	Indexes	Weight
1	Age	3	9	Census registration	2
2	Occupational status	2	10	Amount in bank accounts	4
3	Headship	4	11	Installments expenses	4
4	Job stability	3	12	Credit card expenses	4
5	Education level	4	13	Marital status	1
6	Family income	4	14	Amount of children	1
7	Personal income	4	15	Property values	3
8	History bad credit	3	16	Financial assets	3

3. Credit Evaluation by SVMs Ensemble

3.1 Bagging to Construct Component SVMs

The Bagging algorithm (Zhang, 2021) generates training data sets by randomly re-sampling, but with replacement, from the given original training data set. Each training set will be used to train a component SVM. The component predictions are combined via fuzzy integral.

3.2. SVMs Ensemble Based on Fuzzy Integral

In the following, we introduce the basic theory about the fuzzy integral (Kwak & Pedrycz, 2005).

Let $X = \{x_1, x_2, ..., x_n\}$ be a finite set. A set function $g : 2^X \rightarrow [0,1]$ is called a fuzzy measure if the following conditions are satisfied:

$$1 \cdot g(\phi) = 0, g(X) = 1$$
$$2 \cdot g(A) \le g(B), \text{当} A \subset B \text{且} A, B \subset X \qquad (1)$$

From the definition of fuzzy measure g, Sugeno developed a so-called g_λ fuzzy measure satisfying an additional property:

$$g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B) \qquad (2)$$

For all $A, B \subset X$ and $A \cap B = \phi$, and for some $\lambda > -1$.

Let $h : X \to [0,1]$ be a fuzzy subset of X and use the notation $A_i = \{x_1, x_2, ..., x_i\}$. For being a g_λ fuzzy measure, the value of $g(A_i)$ can be determined recursively as

$$\begin{cases} g(A_1) = g(\{x_1\}) = g_1 \\ g(A_i) = g_i + g(A_{i-1}) + \lambda g_i g(A_{i-1}), g_i = g(\{x_i\}), 1 < i \le n \end{cases} \qquad (3)$$

λ is given by solving the following equation

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g_i) \qquad (4)$$

Where $\lambda \in (-1, +\infty)$ and $\lambda \ne 0$.

Suppose $h(x_1) \ge h(x_2) \ge ... \ge h(x_n)$, (if not, X is rearranged so that this relation holds). Then the so-called Sugeno fuzzy integral e with respect to g_λ fuzzy measure over X can be computed by

$$e = \max_{i=1}^{n} [\min(h(x_i), g(A_i))] \qquad (5)$$

Thus the calculation of the fuzzy integral with respect to a g_λ fuzzy measure would only require the knowledge of the fuzzy densities, where the ith density g_i is interpreted as the degree of importance of the source x_i towards the final decision. These fuzzy densities can be subjectively assigned by an expert or can be generated from training data.

Let $\Omega = \{\omega_1, \omega_2, ..., \omega_c\}$ be a set of classes of interest and $S = \{\text{SVM}_1, \text{SVM}_2, ..., \text{SVM}_m\}$ be a set of component SVMs. Let: $h_k : S \to [0,1]$ be the belief degree of a new sample x belongs to class ω_k, that is $h_k(\text{SVM}_i)$ is the probability in the classification of a new sample x to be in class ω_k using SVMs$_i$. If we get $\{h_k(\text{SVM}_i), i = 1, 2, ..., m\}$ and know fuzzy densities $\{g_k(\{\text{SVM}_i\}), i = 1, 2, ..., m\}$, the fuzzy integral e_k for class ω_k can be calculated using (3) to (5). When the fuzzy integral values $\{e_k, k = 1, ..., c\}$ are obtained, we can get the final decision: $k^* = \arg\max_k e_k$.

3.3. Assignment of Fuzzy Density

While to SVMs ensemble, the critical issue is how to effectively assign the fuzzy densities, which are significant to the values of the fuzzy integral. There are a number of ways to obtain the fuzzy densities (Zhang et al, 2015; Roy & Shaw, 2022), but they were only base on accuracy, and did not consider the uncertainty presented in the process of recognition. The densities values here are generated based on both accuracy and uncertainty. The procedure of calculating the densities is described as table 2.

Table 2: Outputs of the Classifier x_j

Class \ Sample	C_1	C_2	...	C_n
y_1	h_{11}^j	h_{21}^j	...	h_{n1}^j
y_2	h_{12}^j	h_{22}^j	...	h_{n2}^j
..	...	...	...	...
y_l	h_{1l}^j	h_{2l}^j	...	h_{nl}^j

Suppose there are l training samples: $y_1, y_2, ..., y_l$, presume $Y_l = \{y_1, y_2, ..., y_l\}$. For each classifier x_j we can obtain sets of outputs in Table 2

For class C_1, it can be regarded as a fuzzy set definite on Y_l, namely

$$C_l^j = \begin{bmatrix} y_1 & y_2 & ... & y_l \\ h_{1l}^j & h_{2l}^j & ... & h_{nl}^j \end{bmatrix} \qquad (6)$$

Where h_{1l}^j denotes C_1 from the point of view of x_j, and its vagueness can be measured by:

$$E_v(C_1^j) = -\frac{1}{l}\sum_{i=1}^{l} h_{1t}^j \ln h_{1t}^j + (1-h_{1t}^j)\ln(1-h_{1t}^j) \qquad (7)$$

$E_v(C_1^j)$ represents the vagueness of C_1 under the condition of x_j. When $h_{1t}^j = 0.5$ for all $t, t=1,2,...,l$, $E_v(C_1^j) = 1$, which represents the greatest vagueness. When $h_{1t}^j = 0 \; or \; 1$ for all $t, t=1,2,...,l$, $E_v(C_1^j) = 0$, which represents no vagueness. Because the output is obtained by x_j, it can be interpreted as the cognitive ability of x_j with respect to the class C_1. Similarity, we can calculate the cognitive ability that xi towards the other classes, and finally we obtain the following matrix.

$$\begin{bmatrix} E_v^1(C_1) & E_v^1(C_2) & ... & E_v^1(C_n) \\ E_v^2(C_1) & E_v^2(C_2) & ... & E_v^2(C_n) \\ ... & ... & ... & ... \\ E_v^m(C_1) & E_v^m(C_2) & ... & E_v^m(C_n) \end{bmatrix} \qquad (8)$$

Where $E_v(C_1^j) \; i=1,2,...,n; j=1,2,...,m$ represent the sensitivity of x_j with respect to class C_i.

On the other hand, the performance matrix that is a $n \times n$ matrix can be formed base on the output of classifiers.

$$\begin{bmatrix} u_{11}^j & u_{12}^j & ... & u_{1n}^j \\ u_{21}^j & u_{22}^j & ... & u_{2n}^j \\ ... & ...... & ... \\ u_{n1}^j & u_{n2}^j & ... & u_{nn}^j \end{bmatrix} \qquad (9)$$

where $u_{ip}^j (i, p=1,2,...,n)$ represents the number of training data belonging to class C_i being classified as class C_p, by classifier x_j, and the data is classified correctly when $i = p$. Thus:

$$A_i^j = \frac{u_{ii}^j}{\sum_{p=1}^{n} u_{ip}^j} \qquad (10)$$

is interpreted as the degree in which classifier x_j identifies class C_i correctly. Then fuzzy densities can be defined as:

$$g_i^j = A_i^j \times e^{-E_v(C_i^j)} \qquad (11)$$

In the case of $E_v(C_1^j) = 0$, there is no uncertainty, which indicates that classifier x_j has clear conception of the class C_i, in other words, the classifier x_j has the highest discriminability to class C_i, so the fuzzy density is the classification rate; In the case of $E_v(C_1^j) = 1$, there are the largest uncertainty, which indicates classifier x_j has no ability to distinguish the class C_i, so the fuzzy density is the minimal.

After determining the fuzzy densities, the g_λ-fuzzy measure and the fuzzy integral can be computed. The final output class from the combination classifier is the one with the highest integrate value.

4. Experimental Results

In this paper, the experimental data came from a commercial bank, we randomly select selected 1500 customers' samples, in which 1279 samples are good, called 'good' customers, and the other 221 samples' conditions are bad, they will mostly fell back, called 'bad' customers. However, the two categories are unequal, which is not suitable for SVMs ensemble training. SVMs ensemble training needs the two categories' amount similarity. If we train the gained sample directly, the optimal separating hyperplane will prefer to smaller density category (Figure 2), its which will lead to much sorting error. In order to get a good performance sort, we must deal with samples to balance the two categories. We set Different assembled penalties result in different performance.
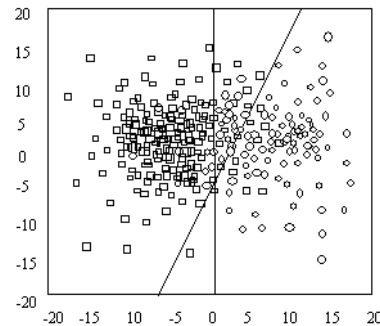

Figure 2: Unequal Samples Classification

In this paper we take the method of constantly approaching to gain C_i ($i = 1, 2$), which is the tow type of samples' penalties. At the process it must satisfy the following formula:

$$\frac{1}{2}(W \bullet W) = \sum_{i=1}^{n} C_i \xi_i \qquad (12)$$

Where

$$C_i = \begin{cases} C_1 & X_i \in \Xi_1 \\ C_2 & X_i \notin \Xi_1 \end{cases} \qquad (13)$$

And C_1, C_2 must satisfy

$$\frac{C_1}{C_2} = \frac{N_2}{N_1} \qquad (14)$$

So in this paper, $\dfrac{C_2}{C_1}$ = 6 (Ξ_1 means the good customers' sample). And we get C_1 =200, and C_2 =1200.

In this paper, we select half of each dataset randomly for training ('good': 639; 'bad': 110), and the residual samples ('good': 640; 'bad': 111) for testing the model's forecasting accuracy.

For bagging, we re-sample randomly 70 samples with replacement from the original training data set. We train three component SVMs based on different penalties independently over the three training data sets generated by bagging and aggregate three trained SVMs via fuzzy integral. Each component SVM uses Polynomial kernel function:

$$K(x, x_i) = [(x \bullet x_i) + c]^q$$
$$K(x, x_i) = \exp(-|x - x_i|^2 / \sigma^2) \qquad (15)$$

And the corresponding parameters are selected by five-fold cross-validation. To avoid the tweak problem, ten experiments are performed and the average performance in training and test datasets is reported in table 3 and table 4. ('0' means 'good', and '1' means 'bad'). As Practical problems are Complex, this accuracy is acceptable for credit evaluation.

The comparison of single SVM, SVMs ensemble via majority voting, single neural network, and fuzzy neural network ensemble is shown in Figure 3. On Figure 3, from left to right: 1: Single SVMs, 2: SVMs ensemble via majority voting, 3: SVMs ensemble via fuzzy integral, 4: Single neural network, 5: Fuzzy neural network ensemble.

Table 3: Training Set Correct Rate

Observation	Forecast		
	0	1	Test Accuracy
0	534	105	83. 6188%
1	22	88	79.8530%
total	32	24	81. 7359%

Table 4: Test Set Correct Rate

Observation	Forecast		
	0	1	Test Accuracy
0	544	96	84.9561%
1	26	85	76.3909%
total	59	52	80.6735 %

In addition to examining average prediction accuracy of SVMs ensemble via fuzzy integral, we compare Type I and Type II of single SVM, SVMs ensemble via majority voting, SVMs ensemble via fuzzy integral, single neural network, and fuzzy neural network ensemble. Table5 presents the average error rate (%) of Type I error and Type II error over the single SVM, SVMs ensemble via majority voting, SVMs ensemble via fuzzy integral, single neural network, and fuzzy neural network ensemble under the training datasets. (•Type I error: This shows the rate of prediction errors of a model, which is to incorrectly classify the bad credit group into the good credit group. •Type II error: Opposed to Type I error, this presents the rate of prediction errors of a model to incorrectly classify the good credit group into the bad credit group.) On average, SVMs ensemble via fuzzy integral is the winner as the best model/classifier architecture for credit scoring. From table 5, SVMs ensemble via fuzzy integral has least Type I error and Type II error. It also proves the proposed method stable, highly accurate, strong robust and feasible.

Table 5: Average Error Rate of TypeⅠ and TypeⅡ Error under the Training Datasets

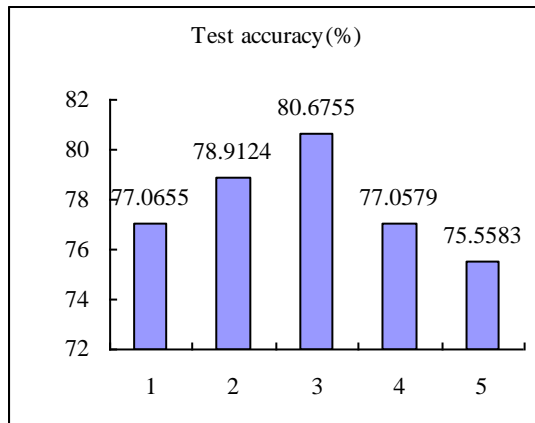| Type	Single SVM	SVMs Ensemble via Majority voting	SVMs Ensemble via Fuzzy Integral	Single Neural Network	Fuzzy Neural Network Ensemble
Type I	22.44%	25.49%	23.61%	21.09%	26.79%
Type II	23.23%	16.69%	13.04%	24.70	24.09

Figure 3: The Results of Classification Accuracy Comparison

Recent studies suggest combining multiple classifiers (or classifier ensembles) should be better than single classifiers. SVMs ensemble proves the conclusion. SVMs training classifiers gain better classification result than single SVM. However, the performance of the single neural network classifier outperforms the (diversified) multiple neural network classifiers, the main reason is the divided training datasets may be too little to make the multiple classifiers for neural network classifier. While SVMs training classifiers performs better on small datasets, so SVMs ensembles accuracy is higher than neural network ensemble.

5. Conclusions

With the development of electronic commerce, increasingly bargaining and trade have been achieved in the websites, so trade security and customer credit are taken into account widely. The problem of how to evaluate personal credit is more and more important. In this paper, we constructed credit index system of customer: there are sixteen indexes in detail. We propose a support vector machines ensemble strategy based on fuzzy integral for classification. The most important advantage of this approach is that not only are the classification results combined but also that the relative importance of the different component SVMs classifier is also considered. Different from previous studies, the fuzzy integral support vector machine integration model constructed in this paper considers the output importance of a single support vector machine classifier. This method has higher stability, accuracy and robustness, which proves the effectiveness of fuzzy integral support vector machine integration in website credit evaluation. The research conclusion of this paper is conducive to further improving the efficiency and

accuracy of credit risk assessment of individual customers of commercial banks, and provides a basis for commercial banks to effectively conduct credit risk assessment in the era of big data.

However, this paper mainly uses the data of personal loan customers of Chinese commercial banks for simulation, which has certain limitations. Future research will extend the sample data to international commercial banks to test the expanded application of the method proposed in this paper in international commercial banks.

References

Abid, L. (2022). A Logistic Regression Model for Credit Risk of Companies in the Service Sector. *International Research in Economics and Finance*, *6*(2), 1-10.

Anil Kumar, C. J., Raghavendra, B. K., & Raghavendra, S. (2022). A Credit Scoring Heterogeneous Ensemble Model Using Stacking and Voting. *Indian Journal of Science and Technology*, *15*(7), 300-308.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., ... & Pham, B. T. (2020). Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides*, *17*(3), 641-658.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541.

Hu, Y., & Su, J. (2022). Research on Credit Risk Evaluation of Commercial Banks Based on Artificial Neural Network Model. *Procedia Computer Science*, *199*(1), 1168-1176.

Jadwal, P. K., Pathak, S., & Jain, S. (2022). Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis. *Microsystem Technologies*,*56*(3), 1-7.

Kim, H. C., Pang, S., Je, H. M., Kim, D., & Bang, S. Y. (2003). Constructing support vector machine ensemble. *Pattern recognition*, *36*(12), 2757-2767.

Khemakhem, S., & Boujelbene, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach. *Accounting and Management Information Systems*, *14*(1), 60-78.

Kwak, K. C., & Pedrycz, W. (2005). Face recognition: A study in information fusion using fuzzy integral. *Pattern Recognition Letters*, *26*(6), 719-733.

Lai, R. K., Fan, C. Y., Huang, W. H., & Chang, P. C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, *36*(2), 3761-3773.

Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2021). Credit risk classification: an integrated predictive accuracy al

gorithm using artificial and deep neural networks. *Annals of Operations Research,40*(7), 1-29.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61-74.

Roy, P. K., & Shaw, K. (2022). Developing a multi-criteria sustainable credit score system using fuzzy BWM and fuzzy TOPSIS. *Environment, Development and Sustainability, 24*(4), 5368-5399.

Srinivasan, V., & Kim, Y. H. (1987). Credit granting: A comparative analysis of classification procedures. *The Journal of Finance, 42*(3), 665-681.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications, 34*(4), 2639-2649.

Wójcicka-Wójtowicz, A., & Piasecki, K. (2021). Application of the Oriented Fuzzy Numbers in Credit Risk Assessment. *Mathematics, 9*(5), 535-536.

Xie, X., Hu, X., Xu, K., Wang, J., Shi, X., & Zhang, F. (2022). Evaluation of associated credit risk in supply chain based on trade credit risk contagion. *Procedia Computer Science, 199*(1), 946-953.

Zhang, L., Hu, H., & Zhang, D. (2015). A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. *Financial Innovation, 1*(1), 1-21.

Zhang, T., Fu, Q., Wang, H., Liu, F., Wang, H., & Han, L. (2021). Bagging-based machine learning algorithms for landslide susceptibility modeling. *Natural hazards,49*(4), 1-24.

Zhao, J., & Li, B. (2022). Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network. *Neural Computing and Applications,34*(8), 12467-12478.