Revista Argentina de Clínica Psicológica
2020, Vol. XXIX, N°4, 489-495
DOI: 10.24205/03276716.2020.855

Research on Active Discovery Model of Medical Insurance Fraud

Di Peng *

Abstract

Medical insurance is a major issue related to the national economy, people's livelihood and national development. The problem of medical insurance fraud seriously threatens the safety of medical insurance funds and hinders the effective implementation of medical insurance policies. Therefore, the active detection of medical insurance fraud is of great importance to the development, improvement and social stability of medical insurance. significance. This paper presents a method of identifying fraudulent behaviors based on BP neural network. For data processing, we chose Excel and Access to summarize and normalize the patient data in Table 1 and the cost schedule in Table 2 based on the patient ID, and eliminate invalid data including incomplete records and format errors. In the process, we discovered that all the consumption records were only for the purchase of medicines, and in this month's consumption records, only a very small number of patients had the behavior of transferring to the department, and some patients were paid at their own expense, and there was no suspected medical insurance fraud, and some patients shared by multiple people. The phenomenon of the medical insurance card directly determines that it is a medical insurance fraud. The consumption records of these patients provide sample support for us to train the BP neural network. For this question, we first used Excel and Access to filter out information useful for fraud identification from a large amount of data, including the patient's age, gender, department, total consumption of the current month, and consumption frequency of the current month, etc. You, and consider There are differences in the consumption situation of different departments, so we calculated the average consumption of each department, and made the relative difference between the consumption of each patient in the current month and the average consumption of the corresponding department. With these fraud factors and the consumption records of self-paid patients and patients who share medical insurance cards, we established a Logistic binary regression model to evaluate the impact of each fraud factor on the probability of fraud and eliminate fraud factors that are invalid for the possibility of fraud. , The fraud factors that have significant influence on the possibility of fraud are retained as the input vector to train the BP nerve, and the trained network is used to identify the fraud of medical insurance patients. In the end, we believe that the patient whose output is 1 is suspected of major medical insurance fraud.

Keywords: Medical insurance fraud, Logistic binary regression, BP neural network data

1. Introduction

Medical insurance is a social insurance system that provides material assistance to citizens or workers for the treatment costs and services after the inability to work due to illness and non-work injuries.

Medical insurance fraud refers to the behavior of violating medical insurance management regulations and policies, using fictitious facts, concealing the truth, and other methods to defraud medical insurance fund management institutions to obtain medical insurance funds or medical insurance benefits. This behavior has two basic characteristics: one is that it is subjectively manifested as direct intention, and the purpose is to illegally occupy the

Jilin University Law School, Jilin University, Changchun, 130012, China
*Correspondence Email: pengdi0898haikou@163.com

medical insurance fund or illegally obtain medical insurance benefits, and the second is that the means of implementation are mainly through fictional facts and concealing the truth, that is, deliberate fiction has never happened. Insurance accidents, or fabricating false reasons for the occurrence of insurance accidents or exaggerating the degree of loss, in order to achieve the purpose of obtaining medical insurance funds or medical insurance benefits by fraud. Since the implementation of the urban employee medical insurance and the new rural cooperative medical system in China, cases of defrauding medical insurance funds have continued to occur. In fact, medical insurance fraud in many countries has caused hundreds of millions of dollars in losses every year, posing a major threat to the safety of medical insurance funds. , Hindering the implementation of medical insurance policies in various countries, so medical insurance fraud has become a social problem that countries attach great importance to [1] Using mathematical modeling methods to analyze medical insurance fraud and establishing a corresponding mathematical model can provide a scientific and powerful basis for discovering medical insurance fraud.

Medical insurance fraud has two basic characteristics: one is that it is subjectively manifested as direct intention, and the purpose is to illegally occupy the medical insurance fund or illegally obtain medical insurance benefits; the second is that the means of implementation are mainly through fictional facts and concealment of the truth, that is, deliberate fiction has never happened In order to achieve the purpose of obtaining medical insurance funds or medical insurance benefits by defrauding the insurance accidents, or fabricating false reasons or exaggerating the degree of loss of the insurance accidents. The common methods used by defrauders to commit medical insurance fraud include fraudulent use of other people's medical insurance certificates and cards for medical treatment; medical personnel in other places forge or falsely issue medical bills to return for reimbursement; "hang up bed" for hospitalization; require the hospital to issue unnecessary diagnosis and treatment items. Or medicines, made or used by others, etc. The following situations are likely to be medical insurance fraud: the cost of a single prescription drug is extremely high, and one card repeatedly takes the drug within a certain period of time.

The BP neural network is a feedforward network trained by the error propagation algorithm. The learning process consists of two processes: the forward propagation of the signal and the backward propagation of the error. In the forward propagation, the pattern acts on the input layer and is processed by the hidden layer Then, the backward propagation stage of the incoming error, the output error is returned to the input layer layer by layer through the hidden layer in a certain form, and "allocated" to all units of each layer, so as to obtain the reference error or error signal of each layer unit , As the basis for modifying the weight of each unit. The process of continuously modifying the weight is the network learning process. This process continues until the error of the network output is gradually reduced to an acceptable level or the set learning times are reached [3][4] .

At present, BP neural network has been widely used in related economic research fields at home and abroad. Scholars in domestic securities, banking and other related fields have begun to use BP network for research. Ye Minghua applies this method to the research of motor vehicle insurance fraud, and Tried the fusion of statistical regression and neural network, which proved that the application of neural network to the recognition of insurance fraud is feasible, and the refined identification factors through regression analysis can make the neural network have better recognition effect.

In this problem, the amount of data is huge, and at the same time, self-pay patients and fraud patients (shared medical insurance cards) provide a large number of samples, and these samples can be used for BP neural network training, and can also be filtered by Logistic binary regression analysis. Fraud factors that have a significant impact combine quantitative and qualitative to make the results more accurate. Therefore, for this problem, the method of combining analytic hierarchy process and Logistic binary regression analysis is used to identify medical insurance fraud.

2. Model establishment and solution

2.1 Preprocess the data

2.1.1 Selection of samples and fraud factors

According to the data processing functions of Excel and Access used in the patient data and consumption records in appendix tables 1 and 2, combined with relevant data, first calculate the average consumption of each department, and then extract the average consumption of the patient's department and the patient's department. Amount, the total cost of the

patient in the current month, the frequency of taking medicines in the current month, age, gender, and 6 fraud factors, and these fraud factors are integrated into the patient ID. See the table in Annex 1.

Table 1 Summary table of fraud factors

Medicare patient ID	Patient department	Average consumption of each department	Frequency of taking medicine in the month	Total cost of the month	age	gender
363050	152	20.19601129	4	2220.05	29	1
627690	187	32.11214846	4	3160.25	52	1
168799	152	20.19601129	6	1018.29	46	2
178614	203	170.467056	9	7510.24	96	2
264972	187	32.11214846	5	1314.26	29	2
199056	187	32.11214846	6	1255.12	46	2
524738	152	20.19601129	3	721.51	48	2
406260	152	20.19601129	4	643.52	48	1
167305	187	32.11214846	6	913.27	47	1
331968	187	32.11214846	8	894.13	30	2
161213	173	151.3563019	5	3752.65	77	1
612657	10	95.28815897	1	2354.68	32	1

2.1.2 Self-paid patients and medical insurance cards share patients

Filter out the patients with the medical insurance card number 1 from the column of the patient information medical insurance card number in the title table 1, extract their patient ID and determine them as self-paid patients. Using the COUNTIF function in the column of the medical insurance card number to filter a multi-purpose patient with one card, it is found that there are 2 people sharing the medical insurance card and 3 people sharing the medical insurance card, and their ID is extracted and determined as a shared medical insurance card patient. According to the ID of the self-paid patients and the patients sharing the medical insurance card, the fraud factor summary table as shown in Table 4.2 is made for further analysis.

2.2 Refining the fraud factor

Use the binary discrete choice model to perform regression analysis on the selected six fraud factors, and obtain significant fraud factors from them. We use the IBM SPSS Statistics 19 software to implement Logistic binary regression analysis of sample data.

The parameters are set as:

Method: Enter

Step probability: enter=0.05, delete=0.1;

The maximum number of iterations: 50

C.I.(X) of Exp(B): 95%.

Table 2 Iteration history

Iteration		-2 log likelihood	coefficient
			constant
Step 0	1	7647.606	-1.935
	2	4584.590	-2.931
	3	3917.985	-3.645
	4	3832.549	-4.013
	5	3829.783	-4.095
	6	3829.779	-4.098
	7	3829.779	-4.098

a. Include constants in the model.

b. Initial -2 log-likelihood value: 3829.779

c. Because the change range of the parameter estimation is less than .001, the estimation ends at the number of iterations 7.

Table 3 is the iteration history. It is estimated that it will terminate after 7 iterations, and the initial -2 log likelihood value reaches 43.927.

Table 3 Classification table

Observed			Predicted		
			Fraud or not		Percentage correction
			0	1	
Step 0	Fraud or not	0	22585	0	100.0
		1	375	0	.0
	Percent of total				98.4

a. Include constants in the model.
b. The cut value is .500

It can be seen in Table 4 that in the input sample, 22585 cases are predicted to be 0, and 375 cases that should be 1 are also predicted to be 0, and the prediction accuracy rate is 98.4%.

Table 4 Significance test

		Score	direction-finding	signal
Step 0	variable	Relatively poor	1.624	1	.202
		Total cost of the month	7.607	1	.006
		Frequency of taking medicine in the month	1.401	1	.237
		age	10.676	1	.001
		Gender (1)	63.674	1	.000
		Patient department	216.120	1	.000
		Average consumption of each department	18.710	1	.000
	Presidential Measurement		300.128	7	.000

Table 5 is a global test of the model, which is a likelihood ratio test. Seven results are given: sig value<0.05 indicates statistical significance. It can be seen that the total cost of the patient in the month, age, gender, the patient's department and the average consumption of the patient's department have a significant impact on the regression, while other factors have no effect. According to this conclusion, a BP network model for medical insurance fraud identification can be established.

2.3 BP network model for medical insurance fraud identification

1) Set the initial weight W(0) to a small random non-zero value.

2) Given a set of input and output samples,$\{u_p, d_p\}_p$

Error index$E_p = \frac{1}{2}\sum_i (d_{ip} - y_{ip})^2$

Total error index$E_{all} = \sum_{p=1}^{P} E_p$

Repeat the following process until the convergence condition is met ($E_{all} \le \varepsilon$)

a) For any sample p, calculate

Forward process:

$$u_p, \cdots, {}^{l-1}O_p, {}^{l}x_p, \cdots, y_p$$

Reverse process:

$$\begin{cases} {}^{l}\delta_{ip} = -(d_{ip} - y_{ip}) \cdot f'({}^{l}x_{ip}) \\ {}^{l}\delta_{ip} = \left(\sum_m {}^{l+1}\delta_{mp} \cdot {}^{l+1}w_{mi}\right) \cdot f'({}^{l}x_{ip}), 1 < l < L \\ \frac{\partial E_p}{\partial {}^{l}w_{ij}} = {}^{l}\delta_{ip} \cdot O_{jp}, 1 < l \le L \end{cases}$$

b) Revised weight

There are two learning methods:

Pattern (Pattern) learning method:

$${}^{l}w_{ij}(t+1) = {}^{l}w_{ij}(t) - \eta \frac{\partial E_p}{\partial {}^{l}w_{ij}}, \eta > 0$$

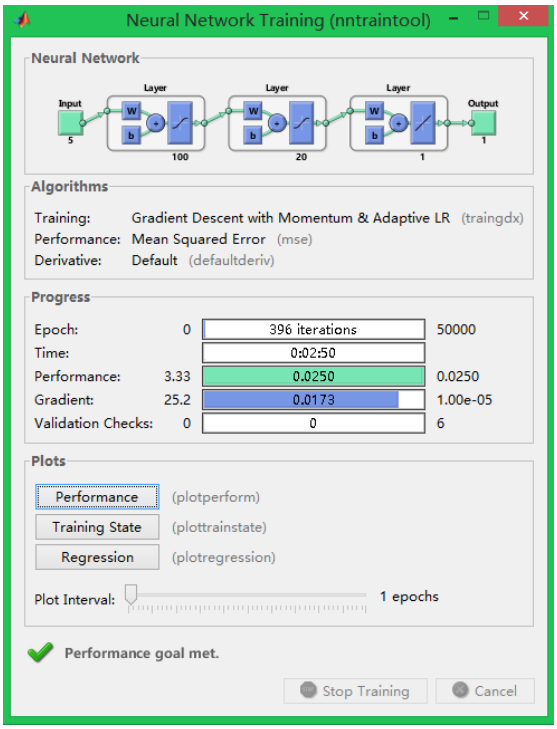
Training (Epoch) learning method:

$$^{l}w_{ij}(t+1)=^{l}w_{ij}(t)-\eta\frac{\partial E_{all}}{\partial^{l}w_{ij}},\eta>0 \quad [5]$$

The network input matrix is composed of 5 fraud identification factor vectors with model significance obtained by Logistic binary regression analysis. The network output vector matrix is a one-dimensional matrix composed of whether the patient is fraudulent (0 and 1), and 0 means that the patient has no fraud. ,1 represents fraud.

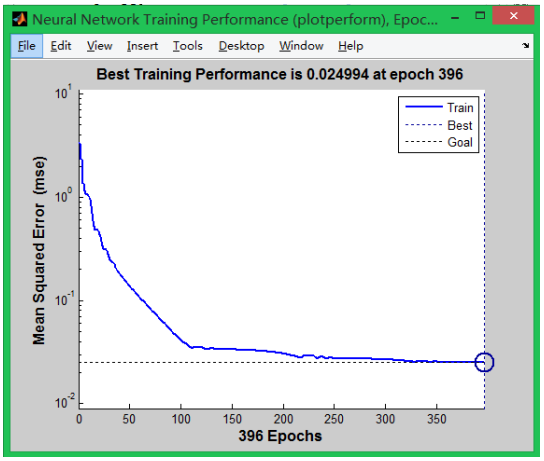
After repeated experiments, based on the principle of minimum error and shortest training time, a BP neural network model with two hidden layers is finally determined. Set the target error to 0.025, the maximum number of iterations is 50000, etc.[6].

Figure 1 BP neural network training diagram



It can be seen from Figure 1 that our BP neural network finally reached the target error of 0.025 after 2 minutes and 50 seconds after 396 iteration cycles. The training process is shown in the figure.

Figure 2 Simulation training process diagram



2.4 Fraud patient identification

According to the trained BP neural network, fraud identification is performed on the patients who

are not sure whether they are fraudulent, and the possible fraud patient IDs are found, and they are sorted according to their fraud possibilities. The higher the ranking, the greater the suspected fraud.

3. Conclusion

This paper adopts the method of combining Logistic regression and BP neural network, and uses Spss software to perform Logistic regression analysis on the samples to extract the identification factors with model significance; the obtained identification factors are used as the input vector of the BP neural network model for training, and the test samples are selected Predictive testing of the model's effectiveness proves the accuracy of the model and the feasibility of using this method to identify medical insurance fraud.

The method based on BP neural network of this model has many advantages: BP neural network has strong nonlinear mapping ability, and mathematical theory proves that a three-layer neural network can approximate any nonlinear continuous function with arbitrary precision. It avoids the process of finding the complex functional relationship between fraud factors and fraud, making the problem easier to solve. Secondly, the BP neural network has a certain fault tolerance. The BP neural network is damaged after its partial or partial neurons are damaged. It will not have a great impact on the overall training results, which means that the system can still work normally even when it is locally damaged.

At the same time, this model also has certain limitations. BP neural network is a local search optimization method. It has to solve a complex nonlinear problem. The weight of the network is gradually adjusted in the direction of local improvement. , This will cause the algorithm to fall into a local extremum, and the BP neural network is very sensitive to the initial network weight. Initializing the network with different weights will often converge to different local minima. This is also the result of multiple trainings. s reason. Secondly, there is no unified and complete theoretical guidance for the selection of BP neural network structure, and it can only be selected by experience. If the network structure is selected too large, the efficiency in training is not high, and over-fitting may occur, resulting in low network performance and reduced fault tolerance. If the selection is too small, the network may not converge. The network structure directly affects the network's approaching ability and promotion properties. In this article, we have adopted a number of experiments to determine the structure of the network, which is subjective.

This model can effectively detect fraud in medical insurance. Based on BP neural network, this model can be easily extended to other types of insurance industries, such as life insurance, motor vehicle insurance, etc. At the same time, although this model gives the ID of patients who are suspected of major medical insurance fraud, unfortunately we have not given the specific probability of each patient's fraud. This is where our model needs to be improved. Our results can provide a valuable reference material for the identification of medical insurance fraud.

References

[1] Lin Yuan. Analysis of the current situation of domestic and foreign medical insurance fraud research[J]. INSURANCE STUDIES, 2010, 12(12): 115-122

[2] Liu Kunkun, Empirical Research on Auto Insurance Fraud Recognition and Measurement Model——Based on Guangdong Province Auto Insurance Historical Claim Data, Journal of Jinan (Philosophy and Social Science Edition), 8:50-55, 2012.

[3] Zhu Daqi, Shi Hui. Principles and applications of artificial neural networks [M]. Science Press, 2006

[4] Tang Wanmei. Research on BP neural network structure optimization problem[J]. System Engineering Theory and Practice. 2005(10)

[5] Liu Caihong. Research on BP neural network learning algorithm [D]. Chongqing: Chongqing Normal University, 2008. 1-76

[6] Duan Chaoxia, Tian Xuemin. Fourier neural network structure selection method based on orthogonal least squares[J]. Petrochemical Industry Automation. 2012(06)

[7] Ye Feiyue. Fuzzy clustering method in the process of data mining[J]. Computer and Modernization. 2003(09)

[8] Edited by Wang Xuemin. Applied Multivariate Analysis [M]. Shanghai University of Finance and Economics Press, 1999

[9] Liao Ningfang, Gao Zhiyun. The best hidden layer structure of BP neural network for function approximation[J]. Journal of Beijing Institute of Technology. 1998(04)

[10] Edited by Cong Shuang. Neural Network Theory and Application Oriented to MATLAB Toolbox [M]. University of Science and Technology of China Press, 1998

[11] Deng Weini. PM10 pollution forecast based on BP neural network and its MATLAB implementation in Xi'an [D]. Xi'an University of Science and Technology 2008