Accepted Manuscript

Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network

Yan Chun, Li Meixuan, Wei Liu, Man Qi

PII: S0304-3975(19)30417-7
DOI: https://doi.org/10.1016/j.tcs.2019.06.025
Reference: TCS 12076

To appear in: *Theoretical Computer Science*

Received date: 22 September 2018
Revised date: 22 May 2019
Accepted date: 6 June 2019

Please cite this article in press as: Y. Chun et al., Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network, *Theoret. Comput. Sci.* (2019), https://doi.org/10.1016/j.tcs.2019.06.025

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network

Yan Chun, Li Meixuan, Wei Liu*, Man Qi

***Abstract*—With the development of the insurance industry, insurance fraud is increasing rapidly. The existence of insurance fraud considerably hinders the development of the insurance industry. Fraud identification has become the most important part of insurance fraud research. In this paper, an improved adaptive genetic algorithm (NAGA) combined with a BP neural network (BP neural network) is proposed to optimize the initial weight of BP neural networks to overcome their shortcomings, such as ease of falling into local minima, slow convergence rates and sample dependence. Finally, the historical automobile insurance claim data of an insurance company are taken as a sample. The NAGA-BP neural network model was used for simulation and prediction. The empirical results show that the improved genetic algorithm is more advanced than the traditional genetic algorithm in terms of convergence speed and prediction accuracy.**

***Index Terms*— genetic algorithm; neural network; insurance fraud**

I. Introduction

Automobile insurance is the first kind of property insurance in our country [1]. With the increase in the number and size of auto insurance policies, the amount of auto insurance claims is increasing, and the number of auto insurance fraud cases is also increasing. The existence of insurance fraud affects the pricing strategy and the social economic benefit of insurance companies in the long term, and it even seriously threatens the development of the insurance industry in China.

In recent years, a variety of artificial intelligence techniques have been introduced into the research of insurance fraud identification. Viaene S. and Dedene G. et al. use a Bayesian Neural Network to verify fraud identification based on claim data of motor vehicle insurance in Massachusetts, USA [2]. Lovro Šubel et al. propose an iterative evaluation algorithm that considers both the attributes and the relationships between entities within an entity [3]. Taking motor vehicle insurance in China as an example, Ye M. H. proposed using a BP neural network to detect insurance fraud [4]. Liu J. L. et al. proposed two evolutionary data mining algorithms for insurance fraud prediction; one is an algorithm that combines a genetic algorithm with a K-means algorithm, and the other is an algorithm that combines a K-means algorithm with a momentum particle swarm optimization algorithm [5]. Tang J .and Mo Y. W. put forward a vehicle insurance anti-fraud detection system model using a support vector machine and a priori data mining technology [6]. Yan C. et al. put forward a vehicle insurance fraud identification model based on a stochastic forest and ant colony algorithm. This model can classify and predict the claim data of automobile insurance more effectively, and it has better accuracy and robustness [7]. The experiments of Wang Y. B. et al. show that the learning performance of deep neural networks is better than that of common machine learning models, such as the stochastic forest model and the support vector machine model. Therefore, a framework for detecting automobile insurance fraud based on the combination of a

Corresponding author: Wei Liu (liuwei_doctor@yeah.net).)

C. Yan is with the College of Mathematics and System Science, Shandong University of Science and Technology, Qingdao, 266590, China (e-mail:yanchunchun9896@163.com).

M. X. Li is with the Postgraduates of Probability Theory and Mathematical Statistics, Shandong University of Science and Technology, Qingdao, 266590, China, (e-mail: llmx9512@163.com).

W. Liu is with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, 266590, China, (e-mail: liuwei_doctor@yeah.net).

M. Qi is with the Department of Computing, Canterbury Christ Church University, Canterbury CT1 1QU, UK, (e-mail: man.qi@canterbury.ac.uk).

ACCEPTED MANUSCRIPT

deep neural network and the latent Dirichlet allocation algorithm is proposed [8]. Li Y. Q. et al. proposed a potential nearest neighbor stochastic forest method based on principal component analysis for the identification of automobile insurance fraud [9]. To curb life insurance fraud, Chudgar D. conducted a detailed survey of 500 clients using cross-form, frequency distribution and regression models [10]. Bhowmik R. et al. used naive Bayesian classification and a decision tree algorithm to solve the fraud identification problem of automobile insurance [11].

In the collection of claim data for insurance fraud, individual sample data may deviate from the actual situation due to a variety of human and random factors, which will cause distortion of the model. The BP neural network has good robustness and fault tolerance. The deviation of individual samples cannot affect the overall performance of the network and can be corrected by learning from the network itself. This is also an advantage of using BP neural network as a recognition model. But the traditional BP neural network has random initial weights, which leads to low learning efficiency, slow convergence speed, and ease of forming local minima without global optimization. Based on the above research, an improved adaptive genetic algorithm combined with a BP neural network algorithm is proposed to identify vehicle insurance fraud. The model combines a genetic algorithm with a BP neural network on the basis of considering that BP neural networks have strong predictive ability and that the genetic algorithm has good ability of searching for optimization.In this paper, the existing vehicle insurance fraud data index is classified and quantified, and then the principal component index of vehicle insurance fraud is selected by principal component analysis, which is used as the input of BP neural network. The improved adaptive genetic algorithm takes into account the degree of population fitness and adaptively adjusts the crossover probability and mutation probability of the genetic algorithm. To improve the convergence efficiency and optimization ability of the genetic algorithm, not only is the sorting and selection strategy added on the basis of the optimal preservation strategy but also the strategy of preserving parents is put forward. The improved adaptive genetic algorithm is used to optimize the initial weight of the BP neural network to achieve the prediction and analysis of vehicle insurance fraud.

Chapter 1 introduces the background significance of automobile insurance fraud and the common methods of identifying automobile insurance fraud in recent years. Chapter 2 introduces the models used in this paper and improves the genetic algorithm. Chapter 3 is the selection of indicators and the fraud test. Chapter 4 is the conclusion.

II. Arithmetic Statement

2.1 BP Neural Network

A BP neural network is a kind of multilayer feedforward neural network with a strong nonlinear mapping ability. The learning rule of the network is to use the steepest descent method to adjust the weights and thresholds of the network through back propagation to minimize the square error of the network. BP neural network learning process (is shown in Fig.1). First, the signal is passed from the input layer through the hidden layer to the output layer, and the error between the output value and the expected value is obtained. Then, the output error is transferred from the output layer to the input layer, and the weight of the neuron layer is adjusted by using the gradient descent function according to the error. Finally, the new weight is used to transmit the signal once more, and the output value is obtained again; the cycle is repeated until the output error reaches the result of satisfying the condition.

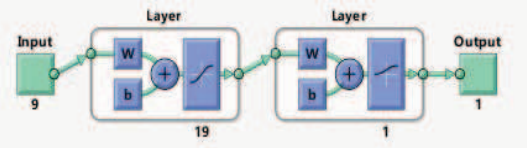


Fig.1 BP neural network structure

In the design of the BP neural network model for vehicle insurance fraud identification, the number of nodes, hidden layers and hidden nodes of the input and output layers should be determined according to the number of vehicle insurance fraud indexes. According to the Kolmogorov theorem, a three-layer BP neural network with an implicit layer can approximate any nonlinear function; thus, the number of hidden layers is

ACCEPTED MANUSCRIPT

set to one. Assuming that the number of nodes in the input layer of a BP neural network with three layers is m, according to the requirement of the problem, the number of hidden nodes p can be obtained according to the empirical formula $p=2*m+1$. W_{ij}^1 and W_{ij}^2 are the weight matrix of the input layer to the hidden layer and the hidden layer to the output layer, and b_j^1 and b_j^2 are deviation vectors of the hidden layer and the output layer, respectively. Suppose $x=(x_{ij})$,$(i=1,2,3,...,n,\ j=1,2,3,...,m)$ is a sample input matrix. The actual output sample of the overall sample is $Y=(y_{ij})$,$(i=1,2,3,...,n,\ j=1,2,\ 3,...,q)$, n is the number of training samples, and q is the number of output nodes. Each row represents a set of training output values, and each set of input training values corresponds to an output value. The excitation function of the hidden layer neurons of the neural network is the S-type tangent function tansig(), and the excitation function of the output layer neuron is the S-type logarithm function logsig(); thus, the net input of the hidden layer is $z_j = \sum_{i=1}^{m} x_i w_{ji} + b_j$ and the output of the hidden layer is $f_j=tansig(z_j)$.

BP neural network has the following advantages：

1、BP neural network algorithm transforms the I/O problem of a group of samples into a problem of finding their non-linear relationship, which can be approximated infinitely by iteration.

2、The thresholds of neuron nodes and the connection weights between neurons in BP neural network can be stored and recorded, so the BP neural network model has memory function.

3、The network architecture of distributed processing of BP neural network enables it to perform a large number of fast operations.

4、 The BP neural network model has strong adaptive learning ability for different training data sets.

2.2 Improved Genetic Algorithms

2.2.1 Flow Of Improved Genetic Algorithms

The traditional genetic algorithm (GA) is a computational model of the biological evolutionary process based on natural selection and the genetic mechanism of Darwinian biological evolution and is a method to search for an optimal solution by simulating the natural evolutionary process. However, for some complex optimization problems, the traditional genetic algorithm easily falls into some local extremum; thus, this paper proposes a new adaptive genetic algorithm called the New Adaptive Genetic Algorithm.

The process of improving the adaptive algorithm NAGA:

(1) Encode the initial group to set the parameters;

(2) Set the fitness function, calculate the fitness value of each individual, and retain the maximum fitness individual;

(3) Judge whether the convergence condition is satisfied or not. If the convergence condition is satisfied, output the result, otherwise proceed to step (4);

(4) If $\pi/12 \le \arcsin(f_{ave}/f_{\max}) < \pi/3$, perform the mutation operation first and then perform the crossover operation (this operation preserves the parent); otherwise, the crossover operation is performed first. Finally, the selection operation is performed;

(5) The result of the selection operation is used to judge whether the convergence condition is satisfied, and if the result is convergent, the output result is outputted. Otherwise, go back to step (2).

The solution flow chart of the NAGA is shown in Fig. 2

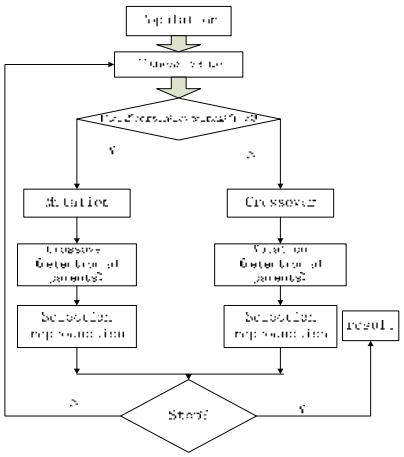


Fig. 2 The flow chart of the improved algorithm

ACCEPTED MANUSCRIPT

The reason why this paper uses $\pi / 12 \le \arcsin(f_{ave} / f_{max}) < \pi / 3$ to determine whether to cross first is because it is easy to have very small fitness in a population; this situation satisfies $f_{ave}/f_{max}<1/2$. If, according to the idea of Yang Congrui's IAGA algorithm[13], this situation is classified as population dispersion, the crossover operation will be carried out first. However, in this case, the population is concentrated in the smaller fitness, and the population difference is not large if the first crossover will slow down the evolution of the population. This outcome can result in slow convergence or difficulty converging. It is in view of this situation that the conditional formula is innovatively changed in this paper, which makes the improved algorithm more comprehensive. When $\arcsin(f_{ave} / f_{\max}) < \pi / 12$, it shows that the average fitness of the population is not close to the maximum fitness of the population. The smaller $\arcsin(f_{ave} / f_{\max})$ is, the easier to judge the more individuals with small fitness. Thus, it shows that the difference between populations is small, the population is not rich, and the crossover operation between individuals can not bring good quality individuals, and it also takes up the execution time of the algorithm. At this time, the adaptive increase of the mutation probability Pm value makes it possible to improve the performance of the algorithm. The algorithm can better jump out of local extremum and search for the optimal solution. When $\pi / 3 \le \arcsin(f_{ave} / f_{\max})$, it shows that the average fitness of the population is close to the maximum fitness value of the population. The larger the fitness value is, the easier it is to judge the more individuals with large fitness. Thus, the smaller the difference between the populations is. At this time, the adaptive increase of mutation probability is more conducive to finding the optimal value.

2.2.2 IMPROVED SELECTION OPERATOR

The selection operator in genetic algorithm performs the survival of the fittest according to individual fitness, which makes the individuals with higher fitness have a greater probability of being inherited to the next generation population, and the individuals with lower fitness have a smaller probability of being inherited to the next generation population. Sorting selection method is commonly used in genetic algorithm. The ranking selection method is based on individual ranking according to fitness to allocate the probability of each individual being selected.

In this paper, individuals are ranked according to fitness from large to small by ranking selection strategy, and the lowest one quarter of the fitness ranking is eliminated; the top one-fourth of the fitness ranking is directly retained as the father of the next generation. Keep the middle 1 / 2 individuals in operation [12] . In this way, bad individuals can be quickly eliminated and the direction of population evolution can be effectively grasped.

Then, calculate the selection probability of the L/2 individual left by the last step:

$$\begin{cases} p_k^N = Q^N(1 - q^N)^{k-1} \\ Q^N = \dfrac{q^N}{1 - (1 - q^N)^{L/2}} \end{cases} \quad (1)$$

p_k^N is the selection probability of the individual in the N generation, k is the ordinal number of an individual in a population, and L is the number of the population. For q, in the early stage of population evolution, there are large differences between individuals. Therefore, to ensure more excellent individuals, large fitness individuals should also have a larger selection probability. As the population evolves, the difference between individuals in the population decreases, and the selection probability of the best individual should be reduced. Therefore, a new q value is proposed, which varies according to the number of iterations.

$$q^N = q_{max} - (q_{max} - q_{min}) \times \frac{N - 1}{M - 1} \quad (2)$$

q_{max} and q_{min} are the choice probabilities of the best and worst individuals defined at the beginning, and M is the total number of iterations. According to the above probabilistic selection choice, L / 4 individuals are retained in new populations as part of the parent group,

ACCEPTED MANUSCRIPT

together with the L / 4 parts, which kept the first step, forming a parent population with L / 2 individuals. To keep the population number constant, the last operation before the selection probability retains the parent generation; in order to prevent the generation from missing better individuals in the intermediate process, the optimal preservation strategy is adopted in this paper [12], which compares the highest fitness of the new population capital with the highest fitness of the previous generation. If the fitness is higher than the highest fitness of the offspring, one individual in the offspring is eliminated at random. Add the highest fitness individuals of the previous generation to the new generation to produce a new population. This approach ensures that the superior individuals of the previous generation will not be destroyed by genetic manipulation, such as crossover mutation.

2.2.3 Adaptive Adjustment Of Crossover Probability And Variation Probability

Cross operation. Cross-operation plays a key role in evolution in genetic algorithm. Good gene recombination produces better individuals in a certain probability, so it can quickly converge to the neighborhood of the optimal solution by crossover and approximate the optimal solution one by one. Here is an example of gene crossover, as shown in Fig.3.

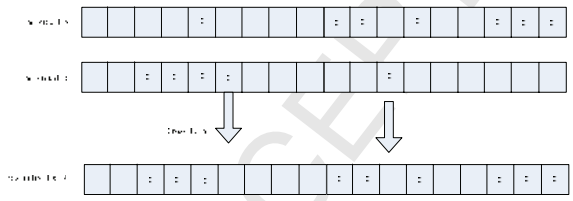


Fig. 3 Diagram of cross operation

Mutation operation. The purpose of mutation is to use mutation operator to break the situation when the population falls into the local optimal solution in the process of evolution and can not jump out only by crossover operator, so as to make the population enter the search for the next optimal value, so as to achieve the purpose of population evolution. Here's an example of gene mutation, as shown in Fig.4.

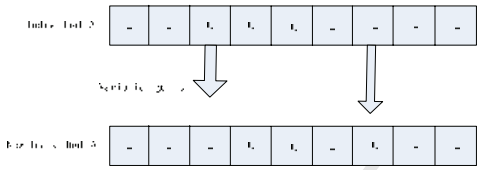


Fig. 4 Diagram of mutation operation

To play the important role in the genetic operation of crossover probability and mutation probability, this paper presents adaptive formulas (3) and (4) for the value of crossover probability and variation probability, respectively.

$$P_c=\begin{cases} k_1\left(1-\dfrac{\arcsin(\frac{f_{ave}}{f_{\max}})}{\pi/2}\right) & \arcsin(\frac{f_{ave}}{f_{\max}})\geq\pi/6 \\ k_4\dfrac{\arcsin(\frac{f_{ave}}{f_{\max}})}{\pi/2} & \arcsin(\frac{f_{ave}}{f_{\max}})<\pi/6 \end{cases} \quad (3)$$

$$P_m=\begin{cases} k_3\left(1-\dfrac{\arcsin(\frac{f_{ave}}{f_{\max}})}{\pi/2}\right) & \arcsin(\frac{f_{ave}}{f_{\max}})<\pi/6 \\ k_2\dfrac{\arcsin(\frac{f_{ave}}{f_{\max}})}{\pi/2} & \arcsin(\frac{f_{ave}}{f_{\max}})\geq\pi/6 \end{cases} \quad (4)$$

In formulas (3) and (4), k_1 is 1, k_2 is 0.05, k_3 is 0.02 and k_4 is 2. The reason why this paper uses $arcsin(f_{ave}/f_{max})$ is because as f_{ave} increases, $arcsin(f_{ave}/f_{max})$ increases faster. In this way, this method can better judge the degree of concentration dispersion between population fitness. The reason why we use π/6 is because $\sin(\pi/6)=1/2$; thus, when $\arcsin(f_{ave}/f_{\max})\geq\pi/6$, this indicates that the average fitness is closer to the maximum fitness. The more $\arcsin(f_{ave}/f_{\max})$ is larger than $\pi/6$, the more general the fitness of individual population increases. With the general improvement of population fitness, the fitness of individual population becomes more concentrated. At this time, the adaptive improvement of mutation probability is conducive to the population

ACCEPTED MANUSCRIPT

getting rid of extreme points.

Table 1 The result of calculation of function f(x)

Dimension f_3	Algorithm	Optimal value	Worst value	standard deviation	Mean convergence algebra
2	GA	9.313225748323190e-10	6.053596739151587e-06	1.376711701389571e-06	419.2
	IAGA	9.313225748323190e-10	2.328306448703445e-08	4.974508104619992e-09	524.23
	PSO	1.266404336525856e-05	8.944264329630682e-04	2.264486177013865e-04	430.73
	ACO	8.429240665235797e-07	6.693584802801848e-05	1.505079966253631e-05	472.07
	NAGA	9.313225748323190e-10	9.313225748323190e-10	0	51.80
5	GA	4.041003995582937e-04	0.007535471117016	0.001795045933530	512.4
	IAGA	1.193429343682162e-04	0.001575146336475	3.771218334428711e-04	567.4
	PSO	0.001721985838132	0.010893586440201	0.002620249266148	484.33
	ACO	8.351667765734251e-04	0.498748652273138	0.122299734029746	579.93
	NAGA	4.190951585769653e-09	5.685724318027496e-07	1.048191850564316e-07	491.93
10	GA	0.039440487520639	0.095536060653518	0.014830395778162	562.03
	IAGA	0.020122783263243	0.086047551614709	0.016063883807190	558.03
	PSO	0.020496745657030	0.054089103635652	0.007864213335203	556.2
	ACO	0.664931371530980	3.086295002314504	0.480979429994666	598.07
	NAGA	7.997453214625416e-05	6.778426469953303e-04	1.316253238873309e-04	543.5

In order to prove the effectiveness of NAGA algorithm, a multidimensional test functions are added in this paper. The IAGA algorithm compared in this paper is an improved algorithm in 2018, and we will discuss this issue more deeply in the future.

From the comparison of tables 1, we can see that the optimal value, worst value, standard deviation and average convergence algebra obtained by NAGA algorithm are smaller than those obtained by GA, IAGA, ACO and PSO algorithms. In function $f(x)$, the optimal value of function solved by NAGA algorithm is 9.313225748323190e-10 in 2-dimensional case, which is much smaller than ACO and PSO algorithm. Although the results are the same as those of GA and IAGA, it can be seen from tables 1 that the convergence algebra of GA and IAGA is obviously higher than that of NAGA. Moreover, the standard deviation of GA and IAGA is obviously larger than that of NAGA. It shows that NAGA is more stable than the other four algorithms.

2.3 NAGA-BP Model

In this paper, the improved genetic algorithm is used to optimize the weights of BP neural networks to solve the problems of slow convergence speed and low accuracy caused by the random initial weights of BP neural networks.

NAGA-BP model flow:

(1) Preprocessing the data, determining the topological structure of the BP network, coding the initial group and setting up each parameter;

(2) Taking the test sample error as the objective function and setting the fitness function of the genetic algorithm;

(3) Calculating the fitness of each individual;

(4) Whether the convergence condition is satisfied or not, if the convergence condition is satisfied, the BP neural network operation is entered; otherwise, go on to step (5);

(5) If $\pi/12 \le \arcsin(f_{ave}/f_{\max}) < \pi/3$, perform the mutation operation first, and then perform the crossover operation (this operation preserves the parent); otherwise, the crossover operation is performed first. Finally, the selection operation is performed;

(6) The result of the selection operation is used to judge whether the convergence condition is satisfied, and if the result is convergent, the output result is outputted, otherwise, return to the second step.

ACCEPTED MANUSCRIPT

The flow chart is shown in Fig. 5.

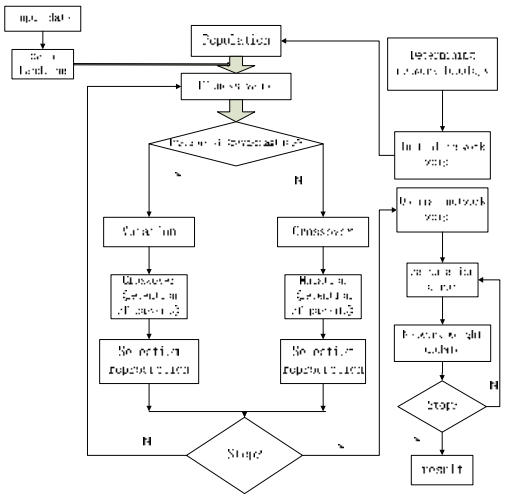


Fig. 5 Flow chart of the improved adaptivegenetic algorithm for optimizing BP neural networks

III. Empirical Analysis

Vehicle insurance fraud is more likely to occur than other types of insurance fraud because of the large number and chaos of vehicle insurance subject matter, the high frequency of vehicle insurance claims and the difficulty of prevention beforehand. To verify the validity of the proposed algorithm in vehicle insurance fraud identification, this paper selects the historical claims of a vehicle

insurance company as an example for the fraud identification analysis.

3.1 Selection of related indicators for vehicle Insurance Fraud Measurement

Before the prediction of vehicle insurance fraud, it is necessary to select the important indexes of the data, and the selected indexes are used as the input vector of the BP neural network. According to the known policyholder information, we select part of the information as the indicators of auto insurance fraud. On the basis of relevant studies, 15 factors affecting the existence of vehicle insurance fraud are preliminarily selected: the credit status of the claimant, the nature of the use of the insured vehicle, the service life of the vehicle under risk, the age of the claimant, the existence of an accident certificate, the risk record, the car inspection, the number of historical claims, the insured amount, the amount claimed, the actual amount paid, etc. Each variable type is described in Table 1.

From the description of the data variables of auto insurance claims in Table 1, we can see that there are non-numerical classification variables and Boolean variables; thus, we need to stratify and quantify these data according to the number of each index variable. The hierarchical results are shown in Table 2.

Table 1 Data set index description Data

variable	type	variable	type	variable	type
Channel source X1	classified variable	Vehicle inspection X7	classified variable	Number of reported replacement parts X13	discrete variable
Use nature X2	classified variable	Is there a report on the scene X8	Boolean variable	Target reported man-hour fee X14	discrete variable
Nature of vehicle X3	classified variable	Risk driver gender X9	Boolean variable	Number of hours reported X15	discrete variable
Automatic protection X4	Boolean variable	Survey type X10	classified variable		
Number of photographs with constant loss X6	discrete variable	Historical risk number X12	discrete variable		

ACCEPTED MANUSCRIPT

Table 2 Stratification of classified variables

variable	lamination
Channel source	The car company is set to 0, the tradition is 1, the agency channel is 2, the new channel is 3, and the comprehensive development is 4.
Use nature	1 for business and 0 for non-business
Nature of vehicle	Enterprise vehicle is set to 1, private vehicle is set to 2, agency vehicle is set to 0
guarantee	Is set to 1, no set to 0
Correct transfer of ownership	Is set to 1, no set to 0
vehicle inspection situation	Untested set to 0, tested to 1, exemption set to 2
Scene report	Is set to 1, no set to 0
Survey type	The first site is set to 1, the unsurveyed site is set to 0, and the replenishment site is set to 2
Target repair shop type	1 for first class shop, 2 for second class shop, 3 for third type shop, 0 for special service station
Whether or not to cheat	Is set to 1, no set to 0

There may be correlations among multiple indicators, and if all the indicators are selected, the recognition efficiency of the model will be affected. Therefore, we need to select the most impactful of vehicle insurance fraud indicators from the 15 indicators.

Principal component Analysis (PCA) is a multivariate statistical method, which can reduce the complexity of the data by reducing the dimension of the multi-dimensional feature matrix, and the reduced data can retain the main information of the original data.Table 3 shows the results of the principal component analysis of 15 fraud identification data.

Table 3 Principal component analysis results

	Initial eigenvalue			Extract the sum of squares		
component	statistics	accumulated %	accumulative total%	statistics	accumulated %	accumulative total %
1	2.496	16.637	16.637	2.496	16.637	16.637
2	1.913	12.752	29.389	1.913	12.752	29.389
3	1.594	10.628	40.018	1.594	10.628	40.018
4	1.376	9.172	49.189	1.376	9.172	49.189
5	1.192	7.944	57.133	1.192	7.944	57.133
6	1.052	7.011	64.144	1.052	7.011	64.144
7	.886	5.909	70.053	.886	5.909	70.053
8	.778	5.185	75.238	.778	5.185	75.238
9	.739	4.929	80.167	.739	4.929	80.167
10	.730	4.865	85.032			
11	.644	4.294	89.326			
12	.601	4.005	93.331			
13	.415	2.768	96.099			
14	.356	2.372	98.472			
15	.229	1.528	100.000			

Extraction method: principal component analysis.

ACCEPTED MANUSCRIPT

According to the results of principal component analysis in Table 3, the contribution rate of the first principal component is 16.63737% and the second principal component is 12.752%. The cumulative contribution rate of the first nine principal components is 80.167%, so the first nine principal components are extracted as input variables of the model.

3.2 FRAUD IDENTIFICATION RESULTS

The nine influence factors are selected as the input of the neural network; thus, the input layer of the BP neural network has nine nodes. According to the formula $p = 2 * m + 1$, we can determine that the number of hidden layer nodes is 19, whether fraud is predicted as the output, where the insurance fraud is 1 or 0, so that the number of nodes in the output layer is 1. In this paper, 79 vehicle insurance fraud samples were divided into two types, of which 70 were training samples and the remaining 9 were test samples. The BP neural network, the BP neural networks optimized by the IAGA algorithm and the NAGA algorithm presented in this paper were trained on the training samples. The test samples are inputted into the trained model to obtain the prediction results of fraud identification. Then, the results are compared with the original data to evaluate the degree of accuracy of each model for the prediction of vehicle insurance fraud.

In this paper, the genetic algorithm toolbox developed by Sheffield University is used to optimize the weight of BP neural networks by using MATLAB. Order $\|Y - T\_test\|$ as the objective function of the genetic algorithms, Y is the predicted output vector of each model and *T_test* is the original data of the test sample. The specific genetic parameters are as follows: the population size of NAGA and IAGA is 30, and the maximum iteration number is 50; the GA population size is 30, the maximum iteration number is 50, the cross probability is 0.7, and the variation probability is 0.1. In the process of optimization, the change of the objective function is shown in Figs. 6 / 7 /8.

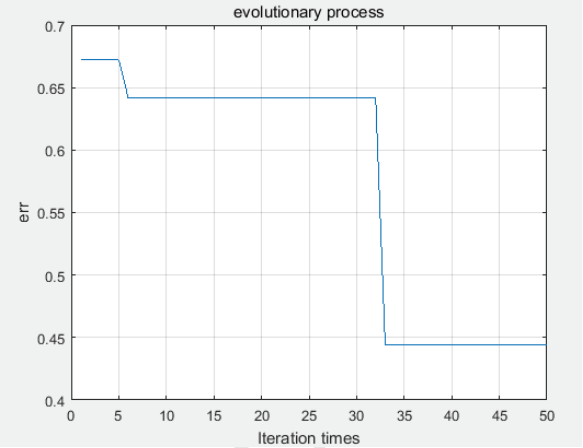


Fig. 6 Error variation diagram of the BP neural network optimized by NAGA

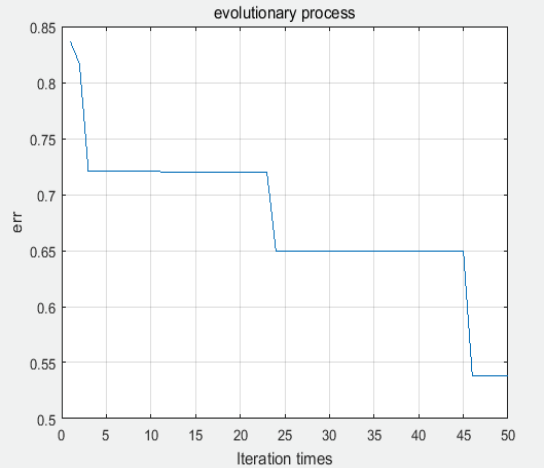


Fig. 7 Error variation diagram of the BP neural network optimized by IAGA

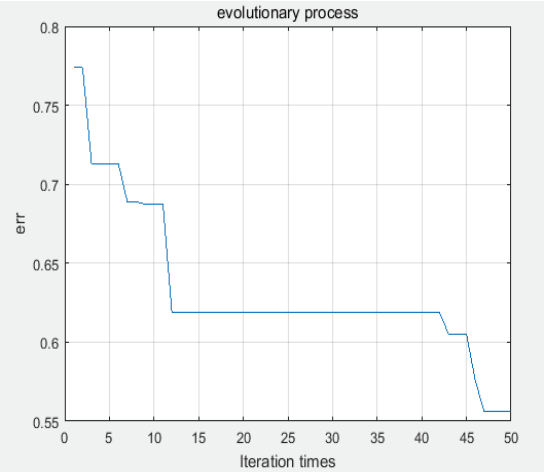


Fig. 8 GA optimized error diagram of the BP neural network

ACCEPTED MANUSCRIPT

It can be seen from Figs. 6 / 7 / 8 that the improved NAGA algorithm is obviously higher than the GA and IAGA algorithms in terms of the value of the optimal solution and the speed of convergence. Therefore, adjusting the crossover rate and the mutation rate of the genetic algorithm adaptively can improve the optimization ability of the genetic algorithm, and the combination of sorting selection and optimal preservation strategy is helpful in accelerating the convergence ability of the genetic algorithm. Therefore, the NAGA genetic algorithm has made great progress in convergence speed and accuracy. It can be seen from the diagram that the NAGA-BP network has reached the convergence state in only 31 steps, while the GA-BP network has not reached convergence at 50 steps. So the NAGA optimization improves the convergence speed of the network to a great extent.

The threshold value is 0.5. When the prediction result of each network is greater than 0.5, the prediction result is shown as 1 (fraud claim). And when the result is less than 0.5, the prediction result shows 0 (honest claim). According to the threshold set in Table 4, the accuracy of the prediction results of each network can be calculated.

As shown in Table 5, the comparison of BP neural network optimized by each genetic algorithm is shown in Fig. 9.

Further Evaluation of errors in experiments by using (MSE) of Predictive Variance.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\right)^2 \quad (5)$$

N, the number of test samples, is 9, and y_i and x_i are the real values and test values of the i samples.

Fig. 9 is a genetic algorithm to optimize the BP neural network prediction sample output value, without a genetic algorithm-optimized BP network output value, and to predict the original value distribution of the sample output. * For the actual situation of whether or not the insurance company determined that fraud occurred, O denotes the predicted value given by the NAGA-BP neural network; when O approximates 1, it represents the prediction of the claim as a fraudulent claim, and when *

Table 4 Algorithm prediction results comparison

sample	test data	BP	GA-BP	IAGA-BP	NAGA-BP
1	0	0.000033019941328	0.003245854173887	0.000000006772449	0.000215821058536
2	0	0.086436268378700	0.004642098875491	0.000533003939674	0.004926822968649
3	0	0.008582406460444	0.003351498620538	0.000000013200197	0.000000073818738
4	1	0.104825485761135	0.495356028603634	0.705259881429877	0.698046981448883
5	1	0.486384307239379	0.717697393740122	0.691628102314683	0.984204252728995
6	0	0.002804657741569	0.000600808927962	0.009366197020168	0.000044449373952
7	0	0.159693623091255	0.135148062884564	0.002599509936519	0.000000001002992
8	0	0.057864877038044	0.047922159707942	0.000000008741928	0.000249037640394
9	1	0.735488525681325	0.903782029391763	0.922606268949739	0.932803767143738

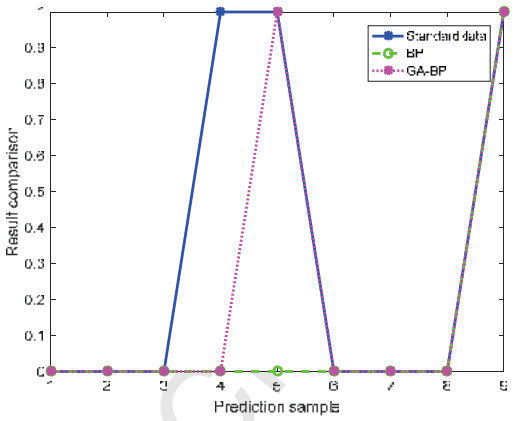
Table 5 Algorithm prediction accuracy and err comparison

	BP	GA-BP	IAGA-BP	NAGA-BP
Accuracy	77.78%	88.89%	100%	100%
MSE	0.1302	0.0405	0.0209	0.0107

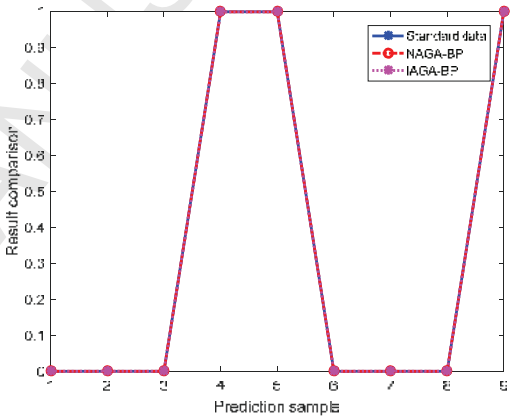
ACCEPTED MANUSCRIPT

approximates 0, it represents the prediction of the claim as an honest claim. At this time, according to the BP neural network trained by 9 influence factors, it can be seen from the prediction results that the fraud predictions and the honest claim predictions except the samples 1, 3, 6, 7, 8, and 9 are true value, and the judgment result of the 4 and 5 samples is wrong.Thus, the simple BP neural network model is not ideal for the identification of vehicle insurance fraud. The recognition of fraud is not ideal. The improved NAGA algorithm is used to train the improved NAGA algorithm to optimize the BP neural network. It can be seen from the prediction results that the nine samples predicted are close to the real values; thus, the vehicle insurance fraud predicted by the NAGA-BP model is more ideal.

Table 5 shows that the prediction accuracy of unoptimized BP neural network and GA optimized BP neural network is only 77.78% and 88.89%, while that of IAGA and NAGA optimized BP neural network is as high as 100%. It is shown that the improved genetic algorithm optimized by BP neural network is superior. In order to further compare the optimization effect of IAGA and NAGA, this paper further evaluates the effectiveness of these two models by using predictive variance (MSE) according to the results of Table 4. Table 5 shows that the error (MSE) of BP neural network optimized by NAGA is smaller than that of IAGA. So the prediction of fraud by NAGA optimized BP neural network based on improved genetic algorithm in this paper is closer to the original data. It is further explained that the improved NAGA genetic algorithm is more prominent in improving the shortcomings of BP neural networks which are prone to fall into local minima and slow convergence speed.



(a)



(b)

Fig. 9 Comparison diagram of the BP neural network optimized by each genetic algorithm

IV. CONCLUSION

In the insurance market of our country, automobile insurance is the first major insurance in terms of property insurance. According to the experience of international insurance, most dishonest claimants take the means of deliberately destroying insurance marks, forging traffic accidents, making false documents and so on, and the deceiver who escapes will certainly commit the crime again until the case is solved. Therefore, we urgently need to propose an effective method to identify vehicle insurance fraud to excavate potential fraudulent customers and to judge whether they are fraudulent or not, according to the customer's claim data, to take appropriate measures to prevent fraud in advance.

ACCEPTED MANUSCRIPT

In this paper, the fraud claim data of a certain insurance company are extracted by correlation tests, and the highly significant indexes are used as the variables of model fraud prediction to verify the ability of the NAGA-BP prediction model to identify fraud. The prediction model of a BP neural network optimized by NAGA is proposed in this paper. Considering the prediction ability of a neural network and the characteristics of searching and optimization of the genetic algorithm, the genetic algorithm is combined with a neural network. To overcome the shortcomings of neural networks, such as slow convergence speed and ease of falling into local minima, the improved adaptive genetic algorithm considers a variety of centralized dispersal degrees of population fitness and adaptively adjusts the crossover probability and mutation probability of the genetic algorithm. At the same time, the strategy of retaining primary parents is added, which not only ensures that the excellent genes in the intermediate process will not be destroyed by the subsequent genetic operation but also eliminates the individuals with low fitness in time to improve the convergence efficiency. In the final empirical analysis, the improved genetic algorithm is compared with the IAGA and GA algorithms in terms of convergence speed and accuracy, and the BP neural network is optimized to predict insurance fraud data. The results show that the prediction data of vehicle insurance fraud obtained by the improved NAGA-BP model are closer to the original data.

ACKNOWLEDGEMENT

This work was financially supported by the Project of National Natural Science Foundation of China (No. 61502280, 61472228).

All authors have no conflicts of interest.

REFERENCES

[1] Zhao Y. Z .Application of data Mining in vehicle Insurance Fraud and Identification . The times finance is 26: 246+249.

[2] Viaene S, Dedene G, Derrig R A. Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications, 2005, 29(3):653-666.

[3] Lovro Šubelj,Štefan Furlan,Marko Bajec. An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems With Applications,2010,38(1):1039-1052.

[4] Ye M. H. Insurance frauds identification research based on the BP neurological network——with China motor insurance claim as an example. Insurance Studies, 2011(03):79-86

[5] Liu J. L, Chen C L. Application of Evolutionary Data Mining Algorithms to Insurance Fraud Prediction. International Proceedings of Computer Science & Information Tech, 2012（7）:17-22.

[6] Tang J, Mo Y. W. Construction of vehicle insurance anti fraud system based on data mining technology . Shanghai insurance 2013 (11):39-42+63..

[7] Yan C, Li Y. Q, Sun H.T. A Research on Automobile Insurance Fraud Identification Based on Random Forest Model and Ant Colony Optimization Algorithm. Insurance Studies, 2017(06):114-127.

[8] Wang Y, Xu W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 2017（7）:87-95.

[9] Li Y. Q, Yan C ,Liu W , Li M. Z . A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing,2017：27-39.

[10] Chudgar, Dhara. "An Empirical Analysis on Drivers that Contribute to Fraudulent Activity with Regard to Life Insurance Fraud." Social Science Electronic Publishing (2017)：50-57.

[11] Bhowmik R.Detecting Auto Insurance Fraud By Data Mining Techniques. Journal of Emerging Trends in Computing ang Information Sciences,2011,2(4):371-377.

[12] Gao H, Xue L.Y. Back Propagation Neural Network Based on Improved Genetic Algorithm Fitting LED Spectral Model. Laser & Optoelectronics Progress, 2017, 54(7):294-302.

[13] Yang C, Qian Q, Wang F, et al. An improved adaptive genetic algorithm for function optimization IEEE International Conference on Information and Automation. IEEE, 2017:675-680.