2018 Citi Financial Innovation Application Competition

Machine Learning Report

Title ： Li Jin--A REITs platform for securitization of housing lease assets

Captain： Chu Tianshuo

Tutor： Gao Ming, Sui Cong

School： Dongbei University of Finance and Economics

Contents

1. Part One: Background Overview

1.1 Research Background

Lijin platform is committed to providing credit and financial support to small and medium-sized real estate enterprises. After packaging the guaranteed assets of small and medium-sized enterprises, it sets up a special plan for REITs, and earns commissions and investment returns from it. Among them, the credit rating assessment of medium and small real estate enterprises with repayment ability largely determines whether the platform can achieve expected earnings and whether it can run steadily. Therefore, a composite neural network model was designed based on the combination of supervised learning and unsupervised learning in machine learning, the model was continuously optimized for the accuracy of classification results,and finally, the credit rating was assessed to determine whether to extend loans to the rated enterprises.

1.2 Supervised Learning

Supervised learning refers to the training data input index to a certain extent, make up machine learning model under the index's standard for the classification of the training set to achieve high accuracy classification of training sets under the specification of this index.The Lijin platform inputs 13 financial data of small and medium-sized enterprises as feature variables into the learning model, takes bad or good credit enterprises as expected classification results of the model, extracts relevant financial information of enterprises with unknown credit rating from sina financial network and financial database as samples, and train the neural network model. Finally, the classification results are obtained to decide whether to allow enterprises to enter this platform.

2.Part Two: Model Selection

2.1 Model Adoption

By comparing the advantages and disadvantages of each rating method, the Lijin platform finally chooses SOM(self-organizing feature Map)-LVQ(Learning Vector Quantization) composite neural network model for customer credit rating.

Firstly, on the basis of referring to the literature of some famous scholars, we established the evaluation index system, used factor analysis to eliminate correlation, and selected representative factors to build the model.

Secondly, we built the model through MATLAB, input the filtered data into the SOM-LVQ composite neural network for training and classification, and finally divide the classification results into grades to complete the rating process.

2.2 Reasons For Selection

(1)The traditional factor analysis method, ratio analysis method and scoring method all make subjective analysis of the indicator system and give artificial weight to the index. Clearly, these methods are too subjective to be persuasive.

(2)Methods such as Logistic regression model and k-nearest neighbor discrimination of cluster analysis usually require relatively strict presupposition, such as normal distribution of sample data.

And a number of studies have shown that:

1) The research of enterprise financial status can be regarded as a classification problem of a series of independent variables.

2)There is usually a non-linear relationship between the influencing factors of a company's financial condition.

3) The predicted variables are usually highly correlated and cannot be analyzed independently.

Those factors further limit the choice of model.

(3)The category of enterprise samples collected by us is unknown in advance, which means that there is no way to conduct directed network learning, such as BP neural network model.

While SOM neural network is an undirected learning model, which can adjust the network parameters adaptively by automatically looking for the characteristics and essential laws of samples. At the same time, the weight obtained by SOM model exactly meets the prerequisite requirements of the supervised learning of LVQ model.

(4) SOM has a strong ability of self-organization and self-learning, which can automatically identify the most obvious features in vector space, with stable performance and effective classification. The advantage of LVQ is that the algorithm is simple, the training time is less, and it has good recognition effect.

(5) SOM and LVQ neural network models are both subordinate to the mode of competitive network, and they have relatively similar characteristics in principle and algorithm. Combining them organically can better realize predictive function.

3.Part Three: Supervised Learning Neural Network And Unsupervised Learning Neural Network

3.1 SOM Neural Network

The full name of SOM neural network is Self Organizing Maps, which can perform unsupervised learning clustering on data. The SOM neural network has only two layers of network structure, namely the input layer and the competition layer/hidden layer: the input layer corresponds to the input data, and the competition layer/hidden layer further optimizes the weight setting of the competition layer by automatically

adjusting the weight of the classification result. In the training method of "competitive learning", each input sample finds a node that matches it best in the hidden layer, called its active node, also called "winning neuron". Then the parameters of the active node are updated by the random gradient descent method. At the same time, the points adjacent to the active node also update the parameters appropriately based on their proximity to the active node.

3.1.1 Operation Steps

(1) Initialization: Each node randomly initializes its own parameters. The number of parameters for each node is the same as the dimension of Input.

(2) For each input data, find the node that best matches it. Assume that the input is D-dimensional, X={xi, i=1,...,D}, Then the discriminant function can be the Euclidean distance:

$$dj(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

(3) After finding the active node I(x), we will update the node that is adjacent to it. Let Sij denote the distance between nodes i and j, and assign an update weight to the nodes adjacent to I(x):

$$T_{j,I(x)} = \exp(-S_{j,I(x)}^2 / 2\sigma^2)$$

Simply, the neighboring nodes are discounted gradually according to the distance.

(4) Then it is to update the parameters of the node according to the gradient descent method:

$$\Delta w_{ji} = \eta(t).T_{j,I(x)}(t).(x_i - w_{ji})$$

Iterate until convergence.

(5) Record the maximum distance and the minimum distance of the last iteration.

3.2 LVQ Neural Network

The LVQ (Learning Vector Quantization) neural network is an input forward neural network for training competitive supervised learning methods. The algorithm is evolved from the Kohonen competition algorithm and can achieve effective nonlinear classification. The LVQ neural network consists of three layers of neurons, the input layer, the competition layer, and the linear output layer. As shown in the following figure:

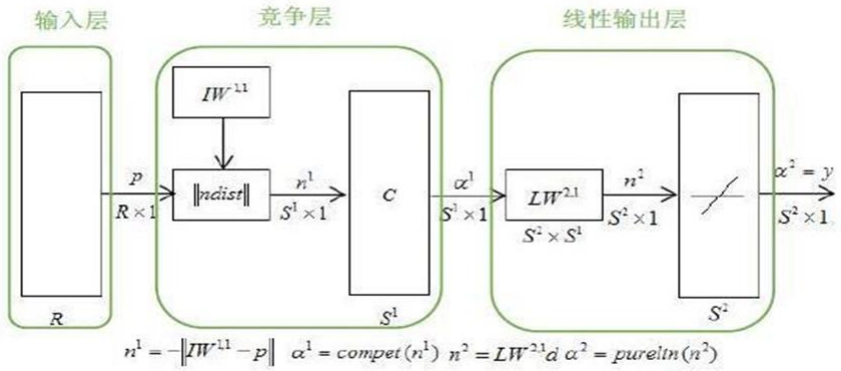

Figure 3-1

P is the input mode of R dimension;

$S^1$is the number of neurons in the competition layer;

$IW^{1,1}$is the weight coefficient matrix connecting the input layer and the competition layer;

$n^1$is the input of competition layer neurons;

$\alpha^1$is the output of competition layer neurons;

$LM^{2,1}$is the connection weight matrix between the competing layer and the linear output layer;

$n^2$is the input of linear output layer neurons;

$\alpha^2$is the output of linear output layer neurons.

4. Part Four: The Specific Implementation Process

4.1 Process Introduction

Firstly, the training vector is input into the SOM network for iterative calculation. By setting the step size, a stable network is obtained. Then the obtained training sample classification result is used as the input stream of the LVQ network, and the network parameters obtained by the SOM network training are used as the initial weight value of the LVQ network. Through the organic combination of SOM and LVQ, the whole process forms a composite network, which largely avoids the errors of training results due to the sensitivity of the neural network to the initial weight assignment and reduces the error of classification, and reduces the training time. . At the same time, the composite neural network uses a single neural network (such as BP neural network), which is more efficient to use and the prediction ability is greatly improved.

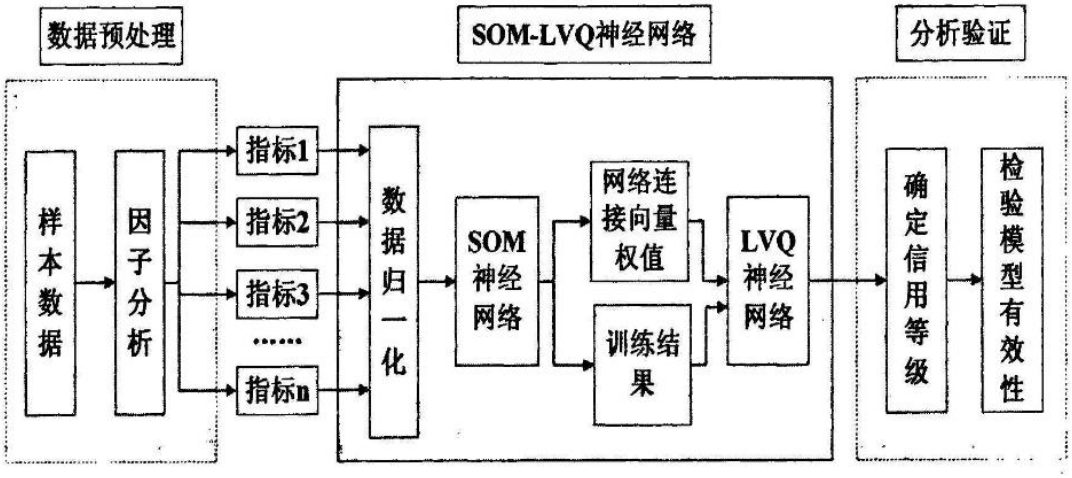

Figure 4-1

4.1.1 Data preprocessing

The collected enterprise data sample data is input into the SPSS software, and the indicators included in the solvency, operational capability, profitability, and growth capability are separately analyzed, and the representative indicators are selected to enter the SOM-LVQ neural network model through screening. Through screening, we selected 13 items such as ROE, current ratio, quick ratio, accounts receivable

turnover, inventory turnover ratio, and debt-to-equity ratio to enter the model for training.

Table 4-1

feature 1：Basic earnings per share	feature 2：Net assets per share
feature 3：Return on equity - weighted average	feature 4：Deducted earnings per share
feature 5：Current ratio	feature 6：Quick ratio
feature 7：Accounts receivable turnover	feature 8：Gearing ratio
feature 9：Net profit rate	feature 10：Rate of return on total assets
feature 11：Inventory turnover	feature 12：Fixed asset turnover
feature 13：Total asset turnover	

4.1.2 SOM-LVQ composite neural network

Since the financial indicator variable is the interval scale variable, the enterprise scale indicator is the ordinal type variable; in addition, the current ratio and the inventory turnover rate in the financial indicator are generally positive, and the asset profit rate can be positive or negative. Therefore, in order to make the different indicators have commonality, the data is first normalized to the [0,1] interval, and then the normalized data is input into the SOM model for training, which will be stable and effective after training. The result is entered into the LVQ model as a network connection vector weight. LVQ takes the input weight as its initial network connection vector weight, and through the self-learning of LVQ neural network and the training result of SOM as the supervision, the sample optimization training is carried out, and finally the classification result is obtained.

Note: Normalized formula: (take the maximum and minimum linear transformation method)

$$y = \frac{x - \mathrm{MinValue}}{\mathrm{MaxValue} - \mathrm{MinValue}}$$

Among them, MaxValue and MinValue are the maximum and minimum values in the sample, respectively.

(1) We use the MATLAB neural network toolbox to generate a SOM neural network using the newsom function to define the input matrix. Among them, the topology function of the neural network, the distance function, the learning rate of the classification stage, and the domain distance are all default values.

(2) Input the normalized sample into the SOM model and train the neural network using the train function and the sim function. To ensure real-time observation of training results, we set the number of training steps to 1000, 2000, 3000, 4000, 5000 and 6000. The code is as follows:

```
P = xlsread('C:\Users\dell\Desktop\huaqi\huizong');
n=13;
for i=1:n
    P(i,:)=(P(i,:)-min(P(i,:)))/(max(P(i,:))-min(P(i,:)));
end
P=P';    %%%取转置矩阵
PP=P(1:1:57,:)';
net=newsom(minmax(PP),[1 4]);
a=[1000 2000 3000 4000 5000 6000];
for i=1:6
    net.trainParam.epochs=a(i);
    net=train(net,PP);
    y=sim(net,PP);    %进行仿真
    yc=vec2ind(y);    %建立索引
end
```

```
18 —    pt=AA;
19 —    C=yc;
20 —    T=ind2vec(C);
21 —    n1=1;
22 —    n2=1;n3=1;n4=1;
23 —    n5=1;n6=1;n7=1;
24 —    n8=1;n9=1;n10=1;
25 —    n11=1;n12=1;n13=1;
26 —    a1=[];a2=[];a3=[];
27 —    a4=[];a5=[];a6=[];
28 —    a7=[];a8=[];a9=[];
29 —    a10=[];a11=[];
30 —    a12=[];a13=[];
31 —    for i=1:100
32 —        if C(i)==1
33 —            a1(n1)=i;
34 —            n1=n1+1;
35 —        end
36 —        if C(i)==2
37 —            a2(n2)=i;
38 —            n2=n2+1;
39 —        end
40 —        if C(i)==3
41 —            a3(n3)=i;
42 —            n3=n3+1;
```

(3) It is found through experiments that when the number of training steps of the neural network is 6000, the classification result is stable and the weight does not change.

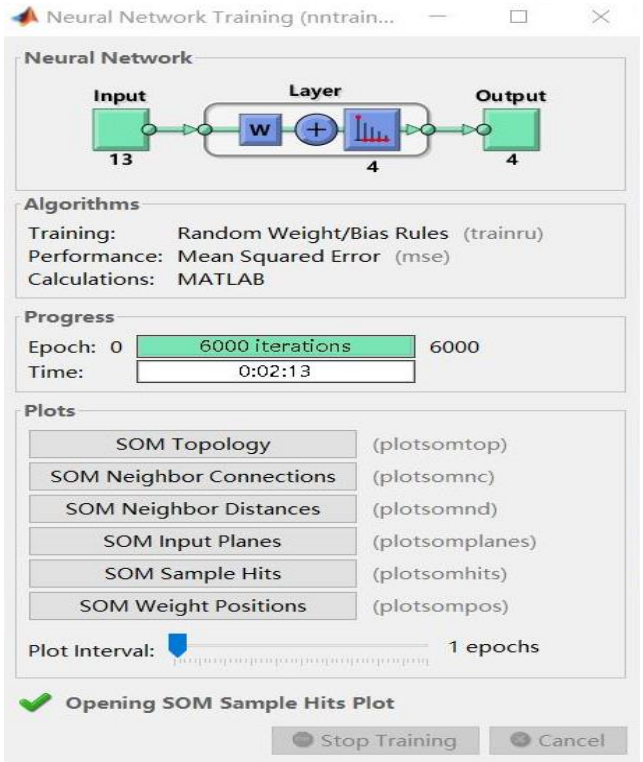

Figure 4-2

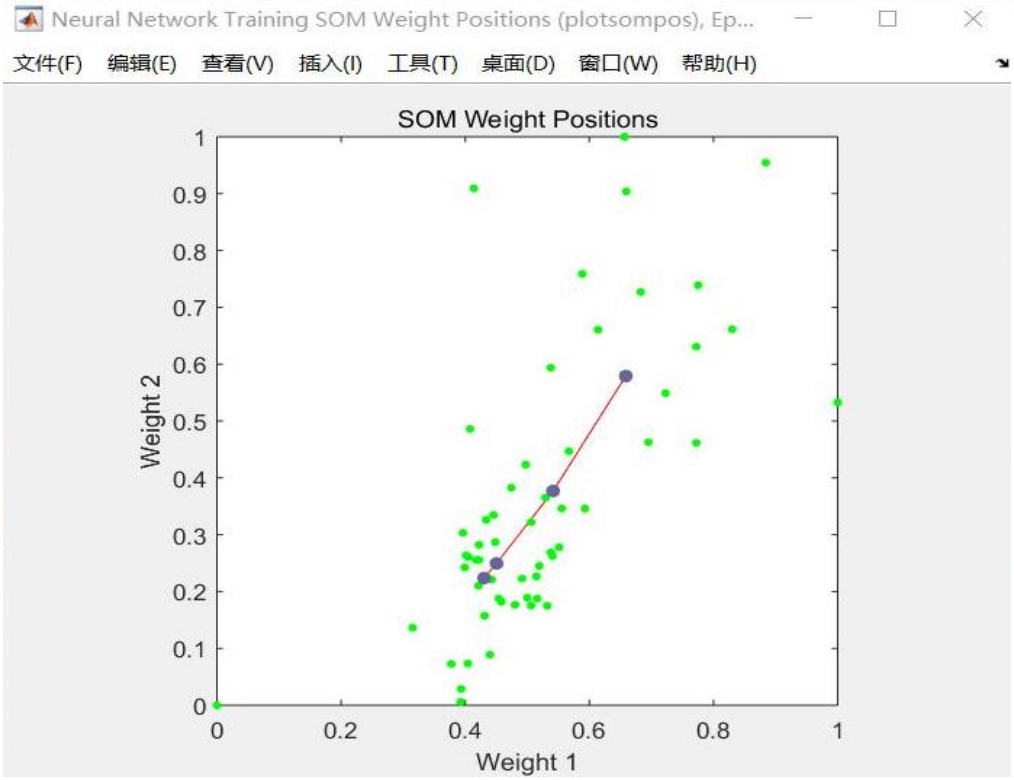

Figure 4-3

(4) Next, use the newlvq function to generate an LVQ neural network, and use the training result of SOM as the input vector of the LVQ network. Similarly, the parameters such as the topology function of the LVQ neural network are also taken as default values.

(5) Use the functions train and sim to simulate the neural network, set the number of training steps to 20, 40, 60, 80, 100. The code is as follows:

```
Pt=PP;
C=yc;
T=ind2vec(C);
n1=1;
n2=1;n3=1;n4=1;

a1=[];a2=[];a3=[];
a4=[];
for i=1:57
    if C(i)==1
        a1(n1)=i;
        n1=n1+1;
    end
    if C(i)==2
        a2(n2)=i;
        n2=n2+1;
    end
    if C(i)==3
        a3(n3)=i;
        n3=n3+1;
    end
    if C(i)==4
        a4(n4)=i;
        n4=n4+1;
    end
end
```

```
B=[(n1-1);(n2-1);(n3-1);(n4-1)]'/57;
lvqnet=newlvq(minmax(Pt),4,B);
lvqnet.IW{1,1}=net.IW{1,1};
a=[20 40 60 80 100];
for i=1:5
    lvqnet.trainParam.epochs=a(i);
    lvqnet=train(lvqnet,Pt,T);
    yy=sim(lvqnet,Pt);
    C=vec2ind(yy);
end

```

(6) Classification results shown in the following figure:

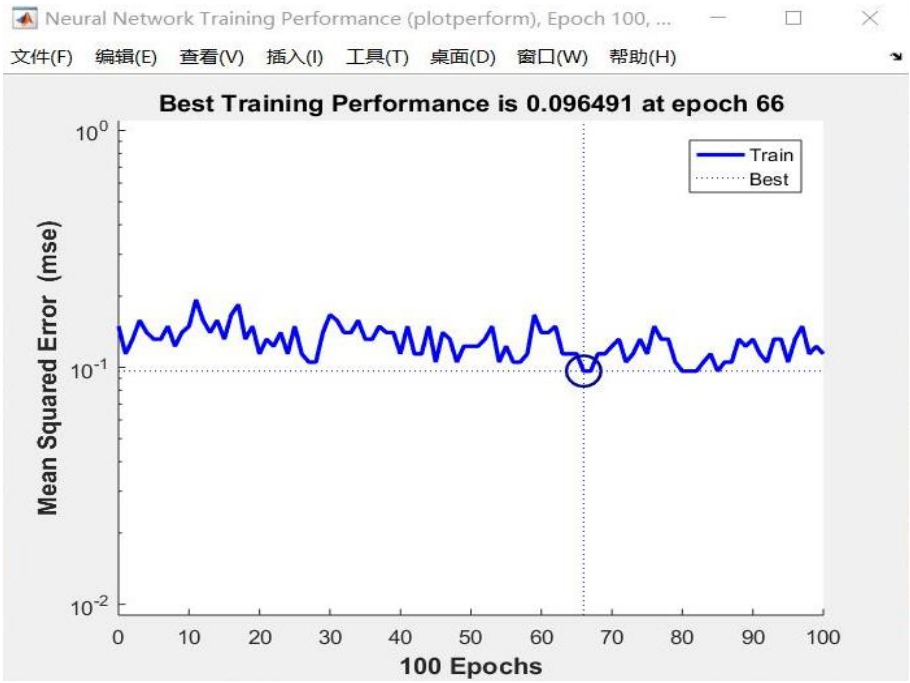

Figure 4-4

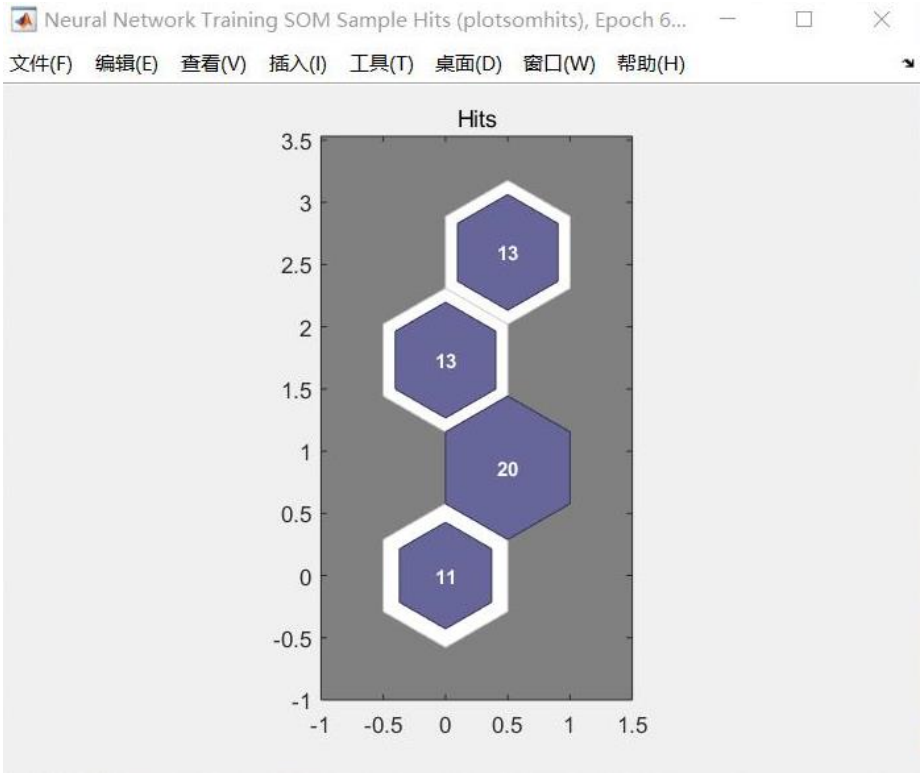

Figure 4-5

It can be seen that we have input a total of 57 enterprise sample data for training, and finally get 4 types of results. Note: data of 57 samples are shown in the following table:

Table 4-2

Serial number	enterprise name	Serial number	enterprise name
1	vanke	30	green-space
2	Gold-ground	31	poly
3	rongfeng-holding	32	new-hualian
4	rongan-property	33	deep-chase
5	huayuan-property	34	financial-street
6	Pan-sea-construction	35	shahe-industry
7	Yu-developer	36	wanze-shares
8	Beijing-urban-construction	37	suzhou-gaoxin
9	yangguangcheng	38	oct
10	Cof	39	shunfa-hengye
11	billion-shares	40	hna-foundation
12	taihe-group	41	wolong-real- estate
13	Wah-capital	42	zhongtian-finance
14	damingcheng	43	Shanghai-wanye
15	qixia-construction	44	sanxiang-impression
16	gorgeous-family	45	dagang-shares
17	Shanghai-lingang	46	new-huangpu
18	Zhejiang-guangsha	47	wolong-real-estate
19	Dadonghai	48	huatian-hotel
20	Jinjiang-shares	49	rongsheng-shares
21	contact-interaction	50	brand-new-good
22	Beijing-investment-development	51	sunshine-shares
23	*ST-tianye	52	ST-rock
24	dongfeng	53	jinghan-shares
25	Suning-global	54	rey
26	Jingneng	55	hotels

27	everbright-garbo	56	imperial-court-international
28	world-joint-line	57	Tibet-city-drop
29	Jin-Lingtong		

4.2 Analysis And Verification

On the basis of the classification result of the second step, the result is divided into different credit ratings; at the same time, the enterprise sample data of the known level is collected and verified as a verification set to ensure the validity of the classification result. The verification results are shown below:

Table 4-3

classification	Company
1	16，27，64，73，85，90，93，95，99
2	29，35，26，66，91
3	11，41，37，46，58，70
4	2，4，10，15，40，38

4.3 Model Evaluation

Through verification of the verification set, we found that the correct rate of classification results reached 80%, especially for enterprises with high credit rating and enterprises with low credit rating. There is a certain deviation in the rating of intermediate enterprises, but the deviation is small and has little influence on the credit evaluation. Thus, our SOM-LVQ composite neural network is effective.