A Two-stage Intelligent Model for Personal Credit Loans Default Prediction by Using SOM and PHM

Qiao-Ling Chen  
Student, Department of Business Management, Nanfang College of Sun Yat-sen University, Guangdong Province, 510970, China

Jian-Bang Lin  
Associate Professor, Department of Business Management, Nanfang College of Sun Yat-sen University, Guangdong Province, 510970, China

Jun-Ya Zeng  
Student, Department of Business Management, Nanfang College of Sun Yat-sen University, Guangdong Province, 510970, China

ABSTRACT: In the era of big data, it is very important for companies to mining the value of vast data in the data warehouse. This study provides a two-stage intelligent predicted model adds the concept of customer relationship management (i.e. customer segmentation). As a result, this study applied SOM to solve the urgent issue and find 3 groups of borrowers, then we predict the lifetime of borrowers by using PHM. The results show that the borrowers can be segmented effectively into different group by SOM and the performance of integrated model is the best.

1 INTRODUCTION

In the era of big data, it is very important for companies to mining the value of vast data in the data warehouse. Many companies have begun to use intelligent procedures finding the hidden information from consumer behaviors along with the rapid development of computer technology. The bank is one of the industries has big data, such as the application data of personal credit loans.

In the past, people in approval department often used their experience to judge whether to accept the applications. However, it not only wastes operation time but easily has biased due to manipulation. The banks will suffer huge losses if they loan to the people who is the potential default. Hence, to provide an intelligent procedure to predict default plays a crucial role in banks.

Single-stage default predicted models such as logistic regressions, discrimination analysis, decision tree, or different kinds of survival models are usually used in the large volume of previous studies. However, the lending behaviors are usually exposed to the different backgrounds of the customers. Thus, this study has a new idea and adds the concept of customer relationship management (i.e. customer segmentation) into intelligent technology to predict the default. This study aimed to build a two-stage default predicted model by using self-organizing mapping (SOM) to segment first, and by proportional hazard model (PHM) to predict the lifetime (default time). It will help banks identify different borrowers and make the rational decisions from big data.

The remainder of this study is organized as follows. In Section 2, the literature on SOM and PHM are briefly reviewed. In Section 3, the research design is proposed and explained. In Section 4, an empirical study is conducted to prove that the proposed model is workable. Finally, Section 5 contains the conclusions of this study.

2 LITERATURE REVIEW

SOM was provided by Kohonen in 1981 and widely applied in various fields. For example, it was used to find out the relationship between users and agents in telecommunications (Chien, 2006); to segment travelers (Li et al., 2010); to analyze students’ evaluations of teaching (Sorrosal Forradellas et al.,
2012); to know purchase intention in e-commerce (Hong & Kim, 2012); to improve the stock trading system (Pham et al., 2014), and in the transportation field (Yang et al., 2014) etc. However, it seems to rarely be applied in segmenting the borrowers of personal credit loans. So, this study attempted to segment the borrowers by SOM base on the past studies. It could achieve good performance in personal credit loans as well.

In addition, this study aims to estimate the lifetime besides predicting whether default or not. As reviewing, the major advantage of PHM among all predicted models is that it can accurately predict the lifetime and be excellent generalization performance in a wide range of problems. PHM first developed by Cox in 1972 and then widely used in prediction of event occurrence, such as the lifetime of online contents (Lee et.al, 2012), expected time of credit recovery (Ha & Krishinan, 2012), and the stability of parameter estimates (Leow & Crook, 2014) etc. Thus, after segmenting the borrowers of personal credit loans, this study applies the PHM to predict the lifetime of different group of borrowers.

3 RESEARCH DESIGN AND METHOD

3.1 Research design

In order to solve the problem of complicated information in personal credit loans, this study combines big data analysis intelligence with CRM concepts. The figure 1. is the description of every step.

![Research process chart](image)

As Figure 1. Shows, this study acquired data after communicating the proposed concept with the bank. Then, we understood the distribution of data and clean the data to confirm the completeness and correction. Furthermore, this study selects the important factors by association analysis (i.e. Chi-square test, Cramer’s V, t-test, Eta², correlation analysis) to select the important factors to default, and clusters the borrowers by SOM. PHM is applied to predict the lifetime of each segment samples and the whole samples, and finally comparing different result of models.

3.2 Data and variables selection

The data of personal consumer loans used in this study is from a bank in Taiwan, and initially there are 15,725 borrowers with 97 variables. The dependent variable (DV) in this study is repayment status including 3,233 default samples and 12,492 non-default ones. All of variables are divided into 3 groups: loan information, payment status, demographic information. According to Chang et al. (2007) suggestion, this study just adopts the group of “demographic information” and finds the important factors as inputs to SOM.

<table>
<thead>
<tr>
<th>category</th>
<th>Variable</th>
<th>p</th>
<th>Cramer’s V/Eta²</th>
<th>SOM</th>
<th>PHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic information</td>
<td>Cross-selling</td>
<td>0.000</td>
<td>0.083</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>0.002</td>
<td>0.022</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.000</td>
<td>0.116</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marital status</td>
<td>0.005</td>
<td>0.018</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Industry</td>
<td>0.000</td>
<td>0.125</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Income*</td>
<td>0.000</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment status</td>
<td>Total paymentamt*</td>
<td>0.000</td>
<td>0.017</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total interestamt*</td>
<td>0.000</td>
<td>0.011</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Loan information</td>
<td>Loans_type</td>
<td>0.006</td>
<td>0.015</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Approval level</td>
<td>0.010</td>
<td>0.016</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Before clustering and prediction, we select important factors to be used in SOM and PHM. This study uses two-steps to select important factors (Cramer’s V > 0.1 or Eta² > 0.01) at first, and then drops one of the variables if the correlation test of Cramer’s V > 0.5 and Eta² > 0.14 at the second step to prevent collinearity (Cohen, 1988) (see Table 1).

3.3 Research methods

3.3.1 Self-organizing map (SOM)

SOM is an unsupervised competitive neural network learning model proposed by Kohonen in 1981. It applies mapping technique to project multi-dimensional inputs to a two-dimensional space (see Figure 2.), and neighboring neurons will have similar response to one variable. The introduction of algorithm is as following:
Figure 2. Structure of SOM.

- At first, the weight \( w_i(0) \) is assigned randomly.
- Selecting vector \( x \) to train from samples. Finding out the winner neuron. Using minimum distance to decide winner neuron in \( n \) times according to equation (1).

\[
i(x) = \arg\min_j \| x(n) - w_j \|, \quad j = 1,2,...,l \quad (1)
\]

- Taking winner neuron as the center and adjusting the weight follow to equation (2).

\[
w_j(n+1) = w_j + \eta(n) h_{j,i(x)}(n) \left( x(n) - w_j(n) \right) \quad (2)
\]

- \( \eta(n) \) is learning rate, \( h_{j,i(x)}(n) \) denotes the neighborhood function of winner neuron \( i(x) \), \( \eta(n) \) and \( h_{j,i(x)}(n) \) will change continually.

Repeating steps when it does not change in the output layer (Kohonen, 1981; Haykin, 2008).

3.3.2 Proportional hazard model (PHM)
This study uses PHM to predict the lifetime. In particular, it is chosen because it is useful to predict the event occurred time with an appropriate baseline and time-dependent parametric. It can provide different survival probabilities in discrete time. It is a particular problem in statistical analysis, and it can be solved in this study. The brief introduction is as following equation (3)-(7) (Ha & Krishinan, 2012):

The probability density function \( f(t) \) means that someone can survive at time \( t \) but happen to die during \( \Delta t \).

\[
f(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t)}{\Delta t} \quad (3)
\]

Accumulative density function \( F(t) \) for the time \( t \) means that someone has died before or at time \( \Delta t \).

\[
F(t) = Pr(T \leq t) = \int_0^t f(u) \quad (4)
\]

\[
h(t) = \frac{1}{Pr(T \geq t)} = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t)}{\Delta t} = f(t) \quad (5)
\]

\[
S(t) = Pr(T \leq t) = 1 - F(t); \quad f(t) = -\frac{dS(t)}{dt} \quad (6)
\]

\[
h(t) = \frac{1}{Pr(T \geq t)} = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t)}{\Delta t} = f(t) \quad (5)
\]

\[
S(t) = \exp\left[- \int_0^t h(u) du \right] = \exp[-H(t)] \quad (7)
\]

Where \( H(t) \) is the cumulative hazard function.

4 RESULT

4.1 Borrowers segmentation
The result of SOM is shown in Figure 3. and this study decided to integrate neighboring group and find 3 groups finally. Neurons named \( x=2, y=2; x=2, y=1; x=2, y=0 \) are combined into one group, and neurons named \( x=0, y=0; x=0 y=1 \) are combined into another group. According to the default rates, the group \( x=0, y=2 \) is the highest (31.19%) that means this group are most likely to default, and this study named it high risk group (HR). Similarly, the group of \( x=2, y=2; x=2, y=1; x=2, y=0 \) is named medium risk (17.29%) group (MR), and the group of \( x=0, y=0; x=0 y=1 \) is named low risk (9.96%) group (LR).

![Figure 3. Output layer of SOM.](image)

4.2 Lifetime prediction model
There are two kinds of models are included in this research: the segment model and the whole model. The lifetime is decided by comparing probabilities with cut-off points in different time. If the probabilities are less than cut-off points, the borrowers are considered that he (or she) has the intense to default.
Taking HR group for example, Table 2 presents the results of 10 prediction cases. The case 2, 3, 4 are predicted to default in their seventh month, case 1 is predicted to default in sixth month, and case 7 is predicted to default in forth month. Moreover, if the probabilities are all higher than cut-point within the period under observation (16 months), these cases are predicted to keep non-default status within their 16 months.

### 4.3 Model performance

In order to confirm the stability of predicted models, this study applies 10-fold cross validation to build the PHM in both segment model and the whole model. This study divided the samples into 10 mutual exclusion groups, and selected one group as test data to validate the model in each time. Remains are train data to construct default predicted model.

In this study, we adopt the revised concept of confusion matrix (Berry & Linoff, 2004) to calculate the models’ accuracy rate (AR). As Table 3 shows, the gray area is successfully predicted. It represents that the predict lifetimes are earlier than actual lifetime. Accordingly, the AR of one of the whole models is 69.62% (=9852/14152 × 100%)

This study provides the best integrate model to prove prediction of the lifetimes of borrowers, as a result, this study decide to select the best AR of each group. From the Table 4, we find that both train data and test data’s average ARs in segment models (68.72%, 68.69%) are above whole models (67.70%, 67.71%). Thus, comparing with the whole model, to segment borrowers can provide more efficient prediction and obtain more information.

Then we select HR and MR in the whole model, when both AR of HR and MR (61.62%, 62.04% and 72.54, 73.14%) are respectively higher than the segment model (56.35%, 56.51% and 71.82%, 71.74%). LR from segment model is selected,
because the AR (82.02%, 81.78%) are higher than the whole model (69.67%, 69.34%). Therefore, HR and MR in the whole model are combined with LR of segment model, to design an integrated model with higher AR (70.95%, 71.26%) than other groups.

Table 4. AR of different groups.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>68.25</td>
<td>67.50</td>
<td>66.82</td>
<td>67.16</td>
<td>68.10</td>
<td>66.75</td>
<td>65.93</td>
<td>68.02</td>
<td>68.91</td>
<td>69.62</td>
<td>67.70</td>
</tr>
<tr>
<td>Test</td>
<td>70.95</td>
<td>70.48</td>
<td>69.04</td>
<td>69.68</td>
<td>70.06</td>
<td>67.71</td>
<td>66.56</td>
<td>68.05</td>
<td>64.35</td>
<td>60.27</td>
<td>67.71</td>
</tr>
<tr>
<td><strong>HR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>62.32</td>
<td>61.39</td>
<td>60.73</td>
<td>61.66</td>
<td>62.28</td>
<td>61.38</td>
<td>59.13</td>
<td>60.36</td>
<td>63.62</td>
<td>63.34</td>
<td>61.62</td>
</tr>
<tr>
<td>Test</td>
<td>63.44</td>
<td>63.95</td>
<td>69.05</td>
<td>65.99</td>
<td>62.59</td>
<td>61.39</td>
<td>69.90</td>
<td>63.10</td>
<td>49.15</td>
<td>51.87</td>
<td>62.04</td>
</tr>
<tr>
<td><strong>MR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>72.89</td>
<td>72.16</td>
<td>71.59</td>
<td>71.73</td>
<td>72.90</td>
<td>70.85</td>
<td>71.18</td>
<td>74.78</td>
<td>72.90</td>
<td>74.45</td>
<td>72.54</td>
</tr>
<tr>
<td>Test</td>
<td>74.30</td>
<td>75.26</td>
<td>71.45</td>
<td>75.31</td>
<td>72.85</td>
<td>75.83</td>
<td>67.08</td>
<td>83.30</td>
<td>75.44</td>
<td>60.60</td>
<td>73.14</td>
</tr>
<tr>
<td><strong>LR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>70.28</td>
<td>69.74</td>
<td>68.88</td>
<td>68.66</td>
<td>69.71</td>
<td>68.74</td>
<td>68.34</td>
<td>69.60</td>
<td>70.90</td>
<td>71.86</td>
<td>69.67</td>
</tr>
<tr>
<td>Test</td>
<td>77.00</td>
<td>71.26</td>
<td>65.70</td>
<td>72.71</td>
<td>76.81</td>
<td>65.46</td>
<td>61.11</td>
<td>54.11</td>
<td>77.48</td>
<td>71.74</td>
<td>69.34</td>
</tr>
</tbody>
</table>

| **Whole model** |     |     |     |     |     |     |     |     |     |     |         |
| **Segment model** |     |     |     |     |     |     |     |     |     |     |         |
| **Integrated model** |     |     |     |     |     |     |     |     |     |     |         |
| **Whole model** |     |     |     |     |     |     |     |     |     |     |         |
| **Segment model** |     |     |     |     |     |     |     |     |     |     |         |
| **Integrated model** |     |     |     |     |     |     |     |     |     |     |         |

5 CONCLUSION

More and more companies realize mining useful information from a large number of data is not easy in the era of big data. They have begun to use intelligent technology and try to find the hidden information from consumer behaviors. This study provides a two-stage intelligent predicted model and adds the concept of borrower segmentation into default predicted model. In order to improve the process of approval departments and ensure the smooth operation of the bank, banks must effectively prevent the occurrence of default. Recognizing bad and good borrowers is essential part to improve the process, from the standpoint of customer relationship management (CRM), predicting lifetime of borrowers is also important to ensure the steadily operate in banks. As a result, this study applied SOM to solve the urgent issue and find 3 groups of borrowers, then we predict the lifetime of borrowers by using PHM.

We divide borrowers into 3 groups according their personal information. The result shows that different groups have significant characteristics. Highest risk group and higher risk group income average are much higher than other groups and most of them work in finance or insurance industry. It means that people with high income may default easily in this study. It is different from common sense that negative correlation between income and default. However, the possible reason is that the loan purpose of the borrowers in this group is investment. On the one hand, borrowers get the loan easily due to their high income; on the other hand, they judge themselves better than others in regard to investment skills or personality attributes according to overconfidence theory (Baker & Nofsinger, 2010).

They often ignore or underestimate some objective information contrary to their knowledge and tend to take enormous risk in investment just depend on their subjective information. It may cause to fail easily in investment and then default.

This research applies PHM to predict lifetimes of two kinds of data. The results show to segment borrowers by SOM can improve the prediction of lifetimes.

We find that the whole models’ ARs of HR and MR in are higher than the segment models’. This study also find that AR of LR in the segment model is higher than AR of LR in overall sample model. As a result, HR and MR in the whole models are combined with LR in the segment model. The AR of integrated model is significantly higher than AR of the whole model.

This study suggest that banks can use integrated models to predict the lifetime of borrowers and design appropriate projects to different groups, it can help banks develop their personal credit loan efficiently. In further study, we will continuously seek more significant factor to model better prediction framework.
REFERENCES


273