Journal of Software Engineering

ISSN 1819-4311

Academic
Journals Inc.

www.academicjournals.com

OPEN ACCESS Journal of Software Engineering

ISSN 1819-4311
DOI: 10.3923/jse.2016.297.301

Research Article
Combined Forecasting Model Based on Cuckoo Search Algorithm for Personal Credit Assessment

Yong-bin Zhu and Jun-sheng Li

Engineering College, Honghe University, 661199 Mengzi City, China

Abstract

In the establishment of a personal credit assessment model the prediction accuracy is very important. For the deficiency of the single model in personal credit assessment, this study puts forward the method of using the combined forecasting model to carry on it. Based on the advantages of different single model in personal credit assessment, the logistic regression and linear regression are chosen to construct this combination model and the improved cuckoo search algorithm (CS) is used to solve the model weights, a combination forecasting model based on CS is constructed for personal credit assessment. The results show that the model construct in this study can effectively reduce the second class misjudgment rate of personal credit assessment and effectively improve the forecasting accuracy, which has better applicability and significant value for commercial banks to control the consumer credit risk.

Key words: Cuckoo search algorithm, combined forecasting, individual credit assessment

Received: January 12, 2016 Accepted: February 29, 2016 Published: June 15, 2016

Citation: Yong-bin Zhu and Jun-sheng Li, 2016. Combined forecasting model based on cuckoo search algorithm for personal credit assessment. J. Software Eng., 10: 297-301.

Corresponding Author: Jun-sheng Li, Engineering College, Honghe University, 661199 Mengzi City, Yunnan, China Tel: 0873-3694084

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Personal credit assessment refers that rating agencies conduct a comprehensive analysis and evaluation for consumer's personal credit information and performance capabilities by using the credit scoring models, to predict default risk for future lenders to provide a scientific basis for consumer credit decisions (Wang and Liu, 2014). In the personal credit system construction earlier country, with the study of personal credit assessment continues to develop and mature, many methods have been applied to the field of personal credit assessment (Zhang, 2005). At present, individual credit assessment methods are mainly two types, one is based on statistical models, including multiple linear discriminate analysis, linear regression, logistic regression and another kind is non-parametric estimation and artificial intelligence methods, including neural networks, classification tree, genetic algorithms, etc. (Li *et al*., 2012). The construction of personal credit system in our country is relatively late, methodology is still not perfect, which seriously restricts the development of consumer credit in our country (Chen and Jiang, 2015). Therefore, to establish and improve the personal credit rating system which is suitable for the situation of our country has important significance. Combination forecasting model based on cuckoo search algorithm (Yang and Deb, 2009) (Cuckoo Search, CS) to solve the model weight is established in this paper for personal credit assessment and compare with the single model to study applicability of the model.

MATERIALS AND METHODS

Basic principle of combination forecasting: The combination forecasting obtains the results by weighted restructuring other prediction methods, Clemen has pointed out that the combination is projected to become one of the mainstream prediction studies (Clemen, 1989). In combination forecasting theory, in the manner of assembly of each single prediction model, it can be divided into linear and nonlinear combinations, thereof, wherein the linear combination forecasting model is the most studied and the most widely used (Ma and Tang, 1998). The basic principle of linear combination forecasting is as follows: For a prediction problem, there are m types of prediction methods, y_t represent actual observed values (sample size t = 1, 2, ..., n), f_{it} represents the predicted value of the i-th prediction method (t = 1, 2, ..., n), w_i is the weight of the i-th method in combination forecasting model, so, the linear combination forecasting can be formulated as follows:

$$f_t = \sum_{i=1}^{m} w_i f_{it} \text{ s.t. } \sum_{i=1}^{m} w_i = 1, t = 1, 2, \cdots, n, E = \sum_{t=1}^{n} (y_t - f_t)^2 \qquad (1)$$

The key is to determine the weight of each individual forecasting model and the least square method is usually used to solve a set of weights to minimize its error sum of squares E. It is generally believed that the result of the combination forecasting method is more accurate than a single one because of using the more information.

Basic cuckoo search algorithm: Cuckoo search algorithm is a type of new search algorithm derived from the biological behaviors of parasitic and reproduction of cuckoo population in nature and combined with the birds and drosophila special Lévy flight mode, global search capability is strong, suitable for multi-objective optimization problem solving. The algorithm is based on 3 ideal rules to construct, the specific rules are also found in the literature (Gandomi *et al*., 2013). The cuckoo searches the path and the position of the bird's nests under the premise of the three ideal rules and the formula is as follows:

$$x_i^{(t+1)} = x_i^{(t)} + \partial \oplus \text{Lêvy}(\lambda), i = 1, 2, \cdots, n \qquad (2)$$

where, $x_i^{(t)}$ represents the position of Gandomi nest of generation t, ∂ represents step-size information, $\oplus$ represents the dot product, Lévy (λ) represents random search path. After a certain amount of P_a was found to be discarded, the same number of new nest was generated by random walk. The position update formula is as follows:

$$\chi_i^{(t+1)} = \chi_i^{(t)} + r\left(\chi_j^{(t)} - \chi_k^{(t)}\right), \left(j, k = 1, 2, \cdots, n\right) \qquad (3)$$

where, r is a random number belongs to (0, 1), $x_j^{(t)}$ $x_k^{(t)}$ is the two random bird nest location in generation t.

Algorithm improvement: In order to solve the problem of that the algorithm is easy to skip the optimal solution and to occur shock phenomenon in later iterations to cause convergence speed and optimization accuracy to reduce, the algorithm was improved according to the method of literature (Qu *et al*., 2014) in this study. At the same time, the algorithm adopts adaptive step walk way method and Gaussian disturbances strategy to improve the search accuracy and convergence speed and the Manteganna equivalent calculation algorithm is used to reduce the computational complexity. The path and position update formula of the nests was modified as follows:

$$\begin{aligned} & x_i^{(t+1)} = x_i^{(t)} + \text{stepsize}_i \oplus \text{randn}(),\ i = 1,\ 2, \cdots, n \\ & \text{stepsize}_i = \alpha \cdot \text{step} \oplus \left(x_i^t - x_{\text{best}}^t \right) \end{aligned} \quad (4)$$

wherein, randn() is a random function to meet Gaussian distribution, *a* represents walk coefficient, step represents walk step. The adaptive step walk adjustment strategy and the dimensional Gauss perturbation strategy are detailed in the literature (Qu *et al*., 2014).

Model building ideas: Individual credit assessment is essentially a kind of classification problem in pattern recognition. The principle of, which is consumer credit applicants can be divided into repaying and default categories, which made the decision to accept or reject credit applications. In the practice of credit assessment, there are usually two kinds of miscarriage of justice: One is to put the customer with good credit as the bad credit to reject the loan application, the other is to accept the loan application of the customer but with bad credit. Generally, in practice, the latter cause greater loss to the credit agency compared to the former. Therefore, the individual credit assessment, at the same time improve the classification accuracy should try to reduce the occurrence of the second miscarriage of justice. In this study, firstly, linear regression and logistic regression were used to establish a single prediction model and then the linear combination forecasting model is built on this basis and CS algorithm is used to solve the weight (Lu *et al*., 2003). With the single model classification results were compared to investigate the applicability of combination forecasting model based on CS algorithm.

Sample data and pretreatment

Sample data selection: The data used in this study come from the consumer credit database of a credit institution. The classification of "Whether to default" is based on the "Default frequency", e.g., the frequency of the repayment of the loan is lagging or insufficient. In the practice of most countries, it is generally believed that in the last year, the number of breach of contract for more than 4 times, it is considered that the customer has a strong tendency to default. This study also classified by the method and the index was excised for missing attribute more serious. When selecting a sample data, the stratified sampling method was used to divide the sample into two types of default and not default and the two kinds of data to keep the same number of data in order to reduce the uneven impact on model classification capabilities. 1108 sample data were selected as the experimental data and the data were randomly divided into two groups, respectively, 554 sample data, including 277 default sample data and 277 non default sample data were respectively used to build models and test models of classification.

Normalized sample data: In order to eliminate the influence of dimension and reduce the influence of data imbalance on classification ability, the training data and test data are normalized (Lu *et al*., 2003). All explanatory variables are divided into two groups of discrete variable and continuous variable. For discrete variables (including x_1, x_3, x_4, x_5, x_7, x_8 and x_9), the minimum-maximum normalization method is used, as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (5)$$

where, X' in (0, 1), indicates the normalized values of variables, X_{min} and X_{max} represent the minimum and maximum value of the variable X, respectively. For continuous variables (including x_2, x_6 and x_{10}), the distribution of variable values are approximative to normal distribution and the following method is used to deal with it.

$$X' = \Phi\left(\frac{X - \mu}{\sigma}\right),\ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad (6)$$

where, μ, σ separately represent the mean and standard deviation of variable X, $\Phi_{(x)}$ represents the cumulative probability of standard normal distribution.

Construction and application of the model

Single statistical model: As the basis of the combination forecasting model, the linear regression and Logistic regression model are established, respectively at first.

Linear regression requires that the distribution of the explanatory variables only obey certain preconditions and the better results can be obtained (Wang and Zhang, 2010).

In these preconditions, an important assumption is that there is no strong correlation between the explanatory variables, namely the absence of multicollinearity.

Therefore, the SPSS software is used to establish the linear regression model and in order to eliminate the influence of colinearity between explanatory variables of the model, the method of variables into the model is stepwise entering method and the results obtained are as follows:

$$y_1 = 0.729 - 0.843x'_4 + 0.290x'_5 - 0.297x'_6 + 0.138x'_7 - 0.130x'_9 \quad (7)$$

Table 1: Classification results contrast

	Training sample			Test samples		
Models	First class error	Second class error	Classification accuracy (%)	First class error	Second class error	Classification accuracy (%)
Linear regression	18 (6. 50)	41 (14.80)	89.35	15 (5.42)	32 (11.55)	91.52
Logistic regression	19 (6.86)	29 (10.47)	91.34	17 (6.14)	26 (9.39)	92.24
Combination forecast	17 (6.14)	23 (8.30)	92.78	14 (5.05)	19 (6.86)	94.04

The regression equation adjusted R^2 of 0.643, the coefficient of formula 6, t-test and F-test, the results show that the regression equation is valid. The regression equation is used to test samples and 0.5 as the classification boundary, that is, if the predicted result is greater than 0.5, it is judged as non default class, otherwise judged to default class, the results obtained are shown in Table 1. The linear regression for personal credit scoring there is a drawback: The right of the value of the regression equation from -∞ to +∞, but the left side of the equation is a probability, it will be within the range (0, 1). If the left side of the equation is a function of P, it can take any value and then the model will be more meaningful (Zhang, 2005).

Logistic regression is based on linear regression. The probability of the Logit, that y = ln (p/(1-p)), overcomes the shortcomings of linear regression (Shi and Zhang, 2005). To establish Logistic regression models with SPSS, variable selection method use backward: Conditional method (with assumed parameters based on probability as the likelihood ratio test, to select arguments by backward stepwise selection), the Logistic regression equation obtained in this study is:

$$\hat{y}_2 = \frac{\exp\left(0.965x_1^{'} - 6.209x_4^{'} + 1.897x_5^{'} - 3.168x_6^{'} + 1.197x_7^{'}\right)}{1 + \exp\left(0.965x_1^{'} - 6.209x_4^{'} + 1.897x_5^{'} - 3.168x_6^{'} + 1.197x_7^{'}\right)} \quad (8)$$

$\hat{y}_2$ represents y as 1, the probability of that is not in default and 0.5 as a classification boundary, prediction classification results of the equation on the training and testing samples are shown in Table 1.

Combined forecasting model based on CS algorithm: According to the principle of combination forecasting, a combined model based on linear regression and logistic regression is constructed in this study as follows:

$$\begin{aligned} & f = w_1\hat{y}_1 + w_2\hat{y}_2 \\ & \text{s.t} \quad w_1 + w_2 = 1 \end{aligned} \quad (9)$$

The results of linear regression and logistic regression model are used as the input vector to establish the combination forecasting odel based on CS algorithm (Wang *et al*., 2012). CS algorithm after 100 iterations, the optimal weights are obtained as follows:

$$w_1 = -0.74985, \quad w_2 = 1.\ 74985 \quad (10)$$

Thus, the combination forecasting model herein was obtained as follows:

$$f = -0.\ 74985\hat{y}_1 + 1.74985\hat{y}_2 \quad (11)$$

wherein, f_1 and f_2 are, respectively predicted results of linear regression and logistic regression model. The prediction results of linear regression and Logistic regression models on the test sample into the formula 11 and 0.5 as the classification boundaries, the classification result is shown in Table 1.

RESULT ANALYSIS

Then, the comparative analysis is made from the classification accuracy and the two types of erroneous judgement between two kinds of single model and combination forecasting model based on CS algorithm. First, in terms of classification accuracy, it can be seen from Table 1, in the modelling and testing samples, combined forecasting model based on CS algorithm is higher than linear regression and Logistic regression model. This shows that the combination model has the advantage of combining the advantages of a variety of single model, which has the advantage of a single model in personal credit evaluation. Secondly, it can be seen from Table 1, either the accuracy, false positive rate of the first category ("Remove true", the "Good" credit misjudge to "Bad" credit) or a second class false positives ("Accept error", the "Bad" credit is judged as "Good" credit), on modelling and testing samples, the combination forecasting model based on CS algorithm are lower than any single model, especially to improve the rate of the second miscarriage of justice ("Accept error") is more obvious, the second category misjudgement of these two single statistical models are higher than the first class misjudgement.

DISSCUSSION

Cuckoo search algorithm is a novel bionic optimization algorithm, which is simple and efficient and successfully applied to the classical theory research and engineering applications. Since, the dimensional mutual interference phenomenon of the basic CS algorithm will affect the partial solution precision capability and convergence speed. Adaptive step walk adjustment strategy and dimensional Gauss perturbation strategy are used in this study to improve the convergence rate and the quality of the solution. The combination forecasting model based on CS algorithm is constructed. From the analysis of experimental results, from the classification point of view, combination forecasting model is superior to a single model (compared with the linear regression and Logistic regression models), especially in avoiding the second miscarriage of justice ("Accept error"), through the comparative analysis of the above data shown in Table 1, the combination forecasting model based on CS algorithm has better applicability for the practice of avoiding the credit risk.

CONCLUSION

In this study, the combination forecasting model is used for personal credit assessment. Based on the construction of linear regression and logistic regression model, a combination forecasting model based on CS is constructed and the weights are solved by using the global search ability of CS algorithm. Through empirical research, this study draws the following conclusions:

- Classification accuracy of combination forecasting model is higher than the two single statistical models; therefore, it has advantage to improve the accuracy of classification
- In the case of that the first false did not increase, the combination model can further reduce the second kinds of false and it is more important for commercial banks to control credit risk

ACKNOWLEDGEMENTS

This study was financially supported by projects of “Basic research on the application of Yunnan Province” (No.2013fz127) and “Science Research Fund of Yunnan Provincial Education Department” (No.2014C136Y).

REFERENCES

Chen, H. and M. Jiang, 2015. [Rethinking of the methods of personal credit behavior evaluation]. Acad. Exchange, 12: 139-143.

Clemen, R., 1989. Combining forecasts: A review and annotated bibliography. Int. J. Forecasting, 5: 559-583.

Gandomi, A.H., X.S. Yang and A.H. Alavi, 2013. Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems. Eng. Comput., 29: 17-35.

Li, T., S. Li and Z. Zhou, 2012. Genetic algorithm applied research in the individual credit portfolio scoring. Proceedings of the 5th Annual Meeting of the Risk Analysis Council of the China Association for Disaster Prevention 2012, October 27-28, 2012, Nanjing, China, pp: 427-431.

Lu, Q., P.L. Gu and S.M. Qiu, 2003. The construction and application of combination forecasting model in chinese energy consumption system. Syst. Eng.-Theory Pract., 3: 25-31.

Ma, Y. and X. Tang, 1998. Research on the problem of optimizing linear combination forecasting model. Syst. Eng.-Theory Pract., 9: 110-114.

Qu, C.W., Y.M. Fan and J. Dai, 2014. On a forecasting model of grey neural network based on improved cuckoo search optimal algorithm. J. Southwest China Normal Univ. (Nat. Sci. Edn.), 39: 131-136.

Shi, C. and M. Zhang, 2005. Analysis of logistic regression models. Comput. Aided Eng., 14: 74-78.

Wang, D. and Z. Zhang, 2010. Variable selecting for linear regression models: A survey. J. Applied Stat. Manage., 29: 615-627.

Wang F., X. He, Y. Wang and S.M. Yang, 2012. Markov model and convergence analysis based on cuckoo search algorithm. Comput. Eng., 38: 180-182, 185.

Wang, T.Q. and X.Q. Liu, 2014. A study on individual credit evaluation for commercial bank based on PCA-GA-BP. Value Eng., 31: 161-163.

Yang, X.S. and S. Deb, 2009. Cuckoo search via Levy flights. Proceedings of the World Congress on Nature and Biologically Inspired Computing, December 9-11, 2009, Coimbatore, India, pp: 210-214.

Zhang, X., 2005. [Vigorously promote the construction of the personal credit system in China]. Dev. Res., 10: 72-74.