Genetic Algorithm Approach to Design Covariates of Binomial Logit Model for Estimation of Default Probability

TEZUKA Masaru, ITO Yoichi, and MUNAKATA Satoshi

Abstract—Credit risk management is one of the most important tasks of financial institutes. Default probability is the probability that a company will go into default, or be unable to fulfill an obligation, and it is a critical information for credit administration. Binomial logit model is widely used for default probability estimation. The formulas for computing covariates used in the model are designed by human experts in trial-and-error way, based on their experience. In this paper, we propose a method to design covariates. Integer-coded GA is employed and a representation of the chromosome is proposed for the purpose of optimizing the covariate. The method optimizes the covariates using the GA and estimates the coefficient of binomial logit model using Broyden-Fletcher-Goldfarb-Shanno method. The method is tested on an actual data provided for evaluation by a bank. The result of the experiment shows the method outperformed the human design.

I. INTRODUCTION

Credit risk management is one of the most important tasks of banks. A primary mission of credit administration departments of banks is to assess debt-paying ability of their customers and to assign credit score to them. According to the credit score, banks make decision on financing to the customers such as the amount of credit limit, interest rate, and security requirements.

Failure of financial institutions has a tremendous social impact both domestically and internationally. Basel Capital Accord[1] calls for retaining adequate capital to endure unexpected losses. In order to allow banks to continue their business and contribute to the economic security, rigorous and appropriate assessment is required.

Default probability is the probability that a company will go into default, or be unable to fulfill an obligation. It is critical information for credit administration. Default probability of company is estimated from historical default records. Binomial and Multinomial logit model, discriminant analysis, survival analysis, and the combination of these methods are used for the estimation. Binomial logit model is most commonly used.

Medium and small size regional banks use rating agencies to obtain default probability of client companies. They purchase probability data quarterly or half-yearly. When a new customer applies for a loan and probability of its default is not included in the data purchased from the rating agency in advance, the credit administration have to assign a credit score to the new customer without default probability. If the regional banks can estimate the unknown default probability of the new customer from the known probability of the other customers, the reliability of their credit score increases.

In section II, binomial logit model is briefly described which is widely used for default probability estimation. A method we propose in this paper also takes advantage of the model. Design of covariates used in the model is a difficult task. In section III, we propose a method to design covariates and to estimate the coefficient of binomial logit model based on the known default probabilities. The result of numerical experiment for evaluating the proposed method is shown in section IV. Section V is the conclusion of this paper.

II. ESTIMATION OF DEFAULT PROBABILITY

A. Item Response Theory and Binomial Logit Model

Binomial logit model, which is based on item response theory, is widely used to estimate default probability in financial industry. In this section, the model is briefly explained.

According to item response theory, a response of an entity to an item is considered to be generated when linear combination of covariates representing the condition of the entity falls below (or exceed) a certain threshold. In the case of default probability, when the linear combination of the covariates representing the financial condition of a company falls below the threshold, the company defaults in response to the deterioration of the financial condition. Some of the covariates are not observable. Such covariates are treated as a stochastic variate.

We have \( n \) companies and \( m \) observable covariates for each company. \( x_{ij} (i = 1, ..., n \text{ and } j = 1, ..., m) \) denotes \( j \)-th covariate of company \( i \). \( b_j \) is \( j \)-th coefficient of the linear combination. According to item response theory, company \( i \) goes into default when

\[
\sum_{j=1}^{m} b_j x_{ij} + \varepsilon < \theta
\]

where \( \theta \) is the threshold, and \( \varepsilon \) is the stochastic variate which represents the unobservable covariates.

For convenience, we replace the threshold \( \theta \) by \(-\theta_0\) and introduce \( x_{i0} \). \( x_{i0} = 1 \) for any \( i \). Then default rate \( d_i \), that
is the probability that the company $i$ defaults, is written as,

$$d_i = P \left( \sum_{j=1}^{m} b_j x_{ij} + \varepsilon < \theta \right)$$

$$= P \left( \varepsilon < \theta - \sum_{j=1}^{m} b_j x_{ij} \right)$$

$$= P \left( \varepsilon < -\sum_{j=1}^{m} b_j x_{ij} \right)$$

$$= P \left( \varepsilon < -1b^T x_i \right)$$

where $b = (b_0, b_1, ..., b_m)^T$ and $x_i = (x_{0i}, x_{1i}, ..., x_{mi})^T$.

When stochastic variate $\varepsilon$ follows standard normal distribution, i.e. $\varepsilon \sim N(0, 1)$,

$$d_i = \Phi(-b^T x_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-b^T x_i} e^{-z^2/2} dz$$

(1)

where $\Phi(z)$ is CDF (cumulative distribution function) of normal distribution. Equation (1) is called binomial probit model. Binomial probit model is not easy to deal with because CDF of normal distribution involved in the model can not be expressed in closed form.

Logistic distribution[2] is generally used to model the stochastic variable. The CDF of logistic distribution with location parameter $\mu$ and scale parameter $s$ is

$$F(z; \mu, s) = \frac{1}{1 + e^{-(z-\mu)/s}}.$$ When $\varepsilon$ follows logistic distribution with $\mu = 0$ and $s = 1$, default probability of company $i$ is

$$d_i = F(-b^T x_i; 0, 1) = \frac{1}{1 + e^{b^T x_i}}$$

(2)

and (2) is called binomial logit model.

Logistic distribution has longer tails and a higher kurtosis than normal distribution, however, binomial logit model is widely used to estimate default probability because it is expressed analytically and is easy to handle. It is known that logistic distribution is very similar to normal distribution and

$$\Phi(z) \approx \frac{1}{1 + e^{-1.7z}}.$$ B. Estimation of Coefficients

In order to estimate coefficients $b$, maximum likelihood estimation method is widely used. Default probabilities of the companies are very low generally. Therefore, least-square method may cause biased estimation and is not used.

On the assumption that the default of each company is independent, the likelihood function is written as

$$L(b) = \prod_{i=1}^{n} \left( d_i^b (1-d_i)^{1-d_i} \right)$$

where

$$\delta_i = \begin{cases} 1 & \text{customer } i \text{ defaults}, \\ 0 & \text{otherwise}. \end{cases}$$

Therefore, log likelihood function is

$$l(b) = \sum_{i=1}^{n} \left( \delta_i \ln(d_i) + (1 - \delta_i) \ln(1 - d_i) \right).$$

(3)

$d_i$ is either (1) when binominal probit model is used or (2) when binominal logit model is used. As mentioned above, binominal probit model is not expressed in closed form, hence time consuming numerical integral is required during coefficient estimation of the probit model. Thus, binominal logit model is generally used. Furthermore, it is known that there are practically no differences between the estimated results of probit and logit model[3].

Estimated coefficient $\hat{b}$ which maximizes $l(b)$ is a consistent estimator. It is known that asymptotic distribution of $\hat{b}$ is normal distribution with large enough samples.

Solving simultaneous equation

$$\frac{\partial l}{\partial b_{ij}} = \sum_{i=1}^{n} (d_i - \delta_i) x_{ij} \quad j = 0, 1, ..., m$$

coefficient $\hat{b}$ is calculated. Non-linear optimization method such as quasi-Newton method is used to solve the equation.

C. Design of Covariates

In the previous sections, two estimation models of default probability, binominal probit model and binominal logit model, are described as the covariates are given.

Usually the covariates are brought from financial statements of the companies. Financial statements including balance sheet, income statement, and cash flow statement are records of company’s financial condition. Balance sheet is a statement of the assets, liabilities and capital of a company at a certain time point. Income statement shows the revenue, expenses, and profit or loss during a certain period of time. Cash flow statement shows the change in cash and cash equivalents during a certain period.

Undisclosed information of the companies is used as well as the financial statements if available as a matter of course.

The value of each item in the statement is rarely used as a covariate directly. Several items are selected and converted into covariates. For example,

$$x_{0i} = \frac{(\text{current asset})_i - (\text{current liability})_i}{(\text{total equity})_i}$$

(4)

where $(\text{item})_i$ is the amount of the item of company $i$. Converted into the ratio to the total equity, the covariate represents the financial condition of a company independently from the size of the company.

The formulas for computing covariates are designed by human experts in trial-and-error way based on their experience. There is no mathematical model which is widely accepted to design appropriate covariates.
D. Practical Problems of Regional Banks

Credit risk management is one of the most important tasks of banks. A credit administration department of a bank assigns credit scores to their customers. According to the score, they make decision on financing to the customers such as the amount of credit limit, interest rate, and security requirements.

The department investigates debt-paying ability of the customers based on their financial statements and other information. Default probability of each customer is the important information for the credit score assignment. The rates are also used to develop financing portfolio of the bank [4].

Medium and small size regional banks do not analyze and estimate default probability by themselves but purchase its data from rating agencies quarterly or half-yearly because of the fact that the banks have neither enough research capacity nor resources for investigation. Usually, the number of customers of the regional banks is relatively small. Accordingly, the banks do not have enough volume of records required for maximum likelihood estimation. That is to say, they have neither a formula or an expertise for designing covariates nor enough records of \( b_i \) in (3).

From the viewpoint of cost effectiveness, utilization of the external rating agency is reasonable. On the other hand, when a new customer applies for a loan and probability of its default is not included in the data purchased from the rating agency, the credit administration have to assign a credit score to the new customer without default probability.

If the regional banks can estimate the unknown default probability of the new customer from the known probability of the other customers, the reliability of their credit score increases.

III. GENETIC ALGORITHM TO ESTIMATE DEFAULT PROBABILITY OF NEW CUSTOMERS

A. Requirement Analysis

We develop a software system for a regional bank to estimate the default probability of new customers. As mentioned above, the bank does not have an expertise for designing covariates. Thus, the system designs or optimizes the formula for designing covariates. The system also estimates the coefficient \( b_i \) from the known default probability of existing customers purchased from the rating agency.

At first, we planned to employ genetic programming (GP) to design covariates. GP is a powerful tool to evolve computer programs, mathematical formulas, and so on [5]. However, it is known that GP often generates very complex formula which human can not understand.

The credit administration of the bank attaches importance to the understandability of covariates. It prefers the covariates be the ratio of several financial items similar to (4).

For that reason, we introduce genetic algorithm (GA) [6] and propose a representation of the chromosome for the purpose of optimizing the covariate of binomial logit model.

B. Representation for Designing Covariates

There are \( t \) items on financial statements and \( y_{ik} \) \((i = 1, \ldots, n, k = 1, \ldots, t)\) denotes the value of \( k \)-th item on the financial statements of company \( i \).

In order to avoid division by zero at later stage, suffixes are given so that \( y_{i1}, y_{i2}, \ldots, y_{it} \) \((u \leq t)\) are nonvanishing and \( y_{it+1}, \ldots, y_{it} \) may become zero.

The GA we employ is a kind of integer coded genetic algorithm. Figure 2 shows the representation of chromosome we proposed. The length of the chromosome is \( 5 \times m \). \( m \) is the number of covariates for each customer. Domain of each element on the chromosome is,

\[
\begin{align*}
    h_j &\in \{1, 2, \ldots, 6\}, \\
    k_{j1}, k_{j2} &\in \{1, 2, \ldots, t\}, \\
    k_{j3}, k_{j4} &\in \{1, 2, \ldots, u\}
\end{align*}
\]

for \( j = 1, \ldots, m \). A set of five elements \( h_{j1}, k_{j1}, k_{j2}, k_{j3}, k_{j4} \)
The length of chromosome is \(5 \times m\)

\[ h_j \quad k_{j1} \quad k_{j2} \quad k_{j3} \quad k_{j4} \]

Five elements are used to design one covariate.

Fig. 2. Representation of chromosome.

is used to design \(j\)-th covariate of all customers.

\[
x_{ij} = \begin{cases} 
\frac{y_i(k_{j3}) + y_i(k_{j4})}{y_i(k_{j3}) + y_i(k_{j4})} & \text{if } h_j = 1; \\
\frac{y_i(k_{j3}) - y_i(k_{j4})}{y_i(k_{j3}) + y_i(k_{j4})} & \text{if } h_j = 2; \\
\frac{y_i(k_{j3}) + y_i(k_{j4})}{y_i(k_{j3}) + y_i(k_{j4})} & \text{if } h_j = 3; \\
\frac{y_i(k_{j3}) - y_i(k_{j4})}{y_i(k_{j3}) + y_i(k_{j4})} & \text{if } h_j = 4; \\
\frac{y_i(k_{j3}) + y_i(k_{j4})}{y_i(k_{j3}) + y_i(k_{j4})} & \text{if } h_j = 5; \\
\frac{y_i(k_{j3})}{y_i(k_{j3})} & \text{if } h_j = 6.
\end{cases}
\] (6)

From the chromosome, the covariates are designed and calculated according to (6). The domain of \(k_{j3}\) and \(k_{j4}\) is \(1, \ldots, n\) and \(y_{i1}, \ldots, y_{i6}\) do not have zero. In addition, the denominators in (6) consist of one item or the sum of two items. Therefore, the denominators are nonvanishing and division by zero is avoided.

At the initial generation, all the individuals in the population are initialized randomly in the range of the domain specified in (5).

Uniform crossover[7] is performed to reproduce next generation population. Mutation is also performed. The value of the element on which mutation is performed is changed randomly in the range of the domain. Elitism is also employed.

C. Fitness Function

Mean squared error of the estimated default probability is

\[
g(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n} \sum_{i=1}^{n} \left( d_i - \frac{1}{1 + e^{b^T \mathbf{x}_i}} \right)^2 \] (7)

where \(d_i\) is the known default probability of customer \(i\). Coefficient \(b\) have to be decided so that equation (7) is minimized. Thus, the fitness of the covariates represented by the chromosome is

\[
f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \arg \min_b g(b; \mathbf{x}_1, \ldots, \mathbf{x}_n). \] (8)

Step 1 Initial search point \(b_0\) is set to 0 and the inverse of approximate Hessian \(H_0\) is set as an unit matrix.

Step 2 Calculate search direction as

\[
d_v = -H_v \nabla g(b_v),
\]

where

\[
\nabla g(b) = \left( \frac{\partial g}{\partial b_1}, \ldots, \frac{\partial g}{\partial b_m} \right)^T.
\]

Step 3 Decide step length \(\alpha_v\) so that the length satisfies Armijo condition.

Step 4 Update search point as

\[
b_{v+1} = b_v + \alpha_v d_v,
\]

and if \(\nabla g(b_{v+1}) = 0\) or \(|b_{v+1} - b_v| < \varepsilon\), \(b_{v+1}\) is the solution \(b_v\) is a parameter to control termination condition.

Step 5 Update the inverse of approximate Hessian by

\[
H_{v+1} = H_{v} - \frac{H_{v} y_{v} y_{v}^T + s_{v} (H_{v} y_{v})^T}{y_{v} y_{v}^T + s_{v} s_{v}^T} + \left( 1 + \frac{y_{v}^T H_{v} y_{v}}{y_{v} y_{v}^T} \right) \frac{s_{v} s_{v}^T}{y_{v} y_{v}^T},
\]

where

\[
s_{v} = b_{v+1} - b_{v},
\]

\[
y_{v} = \nabla g(b_{v+1}) - \nabla g(b_{v}).
\]

Step 6 Repeat from step 2.

Quasi-Newton method is used to minimize (7) given \(\mathbf{x}_1, \ldots, \mathbf{x}_n\). We employ Broyden-Fletcher-Goldfarb-Shanno (BFGS) method as a quasi-Newton method[8]. Sherman-Morrison formula is also applied to calculate inverse matrix of approximate Hessian efficiently. Step length is decided so that the length satisfies Armijo condition[9]. Figure 3 describes the procedure to estimate \(b\).

Partial differential of \(g\) with respect to \(b_j\) used at step 2 in figure 3 is

\[
\frac{\partial g}{\partial b_j} = 2 \frac{1}{n} \sum_{i=1}^{n} e^{b^T \mathbf{x}_i} \left( d_i - \frac{1}{1 + e^{b^T \mathbf{x}_i}} \right) \frac{1}{(1 + e^{b^T \mathbf{x}_i})^2} x_{ij}.
\]

IV. NUMERICAL EXPERIMENT

Proposed method is tested on the data consisting of financial statements and default probability of 240 companies. Financial statements have 60 items of three years. Actually the statements have more than 150 items, however, the data of 60 items are provided for evaluation to us by the bank. The total number of items, \(t\), is 180 (\(=60\) items \(\times3\) years).

The method is evaluated with cross-validation[10]. Evaluation could be more accurate with leave-one-out cross-validation. However, the validation consumes very huge computation time. Because of the limitation of time, we employ 20-fold cross-validation.
TABLE I
PARAMETER SETTINGS.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>500</td>
</tr>
<tr>
<td>Termination generation</td>
<td>500</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.6</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.2</td>
</tr>
<tr>
<td>Elite size</td>
<td>10</td>
</tr>
<tr>
<td>Tournament size</td>
<td>2</td>
</tr>
<tr>
<td>ε (Termination condition of BFGS)</td>
<td>$1.0 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

TABLE II
ACCURACY OF DEFAULT PROBABILITY ESTIMATION ON TEST DATA SET

<table>
<thead>
<tr>
<th></th>
<th>Proposed method</th>
<th>Human designed covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>$2.00 \times 10^{-2}$</td>
<td>$2.41 \times 10^{-2}$</td>
</tr>
<tr>
<td>VAR</td>
<td>$3.57 \times 10^{-1}$</td>
<td>$7.23 \times 10^{-1}$</td>
</tr>
<tr>
<td>COR</td>
<td>0.43</td>
<td>0.25</td>
</tr>
</tbody>
</table>

240 companies are divided into 20 groups or 20 folds. Each group has 12 companies. One group is extracted and is used as a test data set. Remaining 19 groups (228 companies) are used as training data set. Changing the group extracted for a test, the method is trained and tested 20 times. Each time, it is trained on 228 companies and is tested on 12 companies. 20 results from the groups are combined and then mean squared error and correlation coefficient are calculated.

The proposed method is implemented using Microsoft Visual C++ 2005, and is run on Windows XP professional sp 2 PC with an Intel Pentium IV 3.6GHz and 1Gbytes RAM. The parameter settings used in the experiment are shown in table I. It takes about 3 hours 26 minutes (average 206.3 minutes, variance is 2.748) for one optimization run. The computation time depends on the number of companies used for training.

For the purpose of comparison, an experiment to estimate default probability with human designed covariates is also performed. In this case, covariates are designed by the credit administration department of the bank. The coefficients $b$ for binomial logit model are estimated from training data. The covariates are shown in figure 4.

Table II shows the result. MSE shows the mean squared error between estimated default probability and actual default probability. In this paper, actual probability means purchased data from the rating agency, that is, the probability estimated by the agency. Smaller MSE is better. VAR shows the variance of squared error. COR shows the correlation coefficient between estimated and actual default probability. As for correlation coefficient, it indicates the level of linear relationship between estimated and actual default probability and the estimates correspond to the actual completely when the coefficient is one. Table III shows MSE, VAR, and COR on training data set.

The result of proposed method is better than that of human. Welch’s t-test is done and significant probability is $0.055$. We can not reject the null hypothesis with significance level of five percent. However, the proposed method works better enough.

As the article of Hedberg says, GAs consider orders-of-magnitude more factors and options than human in a fraction of the time, finding new and sometimes unexpected designs and trends[11].

Figure 5 shows an example of covariates designed by the proposed method. Figure 6 show the relation between estimates by proposed method and actual default probability.

V. CONCLUSION

In this paper, we propose a method to design covariates of binomial logit model for estimating default probability. We employ integer-coded GA and devise representation of the chromosome for the purpose of optimizing the covariate of binomial logit model. The method optimizes the covariates by the GA and estimates the coefficient of binomial logit model by Broyden-Fletcher-Goldfarb-Shanno method, which is one of quasi-Newton methods.

We implement the method as a software system and it is tested on an actual data provided for evaluation by a bank. A numerical experiment to evaluate the proposed method is performed. The covariates optimized by the method are

\[
\begin{align*}
X_1 &= \frac{\text{current assets of this year}}{\text{total assets of this year}_i}, \\
X_2 &= \frac{\text{current liabilities of this year}}{\text{total assets of this year}_i}, \\
X_3 &= \frac{\text{current income of this year}}{\text{total assets of this year}_i}, \\
X_4 &= \frac{\text{fixed assets of this year}}{\text{total assets of this year}_i}, \\
X_5 &= \frac{\text{revenue of this year}}{\text{total assets of this year}_i}, \\
X_6 &= \frac{\text{equity capital of this year}}{\text{current liabilities of this year}_i},
\end{align*}
\]

Fig. 4. Covariates designed by human

\[
\begin{align*}
X_1 &= \frac{\text{current assets of this year}}{\text{total assets of this year}_i}, \\
X_2 &= \frac{\text{current liabilities of this year}}{\text{total assets of this year}_i}, \\
X_3 &= \frac{\text{current income of this year}}{\text{total assets of this year}_i}, \\
X_4 &= \frac{\text{fixed assets of this year}}{\text{total assets of this year}_i}, \\
X_5 &= \frac{\text{revenue of this year}}{\text{total assets of this year}_i}, \\
X_6 &= \frac{\text{equity capital of this year}}{\text{current liabilities of this year}_i},
\end{align*}
\]
Fig. 5. Example of covariates optimized by the proposed method.

compared to the covariates designed by humans. The result of the experiment shows the method outperformed the human design.

This time we do not take advantage of GPs because of the requirement of the bank and the limitation of cost and time. However, we would like to compare the method and GPs in future work.

BFGS method is utilized in the method to calculate the coefficients of binomial logit model and to evaluate fitness. BFGS method runs until convergence every time for every individual at every generation. Accuracy of fitness evaluation is not necessarily required during early stage of optimization. We are planning to introduce accuracy control techniques\cite{12, 13, 14} in order to improve efficiency.

REFERENCES

\begin{enumerate}
\item Basle Committee on Banking Supervision, “Amendment to the Capital Accord to Incorporate Market Risks”, Bank for International Settlements, 1996.
\item D. S. Madalla, Limited Dependence and Qualitative Variables in Econometrics, Cambridge University Press, 1983.
\item P. E. Gill, W. Murray and M. H. Wright, Practical Optimization, Elsevier, 1983.
\end{enumerate}