Credit default prediction based on genetic algorithm improved neural network

Zhuo Jiang
School of Management Science and Engineering
Central University of Finance and Economics
Beijing, China

Jie Su
School of Management Science and Engineering
Central University of Finance and Economics
Beijing, China

Ling Zhou
School of Social Development and Public Policy
Beijing Normal University
Beijing, China

2023 9th International Conference on Systems and Informatics (ICSAI) | 979-8-3503-8370-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSAI61474.2023.10423366

Abstract—The article delves into the fundamental objective of credit risk management, which entails evaluating the likelihood of borrowers defaulting on their obligations. This statement underscores the significance of precise risk quantification to mitigate losses, as well as the necessity of a robust risk assessment framework that is bolstered by model algorithms and data. The utilization of extensive datasets and the progression of machine learning algorithms have facilitated the examination of borrowers' data, enabling the prediction of default probabilities and the subsequent evaluation of loan risk. This article provides an overview of the progression of credit assessment models, starting from statistical models and advancing to machine learning algorithms and composite algorithms. It emphasizes the use of ensemble learning algorithms and underscores the benefits of composite algorithms in enhancing the accuracy of these models. The paper introduces the ant colony-BP (back propagation) neural network model as a potential remedy for the limitations observed in conventional BP neural network models. Furthermore, the article delineates the sequential phases entailed in the ant colony approach. The paper examines various data processing strategies, including feature reduction, handling imbalance, and resampling methods. Among these techniques, the SMOTE-Tomek algorithm is found to be particularly effective in mitigating data imbalance. The present paper conducts a comparative analysis of the ant colony-BP neural network model and the XGBoost model, highlighting the superior performance metrics exhibited by the former. The essay finishes by highlighting the practical consequences of the research findings and their significance in the context of data preprocessing and model selection within the financial industry.

Keywords—Neural network, Genetic algorithm, Credit Default, Ant colony algorithm, SMOTE Tomek

I. INTRODUCTION

The primary objective of credit risk management is to evaluate the likelihood of borrowers defaulting on their obligations. When borrowers submit loan applications, lending institutions are required to evaluate their risk using efficient procedures to make informed decisions regarding credit approval. The absence of precise measurement of borrowing risk has the potential to result in financial losses, underscoring the importance of implementing a robust risk assessment framework. The attainment of this objective is contingent mostly upon the assistance provided by exemplary algorithms and data. The accessibility of data collecting, and storage has significantly increased in the era of big data, enabling the development of more precise risk models. Simultaneously, there has been significant progress in the development of machine learning and deep learning algorithms, enabling them to effectively process large-scale, high-dimensional datasets. Consequently, these sophisticated technologies have emerged as viable solutions for addressing such challenges. The borrower's loan risk can be evaluated by the analysis of borrowers' consumption, funding, and credit circumstances using machine learning algorithms and a substantial amount of loan data, which enables the prediction of default likelihood. This strategy can efficiently manage risk and mitigate potential losses. Developing a credit assessment model that can effectively anticipate defaults is of paramount importance. This is particularly true when conventional risk assessment models fail to meet the necessary modeling criteria, leading to inadequate fit outcomes and challenges in identifying key factors that contribute to loan defaults.

This study aims to address the problem of credit default prediction by leveraging prior research in the field, specifically by integrating neural network models with intelligent optimization techniques. Furthermore, the study employs a methodology that integrates components of algorithm optimization and data optimization. Enhancing the interpretability of the model can be achieved by integrating methodologies such as random selection and resampling. The research findings offer valuable information sources for financial institutions, specifically banks, that are entrusted with the task of doing credit assessments.

II. LITERATURE REVIEW

To commence, it is imperative to provide an introduction. Fisher's seminal work on credit risk assessment dates back to 1936 when he created the Fisher Discriminant Analysis [1]. Durand made additional enhancements to the effectiveness of classification by categorizing customers into two distinct groups, namely "defaulters" and "good customers" [2]. The credit rating approach developed by Zeidan et al. involves the utilization of customer sample data to construct models and determine default rates [3]. Historically, there has been a preference among financial organizations and researchers for the utilization of statistical models in credit assessment, with examples being Altman's application of Z-score and regression models [4]. Nevertheless, later research has revealed that the

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on August 27,2025 at 11:02:47 UTC from IEEE Xplore. Restrictions apply.

logistic regression model proposed by Wiginton exhibits notable benefits in terms of both efficiency and accuracy [5]. In contrast, it has been found by certain scholars that linear regression models provide superior interpretability when compared to logistic regression models [6].

The evolutionary progression of credit assessment models the proliferation of technology has led to an expansion in the dimensionality of customer sample data. Consequently, researchers have implemented machine learning techniques. According to Galindo's research, it was observed that the performance of the CART decision tree algorithm outperformed other classification algorithms in specific sample dimensionalities [7]. Malekipirbazari (year) conducted a study in which they employed multiple models, such as KNN, SVM, logistic regression, random forest, logistic, etc., on a consolidated dataset. The results of their analysis indicated that ensemble algorithms, specifically random forest, exhibited notable benefits (Malekipirbazari, year, p. 8). Kruppa and Chopra independently corroborated this finding by doing separate analyses on bank loan data and short-term loan data [9][10]. Subsequent research conducted by scholars revealed that the XGBoost model, along with other ensemble algorithms like CatBoost, had notable benefits in the domain of credit evaluation [11][12]. Moreover, a multitude of neural network techniques have been introduced. Altman was a trailblazer in the application of neural networks to forecast firm financial performance [13]. Nevertheless, it has been discovered by academics that in the age of extensive data analysis, individual algorithms are susceptible to overfitting and being trapped in local optima. To address this issue, Jiang Minghui integrated the particle swarm approach with support vector machines, while Gao et al. merged long short-term memory (LSTM) artificial neural networks with XGBoost models. The aforementioned approaches demonstrated superior performance compared to the individual LSTM model [14]. CK Leong integrated the Bayesian discriminant approach with neural network models to develop a Bayesian network model that effectively addresses the challenges of truncated samples and sample imbalance in credit risk assessment [15].

In conclusion, it can be inferred that the information presented supports the notion that the aforementioned argument in brief, the process of credit assessment has undergone the progression from statistical models to machine learning techniques. The field of machine learning has progressed from the utilization of individual classification models to the adoption of ensemble learning algorithms, and subsequently to the implementation of composite algorithms. Presently, predominant models predominantly comprise machine learning algorithms. Nevertheless, the utilization of composite algorithms has been found to enhance model accuracy when working with diverse datasets. Moreover, this study employs a mixed sampling approach to tackle the issue of reduced interpretability of the model resulting from heightened model complexity, hence augmenting the interpretability of the model.

III. Ant Colony-BP Neural Network

The typical backpropagation (BP) neural network models have the disadvantage of gradient disappearing during the gradient descent process. This is a shortcoming that they suffer from. The occurrence of this phenomenon takes place when the gradient values gradually decrease as they propagate forward, which ultimately leads to training outcomes that are below ideal. The phenomena that have been described can be traced back to the implementation of the chain rule in the backpropagation algorithm during the differentiation phase. As a consequence of this, the weight updates for layers that are located further within the architecture of the neural network are somewhat reduced. Furthermore, the starting weights and biases have a substantial impact on the outcome of the process. This is because different initial values might lead to different convergence outcomes. On the other hand, BP neural network models that are optimized by applying ant colony algorithms have robust global search capabilities, which reduces the likelihood of being stuck in a locally optimal solution. Furthermore, they demonstrate adaptive adjustment skills that are comparable to those shown in ant colonies, which contribute to the optimization of the operational efficiency of the model. One of the most significant benefits of these models is that they can avoid the problem of gradient vanishing. This is mostly because they are not dependent on gradient descent. Within the context of the ant colony algorithm, the following is an outline of the procedural sequence:

- The first step involves reading the input and initializing the topology of the backpropagation (BP) neural network, as well as the parameters of the ant colony algorithm.
- The objective is to ascertain the dimension of the solution space and establish the beginning position of the ants. Commence the iterative procedure of the ant colony algorithm.
- The objective is to determine the quantity of pheromones by considering the spatial arrangement of the ants.
- The position of the individual with the highest pheromone is to be updated. Additionally, the position and pheromone value of the ideal individual are to be recorded.
- The position of the ants and the quantity of pheromones are modified by the probability transition rule.
- The process outlined in steps 3 to 5 should be repeated iteratively until the specified termination condition is satisfied.
- The coordinates of the optimized optimal ant site are obtained and subsequently applied to the BP neural network. The ant colony algorithm is utilized to acquire the appropriate starting weight matrix and threshold vector.
- The optimized backpropagation neural network is trained and tested to evaluate its performance in terms of predictive accuracy. A comparison is made between the predictive accuracy achieved before and after the optimization process.

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on August 27,2025 at 11:02:47 UTC from IEEE Xplore. Restrictions apply.

IV. DATA PROCESSING

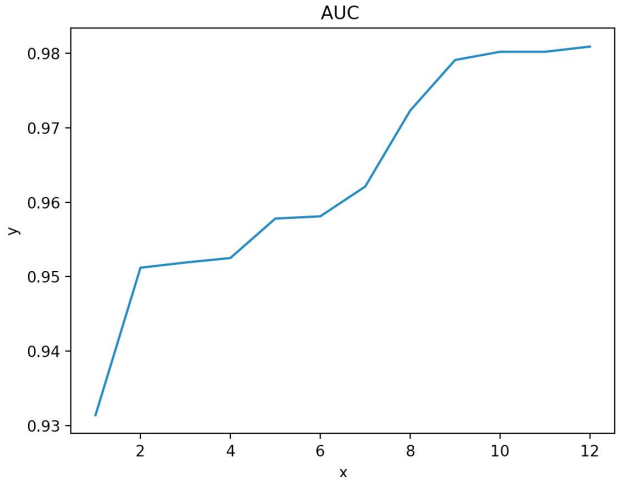
A. Data Source

The dataset utilized in this analysis was acquired from the Karrgle website and was labeled as "credit card." The dataset has credit card information about a total of 284,807 individuals. The number of default cases observed among these clients amounts to a mere 492 instances, or a mere 0.17% of the entire dataset. The percentage in question is of a notably diminutive magnitude. The dataset under consideration exhibits the attributes of an imbalanced dataset, a common phenomenon within this particular setting. This was already referenced in the preceding clause. The dataset does not contain additional explanations of features. Instead, it utilizes feature names such as Time, V1...V30, along with the Class variable, which indicates default circumstances. It is important to note that the variable "Amount" indicates the total loan amount. The data presented in Table 1 unequivocally indicates that the traits V1...V30 exhibit very moderate magnitudes in terms of their variance and extreme values. This is apparent for several reasons. It is crucial to acknowledge, however, that the variable Amount has a substantially wider range and includes extreme values in comparison to the other variables. To mitigate its influence on the outcomes, we have opted to employ Z-score normalization as a means to standardize the quantity variable. Before commencing, it is necessary to examine the characteristics that will be incorporated into the dataset to ascertain the interrelationships among them. The computation of the Pearson correlation coefficient for each variable allows for the determination of the existing correlations between the variables. The analysis reveals that the bulk of the features exhibit non-significant correlations. A considerable proportion of the feature correlations lie within the range of -0.39 to 0.01, as depicted in the figure.

B. Genetic Algorithm Selection

The genetic algorithm demonstrates a strong ability to adapt, allowing it to independently respond to problems that are characterized by different forms and distributions within the feature space. The application of genetic crossover and mutation techniques enables a thorough examination of the search space, hence reducing the likelihood of being confined to local optima. The current research utilizes a genetic algorithm to conduct feature selection. The fitness criteria for this selection process are determined by the area under the receiver operating characteristic curve (AUC). The first step involves the computation of fitness values for each individual, followed by the implementation of the roulette wheel selection method. Following this, genetic processes are performed repetitively until the optimal subset is determined. The assessment of the probability of crossover and mutation procedures is achieved by employing random functions. The initial population size is 50, and the number of iterations is predetermined to be 10. The XGBoost model, which is included in the Sklearn package, is frequently utilized in its default configuration of hyperparameters. Figure 1 illustrates the iteration chart of the genetic algorithm.

Figure 1. Genetic Algorithm Iteration Graph



Following eight iterations, the area under the receiver operating characteristic curve (AUC) demonstrates stability at approximately 0.98. It attains its highest value of 0.9812 during the twelfth iteration. This suggests that the most suitable set of features for the dataset has been determined. The findings from Table 4 indicate that the reduction of dimensions from 30 to 19.

successfully preserves the critical components, resulting in decreased computational complexity and storage demands. This reduction also eliminates extraneous information and noise, enhances the signal-to-noise ratio of the data, and mitigates the potential for overfitting.

TABLE I. REDUCED DIMENSION FEATURES

Before dimension reduction	After dimension reduction
V1-V28, Amount	V2,V4,V6,V9,V10,V11,V12,V13,V14,V15,V16,V17,V18,V19,V21,V22,V23,V24,V27,V28, Amount

C. Handling Imbalance

The presence of an imbalanced dataset and the categorical nature of the final discriminating feature "Class" may result in inferior outcomes if the discrimination algorithm is used directly. Hence, it is imperative to perform dataset resampling to mitigate the problem of data imbalance. Resampling approaches provide as a means to address imbalanced data from a data-centric standpoint. The utilization of resampling strategies can effectively alleviate the influence of imbalanced data. The dataset utilized in this work exhibits a significant imbalance, characterized by a limited number of default samples. Employing the Smote algorithm in isolation will introduce an overwhelming amount of noisy data, hence leading to a decline in the performance of classification. Furthermore, the Smote algorithm is characterized by several limitations stemming from its unsophisticated and aggressive methodology for identifying minority class instances. The utilization of undersampling techniques, such as Edited Nearest Neighbours (ENN), circumvents the issue of generating noisy data by avoiding creating additional samples. In contrast, the hybrid resampling technique known as SMOTE+Tomek

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on August 27,2025 at 11:02:47 UTC from IEEE Xplore. Restrictions apply.

addresses the issue of imbalanced data by augmenting the minority class samples and eliminating neighboring samples. This approach effectively mitigates data imbalance and decreases noise within the dataset. Hence, this work utilizes under sampling, oversampling, and hybrid resampling strategies to mitigate the negative consequences of algorithmic flaws while dealing with data imbalance. The findings are presented in Table 2, as depicted below.

TABLE II. DATA HANDLING METHODS

Over sampling	Under sampling	mixed sampling
SMOTE	***ENN***	***SMOTE+Tomek***

Before employing the techniques delineated in the table, the preprocessed dataset utilized in this work underwent a random partitioning into training and testing sets, with a ratio of 7:3. To assess the effectiveness of the methods stated above, we employed the SKlearn XGBoost model with default hyperparameters for fitting. The assessment metrics employed in this study included Precision, Recall, F1-score, and AUC. The findings are shown in Table 3, as depicted below.

TABLE III. TABLE TYPE STYLES

	Precision	Recall	F1-score	AUC
No process	0.9082	0.6856	0.7865	0.9377
Smote	0.8992	0.6830	0.7763	0.9397
ENN	0.7929	0.7938	0.7852	0.9311
SMOTE-Tomek	0.9043	0.7725	0.7783	0.9314

Upon applying the aforementioned methodologies to the initial dataset, it becomes evident from the tabulated results that there is a lack of significant enhancement in the assessment metrics in comparison to the original dataset. This observation is noteworthy. Conversely, several metrics have exhibited a decline. Since the adoption of the ENN algorithm, there has been a notable decrease in the Precision metric. In contrast, the utilization of hybrid resampling in conjunction with the SMOTE-Tomek technique results in enhancements in all measures, except for the F1-score, to varying extents. In light of this, the SMOTE-Tomek approach will be employed in this study to preprocess the data, so establishing a foundation for further research.

V. CASE STUDY

This paper presents a case study analysis. The dimensionality reduction technique employed in this study has resulted in a reduction of the feature space to 19 dimensions. Consequently, the input layer of the neural network model is composed of 19 nodes, while the output layer consists of a single node. Additionally, the model has five hidden layers, each having a maximum capacity of 10 nodes. Given that the target variable is categorical, the activation function selected for this scenario is the Sigmoid function. The population of ants is established at a value of 50, while the rate of pheromone evaporation is determined to be 1.0. The iteration count is established at 100. Figure 2 and Table 5 present the Receiver Operating Characteristic (ROC) curve and assessment measures, namely Precision, Recall, F1-score, and Area Under the Curve (AUC).

Figure 2. roc curve of ant colony-BP neural network

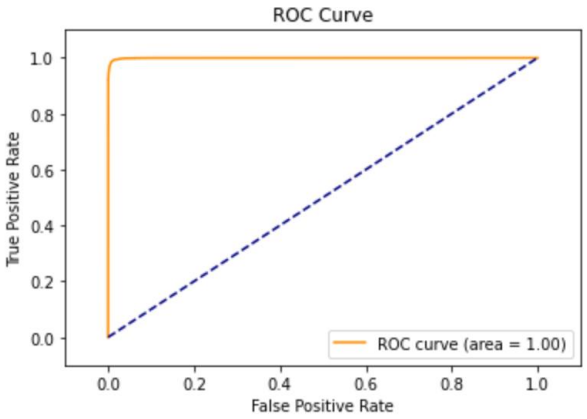


TABLE IV. REDUCED DIMENSION FEATURES

	Precision	Recall	F1-score	AUC
No process	0.9082	0.6856	0.7865	0.9377
SMOTE-Tomek	0.9993	0.9801	0.9871	0.9993

The use of data processing in conjunction with the ant colony-BP neural network model demonstrates enhanced outcomes in comparison to the straight utilization of the XGBoost model without any preprocessing. The improved accuracy of the ant colony-BP neural network model can be attributed to its ability to learn deeper characteristics through adaptive fine-tuning. Furthermore, it has been observed that the XGBoost model exhibits satisfactory performance in straightforward issue domains. However, in the context of intricate problem domains, it is found to be comparatively less successful than the ant colony-BP neural network model. Hence, taking into account the overall performance and practical applicability, the ant colony-BP neural network model emerges as a more favorable option.

VI. CONCLUSION

This article employed Z-score normalization and data visualization approaches to preprocess the dataset and assess the existing imbalance within it. The dataset underwent a reduction process by the implementation of random sampling. This approach was employed to maintain data integrity, while simultaneously mitigating the impact of noise and imbalance. Subsequently, the genetic algorithm GA was employed to select features, leading to a decrease in feature dimensions and noise, while concurrently enhancing the interpretability of the dataset. Additionally, a comparative analysis was conducted on several resampling methodologies, revealing that the SMOTE+Tomek algorithm has shown a significant enhancement in addressing imbalanced data, hence

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on August 27,2025 at 11:02:47 UTC from IEEE Xplore. Restrictions apply.

augmenting the model's data classification capabilities and overall performance. Finally, it was determined that the integration of the ant colony-BP neural network approach resulted in superior performance of the model compared to other models, as evidenced by higher precision, recall, F1-score, and area under the curve (AUC). This research offers valuable guidance and reference for data pretreatment and model selection in real applications, providing a substantial amount of guiding value. ice, with a certain guiding value for practical applications.

REFERENCES

[1] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7.

[2] Durand D. Risk Elements in Consumer Instalment Financing[J]. National Bureau of Economic Research, Inc, 1941.DOI: doi: http://dx.doi.org/.

[3] Zeidan, R., Boechat, C. & Fleury, A. Developing a Sustainability Credit Score System. J Bus Ethics 127, 283 – 296 (2015). https://doi.org/10.1007/s10551-013-2034-2

[4] Altman, E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. J. Financ. 1968, 9, 589–609.

[5] Wiginton J C. A note on the comparison of logit and discriminant models of consumer credit behavior [J]. Journal of Financial and Quantitative Analysis, 1980, 15:757-770.

[6] Chatterjee S, Barcun S. A nonparametric approach to credit screening[J]. Journal of the American Statistical Association, 1970, 65(329): 150-154.

[7] Galindo J, Tamayo P. Credit Risk Assessment Using Statistical and Machine Learning[J]. Computational Economics, 2000, 15(1):107-143

[8] Malekipirbazari M, Aksakalli V. Risk Assessment in Social Lending via Random Forests[J]. Expert Systems with Applications, 2015, 42(10):4621-4631.

[9] Kruppa J, Schwarz A, Arminger G, et al. Consumer credit risk: Individual probability estimates using machine learning[J]. Expert Systems with Applications, 2013, 40(13):5125-5131.

[10] Chopra A, Bhilare P. Application of Ensemble Models in Credit Scoring Models[J]. Business Perspectives and Research, 2018, 6(2):129-141

[11] Xia Y, Liu C, Li Y Y, Liu N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring[J]. Expert Systems with Applications, 2017, 78:225-241.

[12] Ji Q, Ruicheng Y, Pucong W. Application of explainable machine learning based on Catboost in credit scoring[J]. Journal of Physics: Conference Series,2021,1955(1).

[13] Altman E I, Marco G, Varetto F. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience) [J]. Journal of Banking & Finance, 1994, 18(3): 505-529.DOI:10.1016/0378-4266(94)90007-8.

[14] Gao J, Sun W, Sui X, et al. Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model[J]. Discrete Dynamics in Nature and Society, 2021, 2021.

[15] CK Leong. Credit Risk Scoring with Bayesian Network Models [J]. Computational Economics, 2016, 47(3)：423～446.

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on August 27,2025 at 11:02:47 UTC from IEEE Xplore. Restrictions apply.