Migration Letters
Volume: 21, No: 3, pp. 1387-1404
ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)
www.migrationletters.com

Algorithmic Loan Risk Prediction Method Based on PSO-EBGWO-Catboost

Suihai Chen[1], Bong Chih How[2], Po Chan Chiu[3]

Abstract

Loan risk analysis is a common challenge faced by global financial institutions. Under the background of big data, it is of practical significance to prevent loan risk by the machine learning algorithm. Aiming at the characteristics of unbalanced loan data and high noise, this paper proposes an improved Gray Wolf optimization strategy (PSO-EBGWO). PSO-EBGWO is used to optimize the parameters of the CatBoost model. In this method, the Gray Wolf optimized algorithm (EBGWO) is further optimized by particle swarm optimization (PSO), and when combined with it, the convergence performance of the model is improved, the parameters of the model are reduced, and the model is simplified. To a certain extent, it avoids the inefficiency of the Gray Wolf algorithm, balances the ability of local search and global development, and improves the accuracy of the model. Compared with the traditional credit evaluation model, PSO-EBGWO-CatBoost has better accuracy and practical application value.

Keywords: *loan risk; EBGWO; PSO; CatBoost; Accuracy.*

1. INTRODUCTION

With the rapid development of financial economy in our country, credit business has become the main business of banks and financial companies. The quality of customer credit directly affects the business performance of financial companies. To improve their ability to compete in the market, financial companies need to be able to distinguish customers who are likely to be a good investment and those who are not. By identifying and lending to the right customers, they can increase their profits. In fact, there is a large group of customers who fall between being good or bad, and traditional risk assessment methods cannot tell them apart effectively. This is because, the logistic regression algorithm used in traditional risk control model has low accuracy and ability in distinguish good customers from bad customers even though it possesses high speed, high stability, strong explanatory power.

Enhancing the accuracy of a model with high speed, stability, and clear interpretability is the key focus for improving traditional risk control algorithms. The logistic regression algorithm used in the traditional risk control model has the advantages of fast training speed, easy understanding and good model interpretability (Stoltzfus, 2011)[1]. It is suitable for solving linear problems, but its shortcomings are also obvious, such as low accuracy and inability to deal with nonlinear data. Kaastra and Boyd (Kaastra, Boyd, 1996)[2] put forward that the ideal risk control model is able to quickly distinguish good customers (good credit) and bad customers (bad credit) on the premise of good accuracy,

[1] Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), Sarawak, Borneo, Malaysia
[2] Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), Sarawak, Borneo, Malaysia
[3] Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), Sarawak, Borneo, Malaysia

strong stability, strong interpretability, good universality and low consumption of computing resources (enterprise cost). The frequent opinion found is, machine learning algorithms are better than logistic regression algorithms. Many scholars (Ma, Wang, Yang, 2018)[3] currently use XGBoost and LightGBM algorithms for data testing, and the results are significantly better than the traditional logistic regression algorithm. Many scholars (Chang, Chang, 2018)[4] suggest using XGBoost instead of logistic regression. A few scholars (Ma, Sha, Wang, 2018)[5] have also combined logistic regression algorithm with XGBoost algorithm and achieved some results. However, XGBoost also has its drawbacks, such as overfitting, difficulty in parameter tuning, slow learning speed, and LightGBM's algorithm is not as accurate as XGBoost's [6](Liang, Luo, Zhao, 2020). The new CatBoost algorithm compensates for the shortcomings of XGBoost and LightGBM in many ways.

CatBoost algorithm has high accuracy, interpretability and time complexity. Therefore, many experts begin to study it to replace the traditional risk control model algorithm. However, the CatBoost algorithm also has its shortcomings. The algorithm needs to adjust its abundant parameters in order to have high accuracy. Since it has many parameters, so it is impossible to manually adjust one by one. Therefore, it is necessary to use other algorithms to help CatBoost algorithm to automatically Determine the ideal parameters for the Grey Wolf Optimizer (GWO) algorithm introduced by Seyedali Mirjalili. (2014)[7] 2014 has a simple structure and strong advantages in finding the optimal parameters. How to combine these two algorithms to make GWO algorithm quickly find the optimal parameters of CatBoost algorithm, while avoiding GWO algorithm falling into local optimal solution, and balancing the exploration and development ability of the algorithm needs further research.

How to avoid the random search behavior of GWO in the parameter search process, avoid the algorithm falling from local optimum, realize the balance between exploration and exploitation ability, and find the most suitable parameters for the CatBoost algorithm.The main processes for GWO optimization are as follows: The first step requires the improvement of the GWO algorithm, referring to as the "EBGWO" algorithm. The EBGWO algorithm is an enhanced version of the Grey Wolf Optimization algorithm, possessing strong global search capabilities and, to some extent, mitigating the local convergence issues associated with the Grey Wolf Optimization algorithm. It can effectively enhance the predictive accuracy and execution speed of the CatBoost algorithm model. The second step is that although the optimized GWO algorithm (EBGWO) can effectively avoid premature convergence and local optimization, its lack does not consider individual experience and lacks communication between individual and group positions [8] . The core principle of the Particle Swarm Optimization (PSO) algorithm involves the continuous movement of particles at variable speeds within a space. These particles navigate by leveraging their individual memories and group communication to determine the next position. Each particle adjusts its trajectory in pursuit of the optimal individual position. (Reddy, Viswanath, 2018)[9].

Hence, by incorporating the update mechanism of particle positions, the Gray Wolf algorithm can imbue a level of memory in its optimization process, effectively replacing the individual position update for Gray Wolves, and the EBGWO algorithm has the communication ability between the individual and the group position, so as to solve the problem. The following chapters will introduce the detailed process of PSO-EBGWO-CatBoost algorithm, and carry out experimental test verification.