Volatility Forecast of Financial Performance Based on the Logit and LS-SVM Mixed Model

Guan-Hua ZHAO¹,a,*, Qian WANG¹,b

¹Accounting Institute, Shandong University of Finance and Economics, Jinan, Shandong, China
ascarzgh2006@sina.com, b3213276968@qq.com

*Corresponding author

Keywords: Mixed Model, Financial Performance, Volatility Forecast.

Abstract. In order to predict the volatility of the financial performance of the listed companies more effectively, constructed the Logit and LS-SVM mixed model. In the empirical study, to financial performance volatility of Chinese listed companies as the object of study, with 5-fold cross-validation accuracy as the evaluation criteria, it is shown that the prediction accuracy using combined system constructed by the mixed model is significantly higher than that of single model, confirmed the validity of the model.

Introduction

Throughout present situation of the domestic and foreign research, the classified prediction methods mainly include single model method and the method of multi-model combination. Due to the uncertainty of the single model itself, in recent years, many scholars tried to apply multiple model prediction technologies to classification forecasting of financial performance.

There are many methods of multi-model combination, but it can be roughly divided into two categories: the classifier disturbance and the sample disturbance. The classifier disturbance uses multiple models Integrated on a data set to construct the multi-model combination system. The sample disturbance is divided into two categories, one is to use different training sample subset to construct single model, another one is to use alternative methods of multiple features, through accessing to subset with different features to construct single model, then using a certain algorithm to combine with the single model. Sun(2008) and Sun Jie (2009) were two of the earliest ones to study the classifier disturbance, all of them have carried out on weighted voting fusion for the predicted results of each single model, the experimental results all showed that the multi-model combination system on the accuracy and stability is better than that of single model. Tsai (2008) and Jiang Yanxia (2009) and other people have studied the two ways of the sample disturbance, they believe that more than one single model that have high precision of prediction and the large difference between each other can form a good combined and predicted system, in addition, the number of the mixed forecasting model is not the more the better. In recent years, Shi Jiandong(2013), Liu Jian(2013), Xu Hong (2014), Li Song(2015), Zhang Yunong (2015) and other people also have done some systematic studies on mixed model.

In conclusion, when the scholars used the classifier disturbance to construct mixed model, they were all based on the combination of linear or nonlinear way, the setting of weight coefficient in combination had a great deal of subjectivity, and this paper uses the new perspective, using the output of the LS - SVM for the explained variable of Logit regression to provide output probability value, on this account, we can correct the rules of classification boundaries for 0.5 of Logit regression model and increase the accuracy of our predictions. In addition, almost all existing study put “the financial dilemma”(whether the listed companies are given special treatment or not) as the research object, this article put the volatility of financial performance of listed companies as the research object.
The Mixed Prediction Model Based on the Logit and LS-SVM

The Logit Regression Model

Assume for the company i, there is an invisible latent variable designed to depict the volatility characteristics of financial performance of company i, the behavior is determined by the following formula:

\[ Y_i^* = x_i^\prime \beta + \varepsilon_i, i = 1, 2, \cdots, N. \] (1)

Among them, N is the number of companies, also known as the sample size; W is k-dimensional explanatory variable used to describe the observable individual characteristics of the ith company, this article refers to all kinds of financial indicators. The discrete random variables that show values of 0 and 1, define it as follows:

\[ Y_1 = \begin{cases} 1, & \text{when sharp rise in the ith company's financial performance} \\ 0, & \text{when sharp fall in the ith company's financial performance} \end{cases} \] (2)

The relationship between \( Y_1 \) and \( Y_1^* \) is as follows:

\[ Y_1 = \begin{cases} 1, & Y_1^* > 0 \\ 0, & Y_1^* \leq 0 \end{cases} \] (3)

Namely when \( Y_1^* > 0 \), the performance of the company i increases sharply; When \( Y_1^* \leq 0 \), the performance of the company i decreases sharply. Here the critical value is 0, but in fact as long as \( x_1 \) contains the constant term, the selection of critical value is irrelevant, so it might be set to 0, thus:

\[ P(Y_1 = 1|x_1) = P(Y_1^* > 0) = P(\varepsilon_i > -x_i^\prime \beta) = 1 - F(x_i^\prime \beta) \]
\[ P(Y_1 = 0|x_1) = P(Y_1^* \leq 0) = P(\varepsilon_i \leq -x_i^\prime \beta) = F(x_i^\prime \beta). \] (4)

Among them, \( P(\cdot) \) is probability; \( F \) is distribution function about \( \varepsilon_i \), it must be a continuous function and increase monotonically. The type of distribution function determines the type of binary choice, if the error term follows logical distribution, namely \( F(x) = e^x/(1 + e^x) \), binary model is called the Logit model.

The LS-SVM Model

The general description of least squares support vector machine (LS-SVM) is as follows: given a set \( \{(x_i, Y_i), i = 1, 2, \cdots, L\} \) of include L training sample, the ith input data is \( x_i \in R^n \), the ith output data \( Y_i \in \{+1, -1\} \) is binary type variable. LS -SVM is to construct a classification function:

\[ f(x) = sgn(w^T \phi(x) + b). \] (5)

If sample x can be classified correctly, namely solving the following optimization problem:

\[ \min \frac{1}{2} \|w\|^2 + \frac{1}{2} C \sum_{i=1}^{L} \xi_i^2 \] (6)
\[ \text{s.t. } Y_i(w^T \phi(x_i) + b) = 1 - \xi_i. \] (7)

Among them, \( w = (w_1, w_2, \cdots, w_L)^T \) is the weight vector; \( C \) is the penalty factor; \( \xi_i \) is a slack variable whose value is greater than zero, \( \phi(x_i) \) called the mapping function. The dual problem of equation (6) and (7) is Lagrange polymerization:
\[ L(w, b, \alpha) = \frac{1}{2} ||w||^2 + \frac{1}{2} C \sum_{i=1}^{L} \xi_i^2 - \sum_{i=1}^{L} \alpha_i [Y_i(w^T \varphi(x_i) + b) + \xi_i - 1]. \] \tag{8}

Among them, \(\alpha_i\) is Lagrange multiplier, according to equation (8) making partial derivative respectively to \(w, b, \xi\) and \(\alpha_i\) makes it equal to zero, given \(Y_i \in \{+1,-1\}\) at the same time, after finishing we can obtain:

\[
\begin{bmatrix}
\mathbf{1} & 0 & 0 & -Z^T \\
0 & 0 & 0 & -Y^T \\
0 & 0 & \mathbf{C} & -\mathbf{I} \\
Z & Y & \mathbf{I} & 0
\end{bmatrix}
\begin{bmatrix}
w \\
b \\
\xi \\
\alpha
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0 \\
0 \\
1
\end{bmatrix}. \tag{9}
\]

In the formula, \(\mathbf{I} = (1, 1, \ldots, 1)^T, \xi = (\xi_1, \xi_2, \ldots, \xi_L)^T, Z = [\varphi(x_1)Y_1, \varphi(x_2)Y_2, \ldots, \varphi(x_L)Y_L]^T, Y = (Y_1, Y_2, \ldots, Y_L)^T, \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)^T, \mathbf{I}\) is a unit matrix. Eliminating \(\xi\) and \(w\), then reuse Mercer conditions:

\[ \Omega_{ij} = Y_iY_jK(x_i, x_j). \] \tag{10}

After finishing, the formula (10) is converted to:

\[
\begin{bmatrix}
b \\
\alpha
\end{bmatrix} = 
\begin{bmatrix}
0 & -Y^T \\
Y & \mathbf{\Omega} + \mathbf{I}/C
\end{bmatrix}^{-1}
\begin{bmatrix}
0 \\
1
\end{bmatrix}. \tag{11}
\]

Order \(A = \Omega + \mathbf{I}/C\), which is a semi-positive definite matrix, and its inverse matrix exists. To solve the system of linear equations, we can obtain:

\[ b = \frac{Y^TA^{-1}1}{Y^TA^{-1}Y}, \quad \alpha = A^{-1}(\mathbf{I} - Yb). \] \tag{12}

Therefore, the classification function is:

\[ f(x) = \text{sgn}\left(\sum_{i=1}^{L} \alpha_i Y_i K(x_i, x) + b\right). \] \tag{13}

In the formula (13), \(b\) and \(\alpha\) is obtained from formula (12), \(Y_i\) be confirmed according to the nature of the training sample, that is to say when the samples are the companies whose financial performance increase greatly, \(Y_i = 1\), and otherwise, \(Y_i = -1\). \(K(x_i, x)\) is a kernel function. This article uses the Gaussian kernel, its expression is \(K(x_i, x) = \exp(-|x_i - x|^2)/2\sigma^2\), \(\sigma^2\) is kernel Parameter. Because \(f(x)\) is symbol function, when \(f(x) = +1\), the company is considered as increased performance, when \(f(x) = -1\), the company is considered as reduced performance.

**Logit and LS-SVM Mixed Model**

This paper introduces the LS-SVM into traditional Logit model, using the output of the LS-SVM to provide support to the output probability of Logit regression, with, respectively represent mean value of the overall probability of the training sample set, mean value of probability of the positive class samples (increased performance company) and mean value of probability of the negative samples (declining performance company).

\[
\bar{p} = \text{mean}(p_i|Y_i = +1, i = 1,2,\ldots,N) \\
\bar{p}^+ = \text{mean}(p_i|Y_i = +1, i = 1,2,\ldots,N) \\
\bar{p}^- = \text{mean}(p_i|Y_i = -1, i = 1,2,\ldots,N). \tag{14}
\]
According to $\bar{p}^- , \bar{p} , \bar{p}^+ $, we can divide the output probability into four consecutive interval:

$$ I_1 = [0, \bar{p}^-), I_2 = [\bar{p}^-, \bar{p}], I_3 = [\bar{p}, \bar{p}^+), I_4 = [\bar{p}^+, 1] . $$

(15)

In the four interval, the prediction accuracy $f_1, f_2, f_3, f_4$ of the LS - SVM could be obtained and expressed in formula is as follows:

$$ f_i = \begin{cases} \frac{\text{num}\{x_i \in I_j \mid Y_i = -1, i = 1, 2, \ldots, N\}}{\text{num}\{x_i \in I_j \mid i = 1, 2, \ldots, N\}}, & x_i \in I_j, j = 1, 2 \\ \frac{\text{num}\{x_i \in I_j \mid Y_i = +1, i = 1, 2, \ldots, N\}}{\text{num}\{x_i \in I_j \mid i = 1, 2, \ldots, N\}}, & x_i \in I_j, j = 3, 4. \end{cases} $$

(16)

Algorithm steps are as follows:

(1) For the training sample set, calculating the output probability of Logit regression model, calculate, and divide the interval $[0, 1]$ into four continuous intervals.(2) On each interval, the training set is divided into four, recorded as, calculating the prediction accuracy of Logit and LS-SVM model respectively, recorded as.(3) According to the test set, on each interval, similarly, the test set respectively uses Logit and LS- SVM model to calculate.(4) according to the prediction accuracy of second step, establishing the following evaluation criteria: if , using the predicted results of Logit; otherwise, using the predicted results of LS-SVM.

The model changed the classification rules of traditional Logit with 0.5 as boundaries, and in accordance with the prediction accuracy of four intervals divided by the output probability compared with the LS-SVM, namely, using the results of LS-SVM to provide support for fuzzy intervals of Logit, on this account we can amend the prediction results of Logit and increase the prediction accuracy of financial performance.

The Empirical Study

The Sample Source and Selection

All the samples come from CSMAR database, using all manufacturing listed companies 2013-2015 data on Shanghai and Shenzhen Stock Markets, and the lead time of prediction is half a year. Eliminating the ST companies and the companies whose Performance volatility is less than 50%, The final selection of the number of research samples are 264, and randomly selected 70% of them as a training sample (about 176), the remaining 30% as a test sample (about 88), in the training and test samples, the rise and fall of performance accounted for 50% respectively.

Selection and Screening of Initial Indicators

<table>
<thead>
<tr>
<th>Financial Indicators</th>
<th>Financial Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Ratio(X1)</td>
<td>Return on Assets Ratio(X13)</td>
</tr>
<tr>
<td>Acid-test Ratio(X2)</td>
<td>Rate of Return on Total Assets(X14)</td>
</tr>
<tr>
<td>Working Capital to Total Assets Ratio (X3)</td>
<td>Rate of Return on Fixed Assets(X15)</td>
</tr>
<tr>
<td>Debt Asset Ratio(X4)</td>
<td>Increase Rate of Main Business Revenue(X16)</td>
</tr>
<tr>
<td>Long-term Debt Ratio</td>
<td>Increasing Rate of Fixed Assets(X17)</td>
</tr>
<tr>
<td>Debt to Equity Ratio(X6)</td>
<td>Total Assets Growth Rate(X18)</td>
</tr>
<tr>
<td>Long-term Debt to Equity Ratio(X7)</td>
<td>Net Profit Growth Rate(X19)</td>
</tr>
<tr>
<td>Receivables Turnover Ratio(X8)</td>
<td>Operating Income Cash Ratio(X20)</td>
</tr>
<tr>
<td>Inventory Turnover(X9)</td>
<td>Operating Activities Per Share in Cash Flows(X21)</td>
</tr>
<tr>
<td>Fixed Assets Turnover(X10)</td>
<td>Net Cash Flow Per Share (X22)</td>
</tr>
<tr>
<td>Total Assets Turnover(X11)</td>
<td>Sales Received Cash Ratio(X23)</td>
</tr>
<tr>
<td>Operating Profit Ratio(X12)</td>
<td>Net Profit to Cash Ratio(X24)</td>
</tr>
</tbody>
</table>
Based on the reference of related research at home and abroad, this paper finally chose 24 financial indicators, as shown in table 1.

On this basis, using the 24 primary indicators to have significant differences test of single indicator, there are a total of eight indicators have significant differences at 5% significance level, they are: X11, X12, X13, X14, X15, X19, X20 and X21. Finally, using the method of "down" stepwise Logit regression, that is to say, removing variables from the eight indicators one by one, to observe whether the prediction accuracy can be improved or not until the prediction accuracy can't be improved.

**Model Parameter Settings**

The key to predicting the effect of LS-SVM model lies in the choice of kernel function and punish coefficient, and the inappropriate choice will lead to over fitting or insufficient fitting, this article selected Gaussian kernel function with strong nonlinear mapping ability. From the training samples obtained the optimal penalty factor C and Gaussian kernel parameters \( \sigma^2 \) by using 5-fold cross-validation method, the search range is set to \([2^{-10},2^{+10}]\), and the step length are all set as 0.1.

**The Empirical Result Analysis**

On the training samples, the probability intervals are given by Logit are as follows:

- I1: \([0,0.35)\)
- I2: \([0.35,0.47)\)
- I3: \([0.47,0.58)\)
- I4: \([0.58,1]\)

The prediction accuracy of the test sample set is shown in Table 2:

<table>
<thead>
<tr>
<th>Model</th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>62.5</td>
<td>51.0</td>
<td>65.7</td>
<td>86.6</td>
</tr>
<tr>
<td>LS-SVM</td>
<td>62.8</td>
<td>65.5</td>
<td>78.1</td>
<td>80.3</td>
</tr>
</tbody>
</table>

According to the integrated criterion, using the prediction results of LS-SVM on I1, I2, I3 interval, using the prediction results of Logit on I2 interval, and compared with single model, obtaining the prediction results as is shown in table 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall prediction accuracy</th>
<th>The prediction accuracy of increased performance</th>
<th>The prediction accuracy of declining performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>70.3</td>
<td>80.6</td>
<td>60.0</td>
</tr>
<tr>
<td>LS-SVM</td>
<td>74.7</td>
<td>91.8</td>
<td>57.6</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>75.9</td>
<td>88.0</td>
<td>63.8</td>
</tr>
</tbody>
</table>

As is shown in table 3: (1) As to the overall prediction accuracy, compared with single Logit model, LS-SVM has better prediction performance, achieved 74.7%, but in I4 interval, its performance is not as good as Logit model (86.6%), and the mixed prediction model has the best accuracy (75.9%); (2) As to the samples of declining performance, results show that the prediction accuracy of mixed model is highest, achieved 63.8%, increases 3.8% than the Logit model, and 6.2% higher than that of the LS-SVM model; (3) As to the samples of increased performance, the prediction accuracy of mixed model declined compared with the LS-SVM, but have significant increase than that of the Logit model.

**Conclusion**

Based on the logit and LS-SVM integrated mixed model to predict the volatility of the listed companies, the empirical results shows that: The prediction accuracy of the mixed model for the overall sample and the samples of increased/declining performance was better than the single model
of logit and LS-SVM, this demonstrated that the mixed model is effective as proposed in this paper. At the same time, the empirical study also shows that the prediction accuracy of increased performance companies are generally better than declining performance companies in different models, which shows that the selected index data for the distinguishing ability of declining performance companies still exist certain room for improvement.

Acknowledgement
This research is supported by the Shandong Provincial Natural Science Foundation Project (ZR2012GM009).

Reference