A Logistic-BP Classifier for Assessing Personal Credit Risk

Li-Qian WU

School of Mathematics, Sichuan University, Chengdu
LiqianWu1992@163.com

Keywords: Logistic Regression, BP Neural Network, Logistic-BP, Personal Credit Assessment.

Abstract. In view of the present personal credit risk problems in peer-to-peer industry, this paper proposes the Logistic-BP (Back Propagation) combined optimal model. The method combines the traditional statistical method with the artificial intelligence method by building a unified error function. The simulation is carried out with a German commercial banks’ credit data, and the result shows that the Logistic-BP model has higher accuracy and robustness than single model. The average classification accuracy for the test sample is 77.3%, 2-7% higher than the single model and has an obvious promotion effect. In this paper, we compare Logistic-BP with other methods such as LDA, Naive Bayes, RBF-LS-SVM and C4.5, the result shows that Logistic-BP is superior to other algorithms.

1. Introduction

In recent years, along with the rapid development of the internet and the appearance of the big data, the traditional industries are facing the technical revolution. Internet finance appears at this situation. And it not only services the long tail of customers, but emphasizes the personality needs of customers, and has the congenital superiority in efficient. The internet enters in the field of personal credit at the first place. According to statistics, by May 2016, nearly 4000 P2P platforms were put into operation in China. P2P financial is growing far too popular nowadays. In order to control the credit risk to prevent banks and other financial institutions from losing money and service the customers more efficiently as well. It is necessary to establish a scientific system of personal credit evaluation method.

Credit assessment is to discriminate consumers who are assumed to belong to one of two classes, namely good and bad credit risks. Traditional method of credit risk assessment has focused on using the techniques such as the use of human judgement, statistical models, decision tree analysis, mathematical programming and neural networks. The earliest published study dealing with the credit scoring was that by Durand [1] in 1941, under the sponsorship of the National Bureau of Economic Research. Durand employed discriminant analysis for the personal loan accounts. Results showed good prediction of credit repayment in 20 commercial banks and in 9 industrial banking companies. Myers and Forgy [2] introduced the development of numerical credit evaluation systems in 1963, and explained the reasons why numerical rating systems were not in widespread use in that era. Srinivasan and Kim [6] made a comparative analysis of classification procedures of credit granting in 1987. Examined four statistical models: multiple discriminant analysis (MDA), goal programming (GP), logistic regression (logit), and recursive partitioning algorithm (RPA), and a judgemental model based on the Analytic Hierarchy Process (AHP). Henley and Hand [7] used the k-nearest-neighbour (k-NN) method, a standard technique in pattern recognition and nonparametric statistics, to the credit scoring problem. Atiya [8] applied the neural network method for a bankruptcy prediction in 2001. Tsai and Wu [9] used neural network ensembles for bankruptcy prediction and credit scoring in 2008. They investigated the performance of a single classifier as the baseline classifier to compare with multiple classifiers and diversified multiple classifiers by using neural networks based on three datasets.

This paper is organized into four parts. Section 2 details the basic concept for Logistic model and BP neural network, also proposes the combined optimal model: Logistic-BP model. In addition, Section 3 describes the dataset and dose the simulation test. The performance of the combination
model is compared with a range of other discrimination techniques. Finally, Section 4 concludes and summarizes the study.

2. Logistic-BP combined method

2.1. Logistic model

Logistic regression is a kind of generalized linear regression analysis model, which is mainly used to solving the problem of classification. It helps classify samples by predicting probability of the sample type. There are lots of applications in the field of disease diagnosis and economic forecasting. The logistic regression has satisfying good robustness and accuracy, and has no specific requirements to the distribution of the variables. So that logistic model is one of the main models of banks and other financial institutions. By comparison, most other statistical methods require that the data should obey the normal distribution assumption, so from the point of view of statistical methods, Logistic is a kind of ideal model of classification.

Firstly, given a set of customer samples \( T = \{(x_i, y_i)\}_{i=1}^n \), where \( x_i \in \mathbb{R}^p \) is the index vector of customer \( i \), \( y_i \in \{0,1\} \) is a binary variable ( \( y_i = 1 \) represents "good" customer, and \( y_i = 0 \) represents "bad" customer). Logistic model can be used to judge a client default probability, the formula can be expressed as:

\[
P(y_i = 1 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}.
\]

(1)

where \( x_i \) is a \( p \) dimensional vector, \( \beta_0, \beta_1, \cdots, \beta_p \) are the estimator of model parameters. Using vector representation can be represented as the following form:

\[
P(y_i = 1 | x_i) = \frac{\exp(\beta_0^T x_i)}{1 + \exp(\beta_0^T + \beta^T x_i)}.
\]

(2)

where \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \) and \( \beta^T = (\beta_1, \ldots, \beta_p) \). In general situation, researchers solve the above parameters by MLE. But in this paper, in order to combine the BP neural network to build composite method, we define an error function:

\[
\text{Cost}_{i} = \frac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2.
\]

(3)

Our purpose is to get the minimum of the error function, so the parameter estimation problem is equivalent to the MSE estimates, the objective function is:

\[
\text{Cost}_{i} = \frac{1}{2} \sum_{i=1}^{n} (p(y_i = 1 | x_i) - y_i)^2.
\]

(4)

In order to get the minimum of Eq. 4. First of all, put Eq. 2 into Eq. 4, then solve the partial derivative with respect to \( \beta_0 \) and \( \beta \), and let its value equal to zero. We get:

\[
\frac{\partial \text{Cost}_{i}}{\partial \beta_0} = \sum_{i=1}^{n} \left( \frac{1}{1 + \exp(-z)} - y_i \right) \frac{\exp(-z)}{(1 + \exp(-z))^2} = 0.
\]

(5)

\[
\frac{\partial \text{Cost}_{i}}{\partial \beta} = \sum_{i=1}^{n} \left( \frac{1}{1 + \exp(-z)} - y_i \right) \frac{\exp(-z)}{(1 + \exp(-z))^2} x_i = 0.
\]

(6)
where \( z = \beta_0 + \beta^T x \), the Eq. 5 and Eq. 6 can be calculated by iterative algorithm. We use the gradient descent algorithm to estimate \( \hat{\beta} \). Meanwhile, calculate the corresponding probability of estimator \( \hat{p}_j \), which represents the probability forecast of Logistic regression model.

2.2. BP model

Neural network is a kind of artificial intelligent technology that imitates the human brain information processing process. It has the properties of adaptive, self-organizing and strong ability of self-study. There are exist dozens of neural network. BP neural network was proposed by Rumelhart and McClelland in 1986. It is a kind of multi-layers, feedforward perceptron neural network with training of reverse error propagation algorithm, and has made significant achievements in the field of engineering. Its main characteristic is the signal to forward pass, the error back propagation, and the nonlinear learning ability. In the process of the forward pass, the input signal from the input layer through the hidden layers, and get to the output layer at last. Each layer of neurons only has effects on the state of neurons in the next layer. After signal get to output layer, the neurons begin to reverse transmission error if the output layer can’t meet the expected output as a result. The network according to the prediction error and the training data adjusts to the connection weights and thresholds in each hidden layer, in order to gain the best output effect.

BackPropagation \((\text{training}_\text{examples}, \eta, n_{\text{in}}, n_{\text{out}}, n_{\text{hidden}})\) algorithm:

- Create a feed-forward network with \( n_{\text{in}} \) input, \( n_{\text{hidden}} \) hidden units, and \( n_{\text{out}} \) output units.
- Initialize all network weights to small random numbers (eg. Between \(-.05 \) and \(.05\))
- Until the termination condition is met, do
  - For each \((x, y)\) in the \text{training}_\text{examples}, do
    1. Input the instance \( x \) to network and compute the \( o_u \) of every unit \( u \) in the network.
    2. For each network output unit \( k \), calculate its error term \( \delta_k \)
      \[
      \delta_k = o_k(1-o_k)(y_k-o_k) \tag{7}
      \]
    3. For each hidden unit \( h \), calculate its error term \( \delta_h \)
      \[
      \delta_h = o_h(1-o_h) \sum_{\text{outputs}} \omega_{hk} \delta_k \tag{8}
      \]
    4. Update each network weight \( \omega_{ji} \)
      \[
      \omega_{ji} \leftarrow \omega_{ji} + \Delta \omega_{ji} \tag{9}
      \]
The \text{training}_\text{examples} is a pair of the form \((x, y)\), where \( x \) is the vector of network input values, and \( y \) is the vector of target network output values. \( \eta \) is the learning rate, \( n_{\text{in}} \) is the number of network inputs, \( n_{\text{hidden}} \) is the number of units in the hidden layer, and \( n_{\text{out}} \) is the number of output units. The number of the hidden layer can use the empirical formula: \( n_{\text{hidden}} = \sqrt{n_{\text{in}}n_{\text{out}}} \).

The input from unit \( i \) into unit \( j \) is denoted \( x_{ji} \), and the weight from unit \( i \) to unit \( j \) is denoted \( \omega_{ji} \). In this paper, we use the error function:

\[
\text{Cost}_2 = \frac{1}{2} \sum_{i \in I} (y_i - o_i)^2. \tag{10}
\]

Here \( I \) is the set of output units in the network, \( y_i \) is the target value of unit \( i \), and \( o_i \) is the output of unit \( i \). In order to make the target as close as possible with output, that is equivalent to get the solution of the minimum optimization problem of the Eq. 10.
2.3. Optimal combination model

At present, there are kinds of single model in personal credit assessment. Each method has different performance in prediction accuracy, robustness and interpretability. Due to the complicated interactions and casual relationships in the internal system, single method has difficulty in comprehensive interpretation. So at this time, one natural idea is that we could combine these different models to make full use of the advantages of each model, so as to get a better model. In this paper, we combine Logistic model with BP neural network to construct the optimal linear combination model. Let \( \hat{y} \) be the prediction of the combination model, \( \hat{y}_1 \) and \( \hat{y}_2 \) are respectively the predictive value of Logistic model and BP neural network. Their relationship can be expressed as: \( \hat{y} = \omega_1 \hat{y}_1 + \omega_2 \hat{y}_2 \), where \( \omega_1, \omega_2 \) are the weights of model of \( \hat{y}_1 \) and \( \hat{y}_2 \), with \( \omega_1 + \omega_2 = 1 \). Define the error function of combination model:

\[
\text{Cost} = \frac{1}{2} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 .
\]

where \( \hat{y}_i \) is the predictive value of combination model, \( y_i \) is the real target value, \( \mathbf{y} = (y_1, \ldots, y_n)^T \).

Using vector form can be written as follows:

\[
\text{Cost} = \frac{1}{2} (\hat{\mathbf{y}} - \mathbf{y})^T (\hat{\mathbf{y}} - \mathbf{y}) .
\]

\[
\text{Cost}_i = \frac{1}{2} (\hat{y}_i - y_i)^T (\hat{y}_i - y_i) \quad (i = 1, 2) .
\]

We can get the relationship of error function between the combined model and the single model by inference:

\[
\text{Cost} = \omega_1^2 \text{Cost}_1 + 2 \omega_1 \omega_2 \rho + \omega_2^2 \text{Cost}_2 .
\]

Here \( \rho = \frac{1}{2} (\hat{y}_1 - y)^T (\hat{y}_2 - y) \), is defined to be the association of the two single model. So the problem of minimum of Eq. 11 is equivalent to solve the following optimization problem:

\[
\min \; \text{Cost} = \omega_1^2 \text{Cost}_1 + 2 \omega_1 \omega_2 \rho + \omega_2^2 \text{Cost}_2 \\
\text{subject to} \quad \omega_1 + \omega_2 = 1, \quad \omega_1, \omega_2 \geq 0 .
\]

By solving optimization problem (15), we have the optimal solution:

\[
\omega_1^* = \frac{\text{Cost}_2 - \rho}{\text{Cost}_1 - 2 \rho + \text{Cost}_2}, \quad \omega_2^* = \frac{\text{Cost}_1 - \rho}{\text{Cost}_1 - 2 \rho + \text{Cost}_2} .
\]

The optimal objective function value is the optimal value of the combination forecast:

\[
\text{Cost}^* = \frac{\text{Cost}_1 \text{Cost}_2 - \rho^2}{\text{Cost}_1 - 2 \rho + \text{Cost}_2} .
\]

3. Simulation

3.1. Dataset

The data source of this paper are derived from the website at [14]. This is a German credit database of 1000 credit history data, including 20 evaluation indicators, such as gender, credit history, the
loan purpose. Bank customer credit status can be divided into two kinds: good customer and bad customer, among which 700 are good customers, bad customers with 300 cases.

This dataset is adopted by many scholars to study personal credit problems. Tony Van Gestel used the Pol LS-SVM method to this dataset for personal credit assessment, it achieved good results for 76.3% predict accuracy of test data [10]. David Meyer used Linear-LS-SVM to get 75.4% accuracy [11].

3.2. Simulation results

In general literature, most researchers adopt the random sampling to partition the training dataset and test dataset. Due to the proportion of two types of customer is 3:7 in the total dataset, so we adopt a more reasonable method-stratified random sampling, to ensure the same proportion of good and bad customer in the training dataset and test dataset. In this paper, we use matlab software to simulation. First of all, recoding the original data and normalizing the recoding data; then using the stratified random sampling method to partition the dataset; inputting the processing data into our program at last. Here, Logistic model adopt gradient descent method to calculate the estimated parameters. In BP neural network, we use the linear combination function and the sigmoid function as a feedforward function and output function. At the same time, the reverse transmission error is using the algorithm of gradient descent, we can get the results of Table 1.

From the result, we can easily get the performance of BP neural network which is not stable, the highest classification accuracy is 74.0%, the lowest is 66.5%. But the average accuracy of Logistic-BP combination model has a strong robustness in training data and test data. The average accuracy of test data is reaching 77.3%. This is 2-7% higher than single model, so the result of the combination is pretty good.

In this paper, we compare the Logistic-BP with other methods. The result presents in Table2. The linear discrimination analysis (LDA), decision tree algorithm (C4.5), Naive Bayes (NB), polynomial least squares support vector machine (Pol LS-SVM), linear SVM (Lin SVM) and instance based learners (IB) were compared by Tony Van Gestel [10] in 2004. Radial basis function LS-SVM (RBF-LS-SVM), linear-LS-SVM and quadratic discrimination analysis (QDA) were applied by the paper [11] in 2002. Many ADALINE (MADALINE) and RBF-SVM in [12] were compared by Auer, Burgsteiner and Maass in 2002. The average accuracy of the literature

<table>
<thead>
<tr>
<th>method</th>
<th>dataset</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>train</td>
<td>0.7725</td>
<td>0.7700</td>
<td>0.7663</td>
<td>0.7900</td>
<td>0.7800</td>
<td>0.7758</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.7300</td>
<td>0.7650</td>
<td>0.7600</td>
<td>0.7500</td>
<td>0.7500</td>
<td>0.7510</td>
</tr>
<tr>
<td>BP</td>
<td>train</td>
<td>0.6950</td>
<td>0.7213</td>
<td>0.7125</td>
<td>0.6975</td>
<td>0.7675</td>
<td>0.7188</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.6950</td>
<td>0.6900</td>
<td>0.7200</td>
<td>0.6650</td>
<td>0.7400</td>
<td>0.7020</td>
</tr>
<tr>
<td>Logistic-BP</td>
<td>train</td>
<td>0.7825</td>
<td>0.7788</td>
<td>0.7800</td>
<td>0.8088</td>
<td>0.8338</td>
<td>0.7968</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.7650</td>
<td>0.7600</td>
<td>0.7800</td>
<td>0.7700</td>
<td>0.7900</td>
<td>0.7730</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>average accuracy</th>
<th>method</th>
<th>average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic-BP</td>
<td>77.30%</td>
<td>IB 10</td>
<td>72.60%</td>
</tr>
<tr>
<td>LDA</td>
<td>75.90%</td>
<td>RBF-LS-SVM</td>
<td>76.30%</td>
</tr>
<tr>
<td>C4.5</td>
<td>71.40%</td>
<td>linear-LS-SVM</td>
<td>75.40%</td>
</tr>
<tr>
<td>NB-n</td>
<td>74.70%</td>
<td>QDA</td>
<td>72.50%</td>
</tr>
<tr>
<td>Pol LS-SVM</td>
<td>75.40%</td>
<td>MADA-LINE</td>
<td>70.51%</td>
</tr>
<tr>
<td>Lin SVM</td>
<td>75.40%</td>
<td>RBF-SVM</td>
<td>76.35%</td>
</tr>
</tbody>
</table>
4. Conclusions

In this paper, we introduce the two most widely used model in personal credit assessment. One is the Logistic regression model for statistical model, the other is the artificial intelligence model-BP neural network. And we discuss the basic principle of this two kinds of method. After Combining the models by building a unified error function, we get the Logistic-BP neural network combined optimal model. By using the German commercial bank’s credit card data to carry on empirical research. The result shows that the combination model has good properties in prediction accuracy and robustness, and is much better than single model. Considering the small increase of the personal credit assessment can bring huge economic benefits, so the combined optimal model in the field of personal credit assessment will have a good application prospect.

References


