# Personal Credit Rating Based on Partial Least Squares Regression Classification

**Dai Ting-ting[1, a], Shan chang-ji[2, b], Dong Yan -shou[1, c], Bian Yi-duo[3, d]**

[1]School of mathematics and statistics, Zhaotong University, Yunnan, 657000, PRC

[2]Institute of Physics and Information Engineering, Zhaotong University, Zhaotong,657000, Yunnan, PRC

[3]School of Foreign Languages, Zhaotong University, Zhaotong, 657000, Yunnan, PRC

[a]876310867@qq.com

**Abstract.** In view of the current commercial bank credit issues, this paper proposes a personal credit assessment method based on partial least squares classification. The main idea of this method is: first, a matrix is made by quantifying the customer and the corresponding credit indicators; secondly, a partial least-squares regression classification model is established; and finally, a test is conducted using the German credit data in the model and the results are obtained. Simulation results show that this method is simple, feasible and effective.

## 1. Introduction

Any credit institution will have outstanding credit risks in the process of conducting credit business, and it is the primary consideration for commercial banks and other credit institutions to guard against and control risks [1]. Therefore, the assessment of personal credit before loan is particularly important. To increase the efficiency of credit assessment, we must rely not only on subjective experience and the method of combining score systems, but also need a reasonable and scientific assessment method. For this reason, the study of personal credit evaluation is mainly about improvement of various statistical and artificial intelligence methods in China [2]. In the current mainstream research, individual credit evaluation is viewed as a clustering problem, a classification problem, and a regression problem. As different problems, the methods used are different. At this stage, the main assessment methods have experience. Judgment method, linear discriminant method, statistical method, operation research method, artificial intelligence method, and some non-parametric statistical methods. Based on these methods [3], this paper proposes a partial least-squares regression classification of personal credit assessment methods.

## 2. Theory and Method

### *2.1 Establish a personal credit assessment system [4]*

Personal credit assessment is actually a classification issue. According to different circumstances, the customer is divided into several groups (eg, defaulting customers and non-defaulting customers), and the impact of various factors is considered comprehensively to assess the size of the loan applicant's default rate. Credit agency risk size. After the credit assessment system is established, Assume that each customer corresponds to $n$ indicators, Each customer has n indicators, (That is, each training sample point has n dimensions) Thus, all customer information can be expressed in the following

matrix:

$$\begin{bmatrix} x_{11} & \cdots & x_{1n} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mn} & y_m \end{bmatrix} \tag{1}$$

Rows in the matrix represent customers, Columns indicate indicators, and the last column indicates evaluation values. That is, $x_{ij}$ represents the ith index of the ith customer, and $y_i$ represents the evaluation value of the ith customer. among them, $i$=1,2, ...$m$; $j$=1, 2, ..., $n$.

Based on the above, the credit assessment problem can be converted into the following mathematical problem: Finding hypersurface H(x)=0 in n-dimensional space, classifying m points in n-dimensional space into several categories.(Assume here that it is divided into 2 categories, which means that there will be no default and -1 will be breached).When there is an unknown class of points x(Unknown customer),You can use the following decision function to determine(The credit assessment):

$$f(x) = \text{sgn}(H(x)) \tag{2}$$

among them, $sgn()$ is an indicative function:

$$\text{sgn}(x) = \begin{cases} -1, x < 0 \\ +1, x \geq 0 \end{cases} \tag{3}$$

### *2.2 Partial Least Squares Regression Modeling [5]*

(1) Data standardization

The purpose of standardization is to coincide the center of the set of sample points with the origin of coordinates. In this article all standardizations are performed *Z_ score* standardization [6] deal with Its conversion function is:

$$Z(x) = \frac{x - \bar{x}}{S(x)} \tag{4}$$

among them, and S(x) Mean and variance of the sample, respectively. The calculation method is as follows: $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$S(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

For convenience, let P $y_1, y_2 ..., y_p$ $x_1, x_2 ..., x_m$ dependent variables and m arguments all standardized variables because of $n$ normalized observation data matrix of variable group and independent variable group Marked as:

$$F_0 = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \tag{7}$$

$$E_0 = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \tag{8}$$

Extract the first pair of components of the two-variable group and maximize the correlation.

Let the first component be extracted from two sets of variables, $t_1$ and $u_1$ is a set of independent variables Linear combination; $X=(x_1,x_2...x_m)'$ $t_1=w_{11}x_1+\cdots+w_{1m}x_m=w_1'X$ is a $u_1$ variable set $Y=(y_1,...,y_p)'$ The linear combination: $u_1=v_{11}y_1+\cdots+v_{1p}y_p=v_1'Y$ Need for regression analysis, Claim $t_1$ and $u_1$.

Extract each invariant group variation information as much as possible, and the correlation between $t_1$ and $u_1$ reaches the maximum. Normalized observation data array with two sets of variables $E_0$ and $F_0$ can calculate the score vector of the first component, Marked as $\bar{t}$ and $\bar{u}$

$$\bar{t_1}=E_0W_1=\begin{bmatrix}x_{11}&\cdots&x_{1m}\\ \vdots&\ddots&\vdots\\ x_{n1}&\cdots&x_{nm}\end{bmatrix}\begin{bmatrix}w_{11}\\ \vdots\\ w_{1m}\end{bmatrix}=\begin{bmatrix}t_{11}\\ \vdots\\ t_{n1}\end{bmatrix}$$

$$\bar{u_1}=F_0v_1=\begin{bmatrix}y_{11}&\cdots&y_{1p}\\ \vdots&\ddots&\vdots\\ y_{n1}&\cdots&y_{np}\end{bmatrix}\begin{bmatrix}v_{11}\\ \vdots\\ v_{1p}\end{bmatrix}=\begin{bmatrix}u_{11}\\ \vdots\\ u_{n1}\end{bmatrix}$$

The first pair of ingredients and Covariance Available score vector for the first pair of components and Inner product calculation. Therefore, the above two requirements can be turned into mathematical conditional extremum problems:

$$\max<\bar{t_1},\bar{u_1}>=<E_0w_1,F_0v_1>=w_1'E_0'F_0v_1 \tag{9}$$

$$w_1'w_1=\|w_1\|^2=1 \tag{10}$$

$$v_1'v_1=\|v_1\|^2=1 \tag{11}$$

Create a regression of $y_1,...,y_p$ to $t_1$ and a regression of to $x_1,...,x_m$ $t_1$.Let the regression model be

$$\begin{cases}E_0=\bar{t_1}\alpha_1'+E_1\\ F_0=\bar{t_1}\beta_1'+F_1\end{cases} \tag{12}$$

(2) Repeat the above steps using the residual matrix $E_1$ and $F_1$ instead of $E_0$ and $F_0$.

Note, The $\overline{E_0}=\bar{t_1}\alpha_1'$ $\overline{F_0}=\overline{u_1}\beta'$ residual matrix is $E_1=E_0-\overline{E_0}$ $F_1=F_0-\overline{F_0}$ and If the absolute value of the elements in the residual matrix $F_1$ is approximately 0, The artificially created regression with the first component has met the demand and can stop extracting components. Otherwise, use the residual matrix $E_1$ and $F_1$ instead of $E_0$ and $F_0$ to repeat the above steps. available:

$$\begin{cases}E_0=\bar{t_1}\alpha_1'+\bar{t_2}\alpha_2'+E_2\\ F_0=\bar{t_1}\beta_1'+\bar{t_2}\beta_2+F_2\end{cases} \tag{13}$$

(3) Calculate the regression equation

Let the rank of the $n\times m$ matrix $E_0$ be $r\le(n-1,m)$ There are $r$ components $t_1,...,t_r$ so that

$$\begin{cases} E_0 = \bar{t}_1\alpha_1' + \cdots + \bar{t}_r\alpha_r + E_r \\ F_0 = \bar{t}_1\beta_1' + \cdots + \bar{t}_r\beta_r + F_r \end{cases} \tag{14}$$

Substituting $t_k = w_{k1}x_1 + \cdots + k_{km}x_m (k = 1,2,...r)$ into $Y = t_1\beta_1 + \cdots + t_r\beta_r$ gives the partial least squares regression equation of *p* dependent variables:

$$y_j = y_{j1}x_1 + \cdots + y_{jm}x_m \quad j = 1,2,\cdots p \tag{15}$$

*2.3 Partial Least Squares Regression Personal Credit Rating Method [7]*

In this paper, the customer information matrix is used as the input data, the credit situation is used as the classification label, and the attribute feature composition method of the credit evaluation data of the least square regression classification is established. The basic process is to collect the credit assessment data and perform preprocessing to establish the least squares regression. The classification model selects the classification function, trains the data to obtain a least-squares regression (ie, obtains a least-squares regression classifier), performs an accuracy test, and performs credit evaluation on samples of unknown credit conditions. The specific steps are:

**Step 1:** Collect raw data and digitize it;

**Step 2:** Determine the final indicator system and sort out the numerical data set for evaluation;

**Step 3:** Establish a Least Squares Regression Classification Model and train it to obtain a least-squares regression equation;

**Step 4:** Select classification activation function to perform credit evaluation for customers.

## 3. Empirical analysis

*3.1 Determination of the activation function during the experiment*

Since this article discusses a simple two-category case, choose a relatively simple activation function during the experiment [8]:

$$F(x) = \begin{cases} f_1(x), x > T \\ f_2(x), x < T \end{cases} \tag{16}$$

Among them, $f_1(x)$and $f_2(x)$ are credit conditions (that is, classification labels), and is a $T = \frac{f_1 + f_2}{2}$ discrimination threshold, which can be set according to specific conditions. For example, it is worth noting that the threshold may be set to two. When the threshold is exceeded, it is determined as "good credit"; if it is lower than a certain threshold, it is determined as "credit difference".

*3.2 Experimental Results and Analysis*

According to the 23 attributes of the German credit data, the cross-validation method is used to determine the components to be extracted, and the regression equation is finally obtained as:

$$f(x) = 0.0962x_1 + 0.0030x_2 - 0.0432x_3 + 0.0011x_4 - 0.0368x_5 - 0.0131x_6 - 0.0372x_7$$
$$-0.0014x_8 + 0.0243x_9 - .0014x_{10} - 0.0471x_{11} + 0.0326x_{12} - 0.0063x_{13} - 0.0243x_{14}$$
$$-0.0135x_{15} + 0.1479x_{16} - 0.1728x_{17} + 0.0565x_{18} + 0.0901x_{19} + 0.0476x_{20} - 0.0407x_{21}$$
$$-0.0855x_{22} - 0.0150x_{23} + 1.9682$$

Table 1 Evaluation results of German credit data under different discrimination thresholds

| Discrimination | Percentage of judgement result (%) | |
|---|---|---|
| Threshold | Training set | Test set |

| (T) | Accuracy | The first kind of mistake | The second type of error | Accuracy | The first kind of mistake | The second type of error |
|---|---|---|---|---|---|---|
| 1.50 | 76.51 | 14.98 | 7.5 | 78.84 | 16.81 | 4.6 |
| 1.42 | 78.16 | 11.53 | 10.18 | 77.76 | 14.42 | 7.82 |
| 1.41 | 75.85 | 9.4 | 14.82 | 75.34 | 11.42 | 13.22 |

From table 1 it can be quickly seen that the first feature attribute has a very important role in the evaluation of the credit assessment results, while the 7th and 12th feature attributes have a weak interpretation of credit results.

Can be seen from Table 1,When taking the threshold $T=\frac{1+2}{2}=1.5$(1 and 2 are credit conditions, 1 indicates :good credit 2 indicates :bad credit)The first error is greater than the second error, which is as high as about 15%, which may cause a large number of 'bad credit' customers to be evaluated as 'good credit' customers during the credit evaluation process, which will increase the credit risk of credit institutions such as commercial banks big; When the threshold value $T$ is biased to the label of the 'good credit' sample within a certain range, by the time $T=1.41$ is reached, the first type of error is significantly smaller than the second type of error, and the control of the error is very convenient and effective.

## 4. Conclusions and Prospects

Based on the method of partial least-squares regression, this paper proposes a partial least-squares regression classification method and uses it in credit evaluation. It also uses the German credit data to test and proves the rationality of this method. However, this article still has some deficiencies: First, this article only uses one German credit data for experimentation, but does not use many other credit data, and lacks stronger support for the rationality of this method; second, there is no detailed study on Multi-grade or two-dimensional and above credit rating indicators are used for testing. These deficiencies are all future research needs to overcome.

## References

[1] M. Fiedler. Algebraic connectivity of graphs[J]. Czech. Math,1973,298–305.
[2] A. L. Barabas. Thoughts on Constructing a Large Quantity of Personal Credit Evaluation System in China[J]Black River Journal.2013,21(15):42-43.
[3] O ja E. A simp lified neuron model as a princip le component analyzer[J]. J Math Bio l, 1982, (5): 267- 273.
[4] CORTES, VNPAIK V.Support-vector networks[J].Machine Learning,1995,20(3):273.
[5] Durand D. Risk elements in consumer installment financing[M]. New York: The Nationl Bureau of Economic Reseach,1998:145.
[6] Altman E I. Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy[J]. The Joumal of Finance,2001,23(4):589-609.
[7] VAPNIK V. The Nature of Statistical Learning Theory[M]. New York:Springer,2000
[8] YUNG S W. MicroRNA-195 regulates vascular smooth muscle cell phenotype and prevents neointimal formation[J]. Cardiovascular Research,2012,95(4):517.