ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 141

International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017)

Internet Credit Risk Scoring Based on Simulated Annealing and Genetic Algorithm

Ji Hu[1,*] and Jiawen Cai[2]

[1]Business School Sichuan University China

[2]Economics School Sichuan University China

[*]Corresponding author

***Abstract*—Credit risk which can be reduced by credit scoring is the focus of financial risk control. In the construction of Internet credit scoring model, we encounter the problem that variables have high dimension. To solve this problem, feature selection is necessary. Simulated annealing and genetic algorithm can be used to do feature selection. This article gives an empirical analysis of individual credit valuation using data from a microfinance internet platform. Experimental result shows both of the logistic model based on simulated annealing and logistic model based on genetic algorithm have better prediction ability and model interpretability comparing to traditional full variable logistic regression. Also, the logistic model based on simulated annealing is slightly better than logistic model based on genetic algorithm**

Keywords-credit scoring; benchmark experiment; feature selection; simulated annealing; genetic algorithm

I. INTRODUCTION

Credit plays an important role in the financial field. With the continuous expansion of credit scale in recent years, credit risk control becomes the focus of financial institutions and P2P reception platform. By constructing credit scoring model, we can make evaluation on new customers' credit accurately and better manage customer relationship. High credit rating of customers can enjoy more quality, more convenient loan service such as lower loan interest rate and longer loan term.

Over the last decades, there have been lots of classification models and algorithms applied to analyze credit risk, for example decision tree [1], nearest neighbor K-NN, support vector machine (SVM), neural network [2]. But in dealing with Internet credit scoring problem, these models and algorithms show some disadvantages. First, these methods may result in over fitting. Second, when the data set we use to construct a model has a huge system index, these methods need a long computation time.

Data source for traditional credit scoring is limited. Most of the data comes from financial institutions and these data updated slowly, similar to static. But with the widespread use of Internet technology nowadays, the amount of data we can use to make credit scoring increases exponentially which means we need to analyze a huge evaluation index system. However, current research on Internet credit scoring is insufficient. In order to solve the problem of high dimensionality of variables, feature selection in modeling is necessary.

What is feature selection? It means selecting N comparatively more important features out of existing M features. The objective of feature selection is to avoid over fitting and improve the model performance, to reduce variable dimensionality and produce cost-effective model. Feature selection still needs to pay a price for these advantages. To get the best variable subset, it needs to search in the full index system over and over again. In this process, determining coefficients for different variable subsets and evaluating model performance every iteration increase burden of modeling task[4] On the whole, there are three kinds of general feature selection methods including wrapper、filter and embedded methods. Wrapper method means the algorithm has intrinsic feature selection functions while filter methods filters variables before conducting target algorithm. Embedded methods is fusing variable selection task in the process of model training. For instance, branching process in decision tree uses embedded method, which is based on a certain intrinsic metrics for feature ranking. Many scientists have proposed a lot of feature selection approaches and the details can be found in a review [3].

Two of wrapper feature selection methods are simulated annealing and genetic algorithm. Metropolis proposed Metropolis acceptance criteria which paved ways to simulated annealing method [4]. Simulated annealing was first developed and applied to the optimization of functions having many local optimums by Bohachevsky [5]. After that, a lot of scholars put forward improved algorithms of simulated annealing [6]. Genetic algorithm was first proposed by Holland in the early 1970s [7]. It was widely used in many fields soon [8].

Simulated annealing and genetic algorithm are rarely used in credit scoring focusing on feature selection. Hand, David J described a way to find the best partition of each variable using a simulated annealing strategy [9]. Jiang Yi proposes a new credit scoring model based on decision tree and simulated annealing algorithm which proved effective [10]. A simulated annealing based rule extraction algorithm was designed for credit scoring problem [11]. In researches above, the focus is the function of simulated annealing and genetic algorithm in avoiding local optimums. A.J. Cuticchia et al present a method of combinatorial optimization, simulated annealing, to order clones in a library with respect to their position along a chromosome [12]. It explores the feature selection function of simulated annealing. But the applied field is biology. Actually genetic algorithm is commonly used in credit scoring. Chi B W came up with a new hybrid approach to integrate genetic

algorithm into dual scoring mode [13]. But these researches did not integrate GA with logistic regression and there is no comparison between simulated annealing and genetic algorithm.

In this paper, we apply simulated annealing and genetic algorithm based on logistic regression model in Internet credit scoring for the purpose of feature selection and we compare the performance and computing speed of simulated annealing with GA and full variable logistic regression.

II. METHODS

A. Simulated Annealing

Metropolis acceptance criteria is the foundation of simulated annealing. Its' core idea is as following: in a certain temperature, the current state of s generates a new state of s', the energies of both are E(s) and E(s') respectively. If E (s') <E (s), the new state s is accepted as the current state. Or calculate the probability p.

$$p = \exp\left(-\frac{E(s') - E(s)}{T}\right) \quad (1)$$

If P is larger than a stochastic number, accept new state of S' as current state. Otherwise keep s as current state. [6] Based on Metropolis acceptance criteria, simulated annealing is first proposed by Bohachevsky simulating the process of metal cooling [7]. In briefly, it's the iteration process of "update-judge-accept or reject". It correspond an x and an objective function f(x) with a state s and the energy E (s) of the solid annealing process, and a temperature parameter t is introduced to simulate the annealing process. In the process of selecting the better state, the t value is gradually attenuated. Finally it can get the global optimal solution. The specific steps of the algorithm are as follows:

Algorithm 1

1 Generate an initial random subset of predictors

2 for iterations i=1…t, do

Randomly perturb the current best predictor set

[Option]Pre-process the data

Tune/train the model using this predictor set

Calculate model performance (E_i)

if $E_i < E_{best}$ then

Accept current predictor set as best

Set $E_{best} = E_i$

else

Calculate the probability of accepting the current predictor set $P_i^a = \exp[(E_{best} - E_i)/T]$

Generate a random number U between [0,1]

if $P_i^a <= U$ then

Accept current predictor set as best

Set $E_{best} = E_i$

else

Keep current best prediction set

end

end

end

Determine the predictor set associated with the smallest E_i across all iterations

Finalize the model with this predictor set

B. Genetic Algorithm

A distant kin to simulated annealing is genetic algorithm. GA simulates the revolutionary process of human genetic. It is effective at finding optimal solution since it allows the current population solutions to reproduce, generating children which compete to survive. The process of combination and mutation will not stop until some artificially set conditions are met such as reaching the upper limit of generations, newly produced chromosome having the characteristics we want and etcetera. In GA algorithm, volumes of genes make up one chromosome. A gene represents the absence or presence of a predictor in the data. So the chromosome which is a binary vector has the same length as the number of predictors in a dataset. The competition ability of a chromosome is determined by the performance of the model containing predictors indicated by the binary vector.

First, a set of chromosomes are randomly selected from all possible chromosomes. Then we calculate the competition ability of these chromosomes and choose comparatively the best two chromosomes to reproduce. In the reproduction phase, the two chromosomes which stand for two conditions of variable selection begin to split at a random position. The tail of one chromosome is combined with the head of the other one. After crossover, a randomly selected gene in the newly produced chromosome has a certain probability to mutate. That means the binary value of this gene changes from one value to the other value. In other words, the mutation process represents a new predictor is added in the variable subset or one predictor in the original variable subset is removed. Algorithm 2 lists these steps.

Algorithm 2

Define the stopping criteria, number of children for each generation (GenSize), and probability of mutation (p_m)

Generate an initial random set of m binary chromosomes, each of length p

Repeat

For each chromosome do

Tune and train a model and compute each chromosome's fitness

End

For reproduction k=1…GenSize/2 do

Select two chromosome based on the fitness criterion

Crossover: Randomly select a loci and exchange each chromosome's genes beyond the loci

Mutation: Randomly change binary values of each gene in each new child chromosome with probability, p_m

End

Until stopping criteria is met

Only the selected two best can implement the crossover process which may produce subsequent optimal solution. Maybe other chromosomes can reproduce better solution. The mutation process provides chances to escape from local optimum for it makes little change from the reproduced chromosome. Usually the mutation probability is kept low (Pm<=0.05), but the practitioner can set a higher probability if the focus is avoiding local optimum.

GA has been shown to be an efficient feature selection tool in the fields of chemistry [14], image analysis [15], and finance ([16] [17]).

III. EXPERIMENT

A. Dataset

In this paper, the data we use comes from a small credit platform. There includes 5000 valid records in this dataset. Each credit record contains 300 attributes explaining the customer's individual circumstances and the target variable. The customers' individual circumstances mainly comprise five dimensions of information: customer's credit history, behavior preference, debt paying ability, identity and interpersonal relationship. Customer's credit history provides information about loan amount, repayment, guarantee, Hydropower payment, while behavioral preference describes Internet searching preference, shopping preference and so on. Debt paying ability is reflected by the status of customers' assets, income, financial investment, etc. Identity information concludes name, gender, living address and so on. Interpersonal relationship means social networks built on social platforms. The target variable is a binary variable whose value is 0 or 1. 0 means non default while 1 means default. In these 5000 personal credit records, 4574 individuals are defined as non-default customers and the other 426 people are defined as default customers.

TABLE I. VARIABLE INTRODUCTION

	Five dimensions		
	Customers' credit history	Behavior preference	Debt paying ability
Variables under each dimension	loan amount, repayment, guarantee, Hydropower payment, etc.	Internet searching preference, shopping, preference, etc.	customers' assets, income, financial investment, etc.
	identity	***interpersonal relationship***	
Variables under each dimension	name, gender, living address etc.	social networks built on social platforms	

B. Data Preprocessing

There are seven character variables in the above 300 variables and first we turn them to numeric type. Then we explore the missing value in this dataset. Result shows there are 47.7% variables having more than 40% of missing values. We replace missing values with the mean value of each column. The character variables which we changed to numeric variables before need to be changed back to character type or they are ordered. Then we check whether there are same samples in this dataset and the truth is that no same samples exist. After all these steps, we change the target variable into character type in order to obtain AUC value and confusion matrix when modeling.

In the phase of data division, we randomly selected 70% samples from the data set which contains 5000 samples as the training set, and the remaining 30% samples were used as the test set. More precisely, 3501 samples are divided in the train set and the other 1499 samples are in the test set.

The whole data preprocessing steps can be found in the following graph clearly.

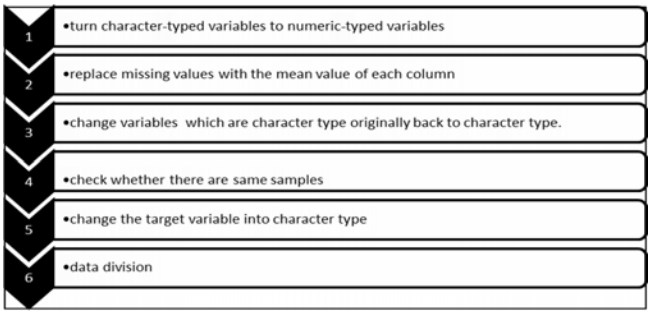

FIGURE I. DATA PREPROCESSING

C. Model Construction and Evaluation

In this paper, we construct traditional full-variable logistic regression model, simulated annealing logistic model and genetic logistic model on this dataset. The 300 predictors are the inputs of models and the target variable is the output. We train models on training set and evaluate models on test set. AUC value, sensitivity, specificity, overall accuracy and the number of selected variables are used for evaluation metrics.

FIGURE 2 shows the process of selecting optimal model using simulated annealing method. We totally do 2000 iterations. At every iteration, we use AUC as evaluation metric. If AUC of new model is higher than that of previous model, we accept the new model directly. Otherwise we accept the new model with probability P. The initial value of t is 0.001.

$$p = \exp(new - old/t) \tag{2}$$

So every iteration, t will decrease by 0.001/(nrounds+1). From the graph we can see that with the increase of the number of iterations, the AUC value fluctuates continuously, but the overall trend is gradually rising. The amplitude of increase is decreasing and when the number of iterations is close to 2000, the amplitude of AUC increase is very small. As what we recorded, Completing 2000 iterations only requires 107 minutes which is much shorter than backward variable selection process. Because using backward variable selection method to construct logistic model on a big dataset which has 300 variables and 5000 samples needs 752 hours on R software. That is time consuming.

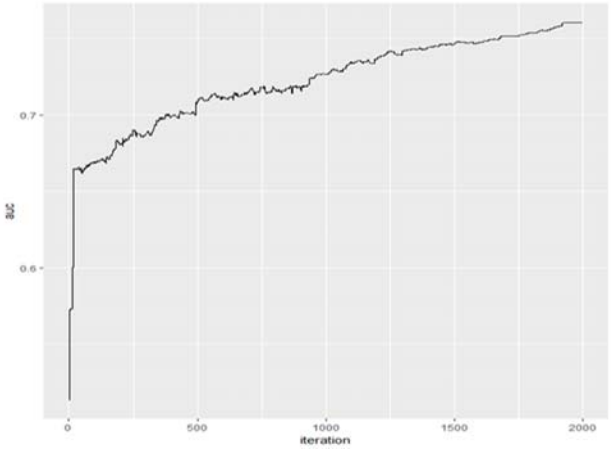

FIGURE II. SIMULATED ANNEALING METHOD

FIGURE 3 represents the process of selecting optimal model using genetic algorithm. We do 30 iterations and at every iteration we keep 30 populations. From these 30 populations we choose 50% of populations as parents which have higher AUC than the remaining populations. Parents begin to cross over and mutate to produce 15 populations. We set the mutation probability to 0.4. The 15 children will replace the population that was eliminated in the previous round. In graph 3, at each iteration, there is a box plot. The box plot shows the overall distribution of AUC values for 30 populations. Actually, as the number of iterations increases, the average of AUC value increases and the difference gap of AUC between 30 populations become more and more narrow. Completing this part of experiment costs 87 minutes.

To calculate sensitivity and specificity of these three models, we should first get the confusion matrix of each model. Table 1 shows the general form of a confusion matrix. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions. Formula (3),(4),(5) give the function of sensitivity, specificity and overall accuracy.

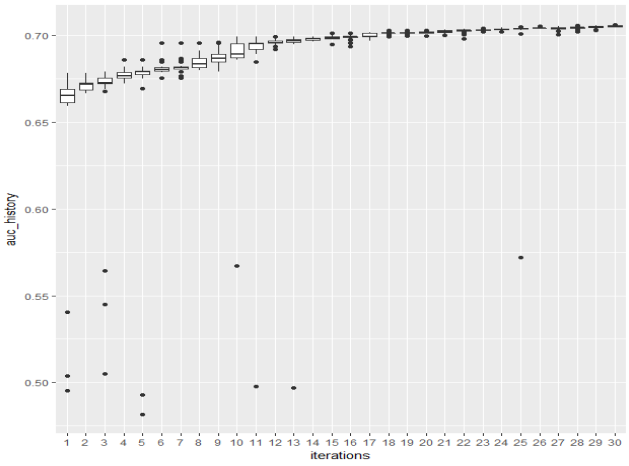

FIGURE III. GENETIC ALGORITHM

TABLE II. CONFUSION MATRIX

True condition	Predicted condition	
	Prediction positive	Prediction negative
Condition positive	True positive(TP)	False negative(FN)
Condition negative	False positive(FP)	True negative(TN)

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

$$Overall\ accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

The ratio of default customers and non-default customers in test set is 299/3501. We use this ratio as threshold to obtain the confusion matrices from these three models. Table 3 presents the result. The accuracy, sensitivity, specificity, AUC and number of selected variables are shown in table 4. It is easy to find that simulated annealing model has highest sensitivity among three models which is 0.6378. Although the sensitivity gap between simulated annealing and genetic algorithm is very small. They are both much higher than that of full logistic model (0.1575). What's more, specificities for simulated annealing logistic model and logistic model based on genetic algorithm are similar. However, they are a little lower than that of full logistic model. In aspect of AUC, the AUC of full logistic regression model is 0.5408 which is much lower than those of the other two models comparing to 0.7604 and 0.7062 separately. Respect to variable selection, the numbers of variables selected by simulated annealing logistic model and

logistic model based on genetic algorithm are very close. Both are nearly a half, as against that of full logistic regression model.

On the whole, simulated annealing logistic model is slightly better than logistic model based on genetic algorithm. Full logistic regression model has good performance in specificity. But its' sensitivity and AUC are the lowest among three models. Credit scoring model has a characteristic that the costs of misclassification for different types of customers are different. The cost of misclassifying a default customer is much higher than that of misclassifying a non-default customer. So, in this research sensitivity is more important than specificity. The specificity of full logistic model is higher than that of simulated annealing model by 25.95%. But the sensitivity gap between those two models is larger (48.03%). So in this regard, simulated annealing model and genetic algorithm model are both better than full logistic model. Also, full logistic regression model has no feature selection function. It requires more predictors as inputs of the model, which means it causes a higher information searching cost in reality.

TABLE III. CONFUSION MATRICES OF THREE MODELS

	Prediction results of three models					
	logistic model		logistic model based on simulated annealing		logistic model based on genetic algorithm	
	Prediction		Prediction		Prediction	
	default	non-default	default	non-default	default	non-default
default	20	107	81	46	75	52
non-default	104	1268	460	912	467	905

TABLE IV. EVALUATION METRICS

models	Evaluation Metrics			
	Sensitivity	Specificity	AUC	Number Of variables
Full logistic	15.75%	92.42%	0.5408	300
Simulated annealing	63.78%	66.47%	0.7604	148
Genetic algorithm	59.06%	65.96%	0.7062	155

IV. CONCLUSION

This article gives an empirical analysis of individual credit valuation using data from an internet microfinance platform. We construct full logistic regression model、logistic model based on simulated annealing and logistic model based on genetic algorithm. Result shows the logistic model based on simulated annealing is slightly better than logistic model based on genetic algorithm. Both of them have better prediction ability and model interpretability comparing to traditional full variable logistic regression. Furthermore, from this empirical research, we can know when dealing with a massive index system, selecting all variables in model will cause poor prediction ability. After variable selection, models can have better performance. Comparing to stepwise logistic variable selection method, simulated annealing and genetic algorithm have faster computation speed.

Future research can focus on combing methods of dealing with sample imbalance with logistic model based on simulated annealing or genetic algorithm in Internet credit scoring.

REFERENCES

[1] Davoodabadi Z, Moeini A. Building Customers` Credit Scoring Models with Combination of Feature Selection and Decision Tree Algorithms[J]. Advances in Computer Science An International Journal, 2015, 4(2).

[2] Khashman A. A neural network model for credit risk evaluation.[J]. International Journal of Neural Systems, 2011, 19(4):285.

[3] Saeys Y, Inza I, Larrañaga P. WLD: review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19):2507-2517.

[4] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equation of State Calculations by Fast Computing Machines[J]. Journal of Biochemical & Biophysical Methods, 1952, 21(6):1087-1092.

[5] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by Simulated Annealing[J]. Science, 1983, 220(4598):671.

[6] Yuting Lu, YuxiaoLin and QiaoziPeng.A summary and exploration of improved simulated annealing algorithm and its parameters. [J]. College Mathematics in Chinese, 2015, 31(6):96-103.

[7] Goldberg D E. Genetic Algorithms and Walsh Functions: Part I{I, Deception and its Analysis[J]. Complex Systems, 1989, 3(2):153--171.

[8] Kumar M, Husian M, Upreti N, et al. Genetic algorithm: Review and application[J]. 2010.

[9] David J. Hand, Niall M. Adams. Defining attributes for scorecard construction in credit scoring[J]. Journal of Applied Statistics, 2000, 27(5):527-540

[10] Jiang Y. Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm[C]// CSIE 2009, 2009 WRI World Congress on Computer Science and Information Engineering, March 31 - April 2, 2009, Los Angeles, California, USA, 7 Volumes. DBLP, 2009:18-22.

[11] Dong G, Lai K K, Zhou L. Simulated Annealing Based Rule Extraction Algorithm for Credit Scoring Problem[C]// International Joint Conference on Computational Sciences and Optimization. IEEE Xplore, 2009:22-26.

[12] Cuticchia A J, Arnold J, Timberlake W E. The use of simulated annealing in chromosome reconstruction experiments based on binary scoring.[J]. Genetics, 1992, 132(2):591-601.

[13] Chi B W, Hsu C C. A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model.[J]. Expert Systems with Applications, 2012, 39(3):2650-2661.

[14] Hibbert D B. Genetic algorithms in chemistry[J]. Chemometrics & Intelligent Laboratory Systems, 1993, 19(3):277-293.

[15] Alippi C, Cucchiara R. Cluster partitioning in image analysis classification: a genetic algorithm approach[C]// Compeuro '92. 'computer Systems and Software Engineering',proceedings. IEEE Xplore, 1992:139-144.

[16] Pereira R. Genetic Algorithm Optimisation for Finance and Investments[J]. Mpra Paper, 2000.

[17] Ni H, Wang Y. Stock index tracking by Pareto efficient genetic algorithm[J]. Applied Soft Computing, 2013, 13(12):4519-4535