Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):157-163

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A new credit risk assessment approach based on artificial neural network

Qian Zhang* and Tongna Liu

North China Electric Power University, Baoding, China

ABSTRACT

For most credit risk assessment models, decision attributes and history data are of great importance in terms of accuracy of prediction. Decision attributes can be classified into two types: numerical and categorical. As these two types have different characteristics, there will be interference if they are used simultaneously in the same model. By applying the case based reasoning (CBR) and artificial neural network (ANN), this study attempts to use numerical and categorical attributes separately in different phases application of the model. For example, if numerical attributes are used in CBR to select similar cases, categorical attributes will be used as inputs of an ANN based on the cases selected. Therefore, interference caused by the different types of attributes is avoided and the accuracy is improved. As only similar history data are selected and input in the ANN, accuracy is improved further. With the idea above, a triple ANN-CBR model is designed in this paper. This model synthesizes advantages of CBR and ANN. Practical examples show that the model established in this paper is feasible and effective. Compared with other models, it has a better precision performance.

Keywords: Artificial neural network; Case based reasoning; Credit assessment

---

INTRODUCTION

Credit risk assessment is very important to credit-granting institutions, in order to discriminate good customers from bad ones. According to some surveys, 80% of measurable risk is determined at the time of sanction of loans, i.e. once the credit is approved, the management can control only the remaining 20% of measurable risk. An improvement of even a fraction of one percent in classification accuracy can help lenders reduce losses of millions of dollars [1].

Due to the importance of credit risk assessment, many models have been developed since the seminal work of Altman [2, 3]. The first group of models comprises statistical analysis and optimization methods, such as linear discriminant analysis [4], logistic analysis [5], linear programming [6] and so on. The second group comprises artificial intelligent techniques, such as artificial neural networks [7, 8], evolutionary computation and genetic algorithm [9] and support vector machine [10] and so on. It is reported that this group is advantageous to statistical analysis and optimization models for credit risk evaluation in terms of empirical results. The third group is the combined or ensemble classifiers, which integrate two or more single classification methods such as neural discriminant model [11], neural-fuzzy model [12], and fuzzy SVM model [13].

No matter what method is applied, decision attributes and the historical data base are crucial to accuracy of the model. Firstly, all existing credit scoring systems incorporate numerical as well as categorical decision attributes. Although methods such as statistics analysis (t-test, regression…) have been developed to choose prediction variables in terms of its risk for distinguishing performance, numerical and categorical attributes are used simultaneously in all the above mentioned methods. As numerical and categorical attributes differs from each other in terms of characteristics and their relationship with the output, they may have bad impacts on each other if they are

used in the same model at the same time. Secondly, as the historical data are large, they are obviously different from evaluation objects. It is easy to understand that model accuracy will be improved if history cases are highly similar with the objects selected to train the model.

This study attempts to overcome the above two shortcomings by applying case based reasoning (CBR) and BP artificial neural network (ANN). Firstly, history cases with similar numerical decision attributes are selected by CBR to train and test an ANN assessment model which uses categorical attributes as inputs. Secondly, categorical attributes are used to choose similar history cases by another CBR, and another ANN with numerical attributes is used for training and testing. Thirdly, similar history cases are selected and a third ANN is acquired with all decision attributes. Finally, assessment results obtained from all the three ANN models for a given object are integrated on the basis of their accuracy. This triple CBR-ANN model is given the abbreviated name TCBR-ANN, and comparisons with other models, such as individual ANN and SVM, are made. The example results show that the TGBR-ANN model is workable and outperforms others in terms of accuracy.

2.Case Based Reasoning Introduction

On the basis of cognitive theories, the Case Based Reasoning (CBR) was presented by Roger Shank, professor of Yale University, in 1982. Just like the analogical reasoning process of human beings, the CBR compares the question case with the former cases, and similar cases are checked out of the case base. Then the remaining similar cases are used to solve the object case. As an incremental learning method, the new object case is retained in the case base for the future. Therefore, the ability of the system to make accurate predictions is improved constantly, and the knowledge and the experiences also increase incessantly [14]. This characteristic differentiates the CBR from other methods of artificial intelligence. A CBR system consists of four parts, the case base, indexing mechanism, matching algorithm and adaptation mechanism. It's working process includes "4Rs", which is retrieve, reuse, revise and retain. [15]

2.1Case representation and Case based building

In the CBR system, case c_i is represented as $c_i = \{X_i, y_i\}$, X_i is its description vector; it usually consists of relevant decision attributes. y_i is its results.

The case base of a CBR system is acquired by Interactive Activation and Competition (IAC) network. IAC is an incremental type neural network model suitable for CBR. IAC has good characteristics, such as generalization, robustness, suitability for information retrieval and so on. So IAC is very suitable for creating the CBR system. A case base is acquired by the following steps in IAC:

Step 1: If it is an old case in the case base, circulation will be stopped; otherwise it runs Step2 to Step 5.
Step 2: Based on the description vector of a new case, a decision is made whether the corresponding group has similar crunodes. If not, a new unit is created.
Step 3: From the input layer to the implied layer, the unit in the OC group is linked to the unit which expresses the corresponding description vector in the input layer; the weight between them is positive.
Step 4: From implied layer to the output layer, the new unit in the OC group is linked to the result unit in the output layer. The weight between them is positive.
Step 5: If a new crunodes is set up in the group, weights of the new crunodes with other crunodes in this group are negative.

As a result, a large size case base could be set up.

2.2 Matching Algorithm

The matching algorithm is used to calculate the similarity between the object case and those in the case base. It is evaluated on the basis of their representation vectors. The theory of the K-most closest Method (KCM) is adopted in the CBR systems proposed in this paper, whose function is to identify and define the similarity among the cases. When a new objective case is given, the KCM will search the case base, and K similar cases will be retrieved. The Most Closest means the maximum similarity between the object case and cases in the case base.

The comprehensive similarity is acquired according to the match function.

Let $C_a = \{(f_1, f_2, \cdots, f_m), \quad y_a\}$ be the objective case, and f_i ($i = 1 \sim m$) is one of its decision attributes (representation vector).

$C_b = \{(a_1, a_2, \cdots, a_m), \quad y_b\}$ is a case from the case base of CBR system, a_j ($j = 1 \sim m$) is its decision attribute(representation vector).

The similarity between C_a and C_b is calculated using

$$SIM(C_a, C_b) = \sqrt{\sum_{k=1}^{m} w_k \times \left[1 - \left(\frac{f_k - a_k}{R_k}\right)^2\right]} \quad (1)$$

Where w_k is the weight of the k^{th} decision attributes, (w_k =1 in the model of this paper) and R_k is the range of the k^{th} decision attributes.

3. The Main Idea of ANN

All the cases retrieved by the CBR system are used to train and test a classifier in the next step, and the artificial neural network (ANN) based on BP algorithm is adopted in this study. It is a network system containing a large number of simple processing elements, which are fully interconnected. In order to make the actual output close to any complex nonlinear mapping, its information processing procedure includes back propagation, forward propagation and weight adjustment. In the process of forward propagation, the sample is input into the input layer, is processed in the implied layer and is then sent to the output layer. The backward propagation process begins from the output layer, whose error function is used to adjust the weight coefficient. ANN of BP algorithms is developed in a 3-layer architecture comprising the input layer, the intermediate layer and the output layer. In this paper the input equates to the influencing factors and the output is the load type. The learning algorithm of the BP Network is as follows:

The first step: Set the initial parameter ω and θ (ω is the initial weight, θ is the critical value); randomly let both of them be a fairly small number.

The second step: Input the known sample to the network and calculate the output value y_j, using:

$$y_j = \left[1 + e^{-(\sum_i \omega_{ij} x_i - \theta_j)}\right]^{-1} \quad (2)$$

Where x_i is the input of junction $(i = 1, \cdots, m)$, ω_{ij} is the connection weight from i to j $(i = 1, \cdots, m, j = 1, \cdots, n)$, and let the initial weight be a fairly small number within [0,1]. θ_j is the critical value and y_j is the calculated value.

The third step: Adjust the weight coefficient ω on the basis of difference $(d_j - y_j)$ between the known output value and the calculated one.

The adjustment is calculated using:

$$\Delta\omega_{ij} = \eta \delta_j x_j \quad (3)$$

Where η is the ration coefficient (learning rate), x_j the input, d_j the actual output of the sample, and δ_j is the output deviation.

Regarding η, it is a small number within the range [0,1]. Under the presupposition that oscillation is not triggered and a fairly high precision is guaranteed, the value of η can be increased step by step until a satisfactory training speed is reached.

Regarding x_j, it is the network input to junctions in the intermediate layer, but to junctions in the output layer, it is the intermediate junctions' output $(j = 1, \cdots, n)$.

Regarding δ_j, it is a value related to output deviation. To junctions in the output layer, it is calculated using:

$$\delta_j = \eta_j (1 - y_i)(d_j - y_j) \tag{4}$$

To junctions in the intermediate layer whose output are hard to compare, its value can be acquired by counter calculation using:

$$\omega_{ij}(t) = \omega_{ij}(t-1) + \Delta\omega_{ij} \tag{5}$$

where t is the learning time.

This algorithm is an interaction process in which all values of ω are adjusted in each round. Such interaction is repeated until the output deviation is less than an acceptable value; then a good network is deemed to have been successfully trained. It is the essence of the BP algorithms to turn the first grade sample input question into a nonlinear optimized question. The gradient decreasing method used in BP algorithms is one of the most common methods used for optimization, while calculation of the weight value by interactive computation is equal to the learning memory.

Since the multilayer feed forward neural network based on BP algorithms has good capabilities in analogue classification, it is used here to simulate mutual relations between predicted variables and the credit risk. It has been proved that a single intermediate layer neural network can be close to any continuous function. Therefore the single intermediate layer neural network model is used in this paper.

4. The Triple ANN Model Based on CBR

Suppose $X = \{X_c, X_n\}$ is the decision attribute set in a credit assessment model, X_c is its subset of categorical attributes, and X_n is the numerical subset. y_i ($\in \{0,1\}$) is the observed result of the credit risk. If it is good, then $y_i = 1$; else, $y_i = 0$.

Then, the triple ANN model based on CBR (TCBR-ANN) can be described as follows.

4.1 Case Base Building

For a given data set, case c_i can be represented by three forms. The first form is c_{iN}, which is shown as:

$$c_{iN} = \{X_{in}, L_i\} \quad L_i = \{X_{ic}, y_i\} \tag{6}$$

The second form is c_{iC}, which is shown as

$$c_{iC} = \{X_{ic}, L_i\} \quad L_i = \{X_{in}, y_i\} \tag{7}$$

The third form is c_{iA}, which is:

$$c_{iA} = \{X_i, L_i\} \quad X_i = \{X_{ic}, X_{in}\} \quad L_i = \{y_i\} \tag{8}$$

According to the first form, c_{iN}, the first CBR system CBR_N is developed and a case base CB_N is created. The numerical attributes of X_{in} are used as its description vector, and $L_i = \{X_{ic}, y_i\}$ is the result, comprising quantitative attributes X_{ic} and the credit result y_i.

Similarly, CB_C and CBR_C, and CB_A and CBR_A are created on the basis of the second and the third representation forms.

Finally, the original data base is transformed into three case bases, and three GBR systems are developed.

4.2 Unitary ANN Model Based on CBR

For a given object case, its first similar case set is acquired by CBR_N . The similar case set is denoted as SC_N , which consists of K cases. Then, 80% of cases in SC_N are selected randomly as the training set TR_N , which is used to train the first ANN model denoted as ANN_C .

As CB_N is based on the first form c_{iN} which uses numerical attributes as descriptions, all selected cases have numerical attributes similar to the object set. Therefore, credits results of the case in SC_N and the object case vary with their categorical attributes to a great extent. It is reasonable to use categorical attributes X_c as the input vector of ANN_C , and credit risk y_i as its output. Then the first assessment result y_C of the object case is acquired by ANN_C .

The balance 20% of cases SC_N are used to test ANN_C accuracy, which is measured by percentage of correctly classified (PCC).

$$PCC = \frac{Gg + Bb}{Gg + Gb + Bg + Bb} \times 100\% \tag{9}$$

Similarly, SC_C and ANN_N are developed. This time the numerical attributes X_n are used as input vector of ANN_N because all cases in SC_C are of similar categorical attributes. Also, credit risk y_i is used as its output. Then the second assessment result (y_N) and its accuracy PCC_N are acquired.

Similarly, SC_A and ANN_C are developed on the basis of c_{iA} , which includes all decision attributes. Then, y_A and its accuracy PCC_C are acquired.

4.3 Global Result

According to the accuracy of the ANN model, the above assessment results are weighted, using

$$w_l = \frac{PCC_l}{PCC_N + PCC_C + PCC_A} \quad l = N, C, A \tag{10}$$

By integrating on the basis of their weights, the global assessment result is acquired as follows:

$$y = \begin{cases} 1 & if \quad w_N.y_N + w_C.y_C + w_A.y_A \geq 0.5 \\ 0 & if \quad w_N.y_N + w_C.y_C + w_A.y_A < 0.5 \end{cases} \tag{11}$$

5. Example Application And Analysis

The German credit card dataset at UCI Machine Learning Repository is used to test the model in this paper. It contains 1000 data, with 700 good cases (granted credit card) and 300 cases where credit cards were refused. In these instances, 20 decision attributes are listed, i.e. the decision attributes set $X = \{X_c, X_n\}$ consists of 20 decision attributes.

The categorical attributes subset X_c consists of 13 attributes : (XC1) Status of existing checking account; (XC2) Credit history; (XC3) Purpose; (XC4) savings account/bonds; (XC5) Present employment since; (XC6) Personal status and sex; (XC7) Other debtors/guarantors; (XC8) Property; (XC9) Other installment plans; (XC10) Housing; (XC11) Job; (XC12) Have telephone or not; and(XC13) Foreign worker.

The numerical subset X_n consists of 7 attributes: (XN1) Duration in months; (XN2) Credit account; (XN3) Installment rate as percentage of disposable income; (XN4) Present residence since; (XN5) Age in years; (XN6) Number of existing credits at this bank; and (XN7) Number of people being liable to provide maintenance for.

In this study, 900 data are selected randomly as the data set to develop CBR_N, CBR_C, and CBR_A. The balance 100 data are used to test the model presented in this paper.

For CBR_N, K=500, which means that the 500 most similar cases will be selected for a given assessment object. Therefore, the SC_N consists of 500 selected cases, and 400 cases will be selected randomly as the training set TR_C for ANN_C, and the balance 100 data will be used to test its accuracy PCC_C. ANN_C is a BPNN model, a three-layer feed-forward BP network. Categorical decision attributes in X_c are used as its inputs. Therefore, ANN_C is designed with 13 inputs, 8 TANSIG neurons in the hidden layer and one PURELIN neuron in the output layer. The training algorithm is Levenberg-Marquardt algorithm. Learning and momentum rates are set at 0.1. The accepted average squared error is 0.001 and the training epochs are 1000.

For CBR_C, K=500, SC_C = 500, TR_N = 400, and the balance 100 data are used to test accuracy PCC_N of ANN_N. Numerical decision attributes in X_n are used as inputs. Therefore, ANN_N is designed with 7 inputs, 8 TANSIG neurons in the hidden layer and one PURELIN neuron in the output layer. Other parameters are the same as ANN_C.

For CBR_A, K=500, SC_A = 500, TR_A =400, and the balance 100 data are used to test the accuracy PCC_A of ANN_A. All decision attributes in X are used as inputs. Therefore, ANN_A have 20 inputs, 7 TANSIG neurons in the hidden layer and one PURELIN neuron in the output layer. Other parameters are the same as ANN_A.

Assessment results of ANN_A, ANN_C, ANN_N and TCBR-ANN are compared with other models, including the traditional ANN and SVM. The results are shown in Table 1.

Table.1 Accuracy results and comparison

Methods	Sensitivity	Specificity	PCC
ANN	75.5	43.03	75.00
SVM	77.37	44.90	78.00
ANNC	82.61	46.02	83.00
ANNN	83.62	45.34	85.00
ANNA	79.02	44.91	79.00
TCBR-ANN	89.24	54.09	88.00

From Tab. Ⅰ, it is found that the assessment accuracy of ANN_N, and ANN_C, is better than ANN_A, as numerical and categorical attributes are used separately in ANN_N and ANN_C, which avoids the interference of varied data types. Also, the accuracies of ANN_A, ANN_C, and ANN_N are better than the traditional ANN and SVM because only similar history cases of varied types are selected to train the model. The best of the above models is the TCBR-ANN model which is a combination of ANN_A, ANN_C and ANN_N. The comparison shows that the proposed model is obviously superior to the traditional individual SVM and ANN models in respect of prediction precision.

CONCLUSION

A TCBR-ANN model for credit risk assessment is designed and tested in this paper.

The history data are preprocessed by CBR, only similar cases are selected for credit risk assessment; numerical and categorical attributes are used separately in different phase of the model, and their interference is overcome. This

model synthesizes the advantages of CBR and ANN as mentioned above. The practical examples show that the model established in this paper is feasible and effective. Comparison of forecasting results of different models shows that the proposed model is obviously superior to the traditional individual SVM and ANN models in respect of prediction precision.

REFERENCES

[1] Wang, Y.Q., Wang, S.Y., Lai, K.K., *IEEE Transactions on Fuzzy Systems* 13, 820‑831, **2005**.
[2] Thomas, L.C., *International Journal of Forecasting* 16, 149–172, **2012**.
[3] Thomas, L.C., Oliver, R.W., Hand, D.J., *Journal of the Operational Research Society* 56, 1006–1015, **2009**.
[4] Fisher, R.A., *Annals of Eugenics* 7, 179–188, **1936**
[5] Wiginton, J.C., *Journal of Financial Quantitative Analysis* 15, 757–770, **2013**.
[6] Glover, F., *Decision Science* 21, 771–785, **1990**.
[7] Lai, K.K., Yu, L., Wang, S.Y., Zhou, L.G., *Lecture Notes in Computer Science* 4132, 682–690, **2012**.
[8] Lai, K.K., Yu, L., Zhou, L.G., Wang, S.Y., *Lecture Notes in Computer Science* 4113, 403–408, **2009**.
[9] Chen, M.C., Huang, S.H., *Expert Systems with Applications* 24, 433–441, **2013**.
[10] Van Gestel, T., Baesens, B., Garcia, J., Van Dijcke, P., *Bank en Financiewezen* 2, 73–82, **2003**.
[11] Lee, T.S., Chiu, C.C., Lu, C.J., Chen, I.F., **2012**. *Expert Systems with Application* 23 (3), 245‑254.
[12] Malhotra, R., Malhotra, D.K., *European Journal of Operational Research* 136, 190‑211,**2012**.
[13] Wang, Y.Q., Wang, S.Y., Lai, K.K., *IEEE Transactions on Fuzzy Systems* 13, 820‑831, **2008**.
[14] Guo Yanhong, Deng Guishi *Computer Engineering and Applications*, **2004**, 40(21), 1-4
[15] Pal S., Simon C. Foundations of Soft Case-based Reasoning[M]. A Wiley-Interscience publication, **2009**.