Corporate Credit Rating based on Bankruptcy Probability Using AdaBoost Algorithm-based Support Vector Machine

Taeksoo Shin
Division of Business Administration,
College of Government and Business,
Yonsei University
(tsshin@yonsei.ac.kr)

Taeho Hong
Department of Business Administration,
School of Business,
Pusan National University
(hongth@pusan.ac.kr)

Recently, support vector machines (SVMs) are being recognized as competitive tools as compared with other data mining techniques for solving pattern recognition or classification decision problems. Furthermore, many researches, in particular, have proved them more powerful than traditional artificial neural networks (ANNs) (Amendolia et al., 2003; Huang et al., 2004, Huang et al., 2005; Tay and Cao, 2001; Min and Lee, 2005; Shin et al., 2005; Kim, 2003).The classification decision, such as a binary or multi-class decision problem, used by any classifier, i.e. data mining techniques is so cost-sensitive particularly in financial classification problems such as the credit ratings that if the credit ratings are misclassified, a terrible economic loss for investors or financial decision makers may happen. Therefore, it is necessary to convert the outputs of the classifier into well-calibrated posterior probabilities-based multiclass credit ratings according to the bankruptcy probabilities. However, SVMs basically do not provide such probabilities. So it required to use any method to create the probabilities (Platt, 1999; Drish, 2001).

This paper applied AdaBoost algorithm-based support vector machines (SVMs) into a bankruptcy prediction as a binary classification problem for the IT companies in Korea and then performed the multi-class credit ratings of the companies by making a normal distribution shape of posterior bankruptcy probabilities from the loss functions extracted from the SVMs. Our proposed approach also showed that their methods can minimize the misclassification problems by adjusting the credit grade interval ranges on condition that each credit grade for credit loan borrowers has its own credit risk, i.e. bankruptcy probability.

Received : July 27, 2011 Revision : August 3, 2011 Accepted : August 10, 2011
Type of Submission : English Fast-track Corresponding author : Taeho Hong

1. Introduction

Corporate credit rating analysis has attracted lots of research interests in the literature since Altman(1968) introduced discriminant method to predict the bankruptcy with financial data. Credit is the major component of a financial institute's activities which is characterized by high risk. In addition, corporate credit rating requires the most rapid analysis while very detailed and careful analyses of credit application are performed simultaneously. Thus, both many studies and financial institutes have focused on constructing automatic corporate crediting rating systems using data mining techniques, which are divided into the statistical and artificial intelligence (AI) techniques in general.

Corporate credit rating analysis concerns two research disciplines which are namely bankruptcy prediction and corporate bond rating. Among the application of data mining techniques, the prediction of corporate failure using past financial data is also a well-known topic. Early studies of bankruptcy prediction used statistical techniques such as multiple discriminant analysis, logit, and probit. As new alternatives of statistical techniques for bankruptcy prediction, AI techniques were introduced in early 1990. Various AI techniques such as ID3, genetic algorithms, and artificial neural networks (ANNs) have been applied to the bankruptcy prediction problem. However, the automatic systems using data mining methods give only the success or failure of business in future although the decision makers, in charge of credit rating in financial institute, are forced to answer to what is the rating for a credit application. In corporate bond rating studies, which could give them the answers, the data mining methods including statistical and artificial techniques also have played an important role of the prediction of bond ratings in future. However, bond ratings which are generally regarded as credit ratings can be only obtained when the bond rating agency such as Moody's and Standard and Poor's disclosure them by the help of human raters' evaluation. However, the model building of the credit ratings is more complex than that of a bankruptcy prediction and the prediction of the credit rating is a very difficult work because the credit ratings are basically a multiclass prediction problem.

The purpose of this paper is to apply AdaBoost algorithm-based support vector machines (SVMs) into a bankruptcy prediction as a binary classification problem for the IT companies in Korea and then perform the multi-class credit ratings of the companies by making a normal distribution shape of posterior bankruptcy probabilities from the loss functions extracted from the SVMs. Many studies have shown that AI methods such as ANNs achieved better performance than traditional statistical methods. Recently, SVMs are being recognized as a competitive tool as compared with other data mining techniques for solving pattern recognition or classification decision problems. Furthermore, many researches, in particular, have proved them more powerful than traditional ANNs (Amendolia et al., 2003; Huang et al., 2004, Huang et al., 2005; Tay and Cao, 2001; Min and Lee, 2005; Shin et al., 2005; Kim, 2003).

The classification decision, such as a binary or multi-class decision problem, solved by the classifier, i.e. data mining techniques is cost-sensitive for their results. The cost which results from misclassification is critical to a decision maker if the data mining techniques classified company's credit rating as sound although it is not in fact. Therefore, it is necessary to convert the outputs of the classifier into well-calibrated posterior probabilities-based multiclass credit ratings according to the bankruptcy probabilities. Zadrozny and Elkan (2001) addressed this problem and introduced new methods for estimating the probabilities from naïve Bayes, decision tree classifiers. However, SVMs basically did not provide such probabilities although it has shown an outstanding performance in dichotomous decision. So it is required to use any method to create probabilities (Platt, 1999; Drish, 2001). Sollich (2002) addressed this issue as the estimation of class probabilities for the predictions of a trained SVM classifier. In this study, our proposed method is used to give decision makers the bankruptcy probability from the predictions of a trained SVM classifier.

IT companies analyzed in this study are the sector which is most attractive to investor, government and business decision makers since Korea overcame the economic crisis of 1997. Nowadays technology has become not only an important dimension to the development of the IT companies, but it is also essential for survival. Investors want to know which companies have a competitive edge without the analyses of technology. Similarly, the decision makers who are in charge of lending section in financial institute need to identify which credit ratings are assigned to the companies. Therefore, this study presents a method to estimate the bankruptcy probabilities from outputs of SVMs for bankruptcy prediction and then suggests credit rating methods based on the probabilities for bank's loan decision making.

The remainder of this paper is organized as follows. The following section provides a literature review related to classification for corporate credit rating analysis. The third section describes a brief overview of SVM and AdaBoost algorithm. The fourth section reports the model development and the results of experiments. The conclusions are presented in final section.

2. Classification Techniques for Corporate Credit Rating Analysis

Corporate credit rating analysis concerns two research disciplines which are namely corporate bond rating and bankruptcy prediction. Bond ratings characterize the risk of the investments and affect the costs of borrowing for the issuer. Bonds are rated by independent rating agencies such as Moody's and Standard and Poor's. Since market yields correspond to bond ratings, indicating an association between rating and risk, the study of the rating process is of interest not only to bond issuers but also to investors.

Numerous bond rating studies have traditionally used statistical techniques such as multiple discriminant analysis (Baran, Lakonishok and Ofer, 1980; Belkaoui, 1980; Pinches and Mingo, 1975), regression (Horrigan, 1966; Pogue and Soldofsky, 1969; West, 1970), probit (Kaplan and Urwitz, 1979) and logit (Ederington, 1985) models. In 1990s a number of studies have demonstrated that AI approaches such as neural networks (Kwon et al., 1997; Maher and Sen, 1997; Singleton and Surkan, 1995), rule-based system (Kim and Lee, 1995) and case-based reasoning (Buta, 1994; Shin et al., 1997) can be alternative methodologies for business classification problems. However, bond rating studies have the following limitation. That is, the target variables used for the study are ratings by credit analysts of bond rating agency. This means that, if those historical ratings that human raters evaluated do not correspond to the correct credit level of the company, the system also cannot provide correct credit information to the users.

Prediction of corporate failure using past financial data is also a well-documented topic. Early studies of bankruptcy prediction used statistical techniques such as multiple discriminant analysis (Altman, 1968), logit (Olson, 1980), and probit (Zmijewski, 1984). Recently, however, numerous studies have demonstrated that AI can be an alternative methodology for classification problems to which traditional statistical method have long been applied. Various AI techniques are applied to the bankruptcy prediction problems such as ID3 (Chung and Tam, 1992; Messier and Hansen, 1988; Park, 1993), genetic algorithms (Kingdom and Feldman, 1995) and neural networks (Barniv et al., 1997; Bell, 1997; Boritz and Kennedy, 1995; Chung and Tam, 1992; Etheridge and Sriram, 1997; Chen et al., 1995; Fletcher and Goss, 1993; Jo et al., 1997; Markham and Ragsdale, 1995; Odom and Sharda, 1990; Salchenberger et al., 1992; Tam and Kiang, 1992; Wilson and Sharda, 1994). Although bankruptcy prediction studies demonstrate that one technique outperforms the others for a given data set, it is hard to tell a priori which of these techniques will be the most effective in solving a specific classification problem. Thus, we may try several different techniques and select one that seems to provide the most accurate results for the specific problem. Alternatively, it has been suggested that the better approach to classification problem might be to integrate several different forecasting techniques by combining their results (Batchelor and Dua, 1995; Markham and Ragsdale, 1995).

Recently, SVMs are fast replacing neural networks as for solving pattern recognition problems (Amendolia et al., 2003). Some of these studies investigated the effectiveness of SVMs for the corporate bankruptcy predictions (Fan and Palaniswami, 2000; Härdle et al., 2003; Min and Lee, 2005; Shin et al., 2005) and credit rating (Huang et al., 2004). These studies showed that SVMs were competitive and outperformed other classifiers including ANNs and linear discriminant classifier in terms of generalization performance.

3. AdaBoost Algorithm-based Support Vector Machine

3.1 Support Vector Machine

The Support Vector Machine (SVM) is a popular technique for solving data classification problems. We employed SVM to estimate the credit ratings of IT companies. The SVM method was developed by Vapnik (1995). SVM, one of many machine learning techniques, is based on statistical theory. It has shown good performance and a generalizing capacity in classification tasks. It is applied to the many areas of business.

SVM is the algorithm that finds the maximum margin hyperplane, which is the maximum separation between classes. In here, support vectors are the closest to the maximum margin hyperplane. If it is impossible to divide into two classes, we can use the kernel function. In the case of nonlinear class boundaries, we can transform the inputs into a high-dimensional feature space. This is the original input space and is mapped into a high-dimensional dot-product space.

In the separating case, we can presume the function $f : R^n \rightarrow \{\pm 1\}$ using a training set. In the separated two classes, A is defined as $x_i \in R^n$, $y_i = 1$, B is defined as $x_i \in R^n$, $y_i = -1$. If it is possible to separate them linearly, they can be represented in equations (1) and (2).

$$w \cdot x_i + b \geq +1, \ \forall x_i \in A \tag{1}$$

$$w \cdot x_i + b \leq -1, \ \forall x_i \in B \tag{2}$$

Where x is the input vector, w is the weight vector and b is bias. w and b represent the parameters used to determine the hyperplane.

Using equations (1) and (2), we can derive equation (3) as follows :

$$y_i(w \cdot x_i + b) \geq 1, \ \forall x_i \in A \cup B \tag{3}$$

The maximum margin classifier optimizes data within the maximum margin hyperplane. This is an optimization problem expressed as equation (4) :

$$\min_{w,\ b} \frac{w \cdot w}{2} \quad s.t.\ y_i(w \cdot x_i + b) \geq 1 \tag{4}$$

Finally, the equation for an optimal separating hyperplane is shown in equation (5).

$$f(x,\ \alpha_i,\ b) = \sum y_i \alpha_i (x \cdot x_i) + b \tag{5}$$

Where α_i and b are parameters for determining the separation of the hyperplane. x is the training data, and x_i is the support vector.

In the case of nonlinear class boundaries, we can implement the idea by transforming the inputs into the high-dimensional feature space. A nonlinear separating case is represented in equation (6).

$$f(x,\ \alpha_i,\ b) = \sum y_i \alpha_i K(x,\ x_i) + b \qquad (6)$$

Where $K(x,\ x_i)$ is called the kernel function. The examples of the kernel functions are the polynomial kernel, $K(x,\ x_i) = (x \cdot x_i + 1)^d$, and the Gaussian radial basis function $K(x,\ x_i) = \exp(-\frac{1}{\delta^2}(x - x_i)^2)$.

3.2 The AdaBoost Algorithm

SVMs basically did not provide such probabilities although it has shown an outstanding performance in dichotomous decision. So it is required to use any method to create probabilities (Platt, 1999; Drish, 2001). SVMs can be viewed as binary margin-based learning algorithm which seeks to achieve small empirical risk for the loss function. The margin of a training example is a number that is positive if and only if the example is correctly classified by a given classifier and whose magnitude is a measure of confidence in the prediction. Several well known algorithms work directly with margins. Furthermore, there are many algorithms that attempt to minimize some loss function of the margin. AdaBoost (Allwein et al., 2001; Freund and Schapire, 1997; Schapire and Singer, 1999) is one example: it can be shown that AdaBoost, called also as adaptive boosting algorithm, is a greedy procedure for minimizing an exponential loss function of the margins. Its function plays an important role in estimating classification probability.

The algorithm AdaBoost builds a hypothesis *f* that is a linear combination of *weak* or *base hypotheses* h_t :

$$f(x) = \sum_i \alpha_i h_t(x) \qquad (7)$$

The hypothesis *f* is built up in a series of rounds on each of which an h_t is selected by *a weak or base learning algorithm* and $\alpha_t \in$ R is then chosen. It has been observed by Breiman (1997a, 1997b) and other authors (Collins et al., 2000; Friedman et al., 2000; Mason et al., 1999; Schapire and Singer, 1999) that the h_t's and α_t's are effectively being greedily chosen so as to minimize equation (8).

$$\frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} \qquad (8)$$

Thus, AdaBoost is a binary margin based learning algorithm in which the loss function is L(z) = e^{-z}. z in the function is a margin of SVMs and is also expressed as $y_i f(x_i)$. For example, suppose $(x_1,\ y_1)$, ⋯, $(x_m,\ y_m)$ as input binary labeled training examples. The instances x_i belong to some domain X and the labels $y_i \in \{-1,\ +1\}$. The margin of an example (x, y) applied to the SVM model is $yf(x)$. The margin is positive if and only if the sign of f(x) agrees with y.

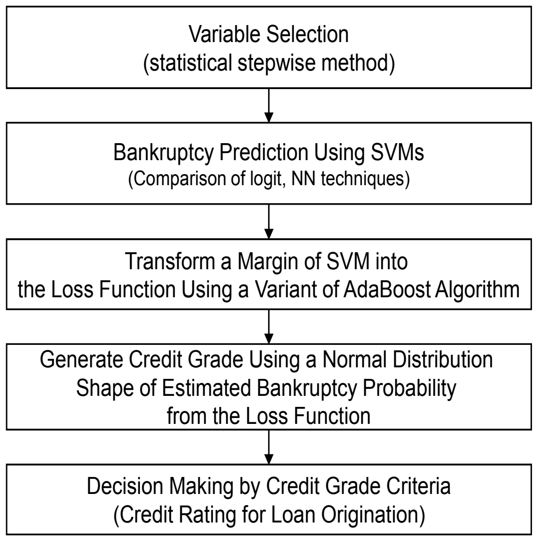
This study uses a variant of AdaBoost (Allwein et al., 2001) to output predictions in which the predicted label of a new sample x is chosen randomly to be +1 with $1/(1+e^{-2f(x)})$. The loss suffered then is the probability that the randomly chosen predicted label disagrees with the correct label y. Let $p(x)=1/\left(1+e^{-2f(x)}\right)$. Then the loss is p(x) if y = -1 and 1-p(x) if y = +1. Using a simple algebraic manipulation, the loss can be shown to be $1/\left(1+e^{2yf(x)}\right)$. So for this variant of AdaBoost, we use $L(z)=1/(1+e^{2z})$ set by Allewein et al. (2001). $L(z)$ is a measure of confidence for a given classification. With such a measure, we can reject classifications with high error probability and make our predictions more reliable.

4. Proposed Model and Experimental Results for Credit Rating Using a Variant of SVMs

4.1 Proposed Model

In this study, we show a multi-class credit rating approach by generating estimated bankruptcy probability from an AdaBoost algorithm-based SVM. This approach is able to gain the credit ratings for the evaluation of the corporate loan origination. The bankruptcy prediction model basically gives information such as good or bad credit. On the other side, the bankruptcy probability could be utilized to assess the corporate credit ratings.

<Figure 1> shows the procedure of our proposed method. The procedure is summarized as follows. First, we selected the variables using logit model with stepwise method. Second, after the feature selection, the data mining techniques for bankruptcy prediction were applied with real data.



<Figure 1> The Procedure of Proposed Method for Credit Rating Using Estimated Probability From AdaBoost Algorithm-based SVM

To validate the performance of SVM, we applied the data to SVM, NN, logit model through 5-fold cross validation. We looked into the performance of SVM comparing to logit and NN by the 5-fold cross validation. Third, a margin of SVM is transformed into the loss function using a variant of AdaBoost (Allwein et al., 2001) because SVMs basically did not provide such probabilities although it has shown an outstanding performance in dichotomous decision. Fourth, credit grades are generated using a normal distribution shape of estimated bankruptcy probability from the loss function. Finally, the credit rating for the cases applied in the study, is assigned using the credit grade criteria. The credit ratings obtained here are very useful for the decision making of loan evaluation in the financial institutions.

4.2 Data Summary

This study focuses on corporate credit analysis for IT-based industry in Korea. IT-based industry includes the followings : (1) manufacture of computers and office machinery, (2) manufacture of electronic components, radio, television and communication equipment and apparatuses, (3) computer and related activities (computer system design and consultancy; software consultancy and supply; data processing and computer facilities management services; data base activities and on-line information provision services; other computer activities).

The companies related to IT have some characteristics as follows. First, there is no considering the size and the type of business of firms. For example, the sales amount is diversified among firms. Also, various type of companies such as manufacturing, software, service etc. are mixed in our sample. So it is not easy to make a model with a traditional data mining technique. Second, the companies in IT-based industry have a strong risk in the market and fierce competition. A company which has good condition in financial ratio suddenly goes bankrupt.

Our proposed method was applied to 209 Korean firms in IT-based industry. The sample consisted of 209 companies and 74 companies in our sample were bankrupt cases and 135 companies are non-bankrupt cases. For both the bankrupt and the non-bankrupt firms, data were collected during the period from 1999 to 2001. From the financial statement, we calculated 60 financial variables.

4.3 Variable Selection

We utilized a logit model to select variables among 60 variables with the above data. By the stepwise method, we obtained 5 variables, i.e. sales coefficient of variation, financial expenses to liabilities, cash ratio, Liabilities to EBITDA, and Payables turnover. The input variables are described in <Table 1>. With the selected variables, we applied logit, neural network, and SVM to 209 companies. To acquire a more reliable approach, we used k-fold cross-validation method with k=5. For our sample, 209 test cases, 5 repetitions are used. In each repetition 167 cases are used for the training set and 42 cases as a holdout set for testing. The 5^{th} fold consists of 41 cases for the testing set in order to cover the whole sample because 167 is not 5 times in correct. Holdout sets are selected so that their union over all repetition presents the entire training set.

<Table 1> Selected Variables of Logit Model

Variables Name	B[1]	S.E.[2]	Wald[3]	P-value[4]
Sales coefficient of variation	−0.033	0.008	18.900	0.000
Financial expenses to liabilities	−0.279	0.077	13.152	0.000
Cash ratio	0.046	0.013	12.105	0.001
Liabilities to EBITDA	−0.060	0.024	5.939	0.015
Payables turnover	−0.011	0.004	5.699	0.017
Constant	3.323	0.737	20.323	0.000

Note : 1) coefficient estimates, 2) standard error, 3) Wald statistics, 4) significant probability.

4.4 Experimental Results

<Table 2> shows the results of SVM using radial based function kernel. We have a lot of simulation changing the parameters C and δ^2. We compared the results of SVM model in <Table 2> with those of logit and ANN model (<Table 3>). SVM model has 74.66% as an average hit ratio, which is better than the performance of logit and neural networks model in <Table 3>. We found that the SVM classifier is superior to other data mining techniques.

Corporate credit ratings generally characterize the loan risks and affect the costs of borrowing for the loan applicants. The credit grades

<Table 2> The Results of SVM using RBF

(Unit : %)

δ^2	C	Set1		Set2		Set3		Set4		Set5	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
1	1	74.9	66.7	77.8	76.2	75.5	61.9	77.3	71.4	74.4	78.1
	25	76.7	73.8	79.0	73.8	80.8	69.1	77.8	76.2	77.4	80.5
	50	**79.6**	**71.4**	79.6	73.8	80.2	73.8	79.0	76.2	76.8	82.9
	75	79.6	69.1	80.8	73.8	80.2	73.8	79.6	78.6	77.4	80.5
	100	79.0	69.1	**82.0**	**73.8**	**80.8**	**71.4**	**80.8**	**78.6**	**79.2**	**78.1**
25	1	73.7	66.7	79.0	76.2	74.9	57.1	73.1	69.1	69.6	75.6
	25	74.9	66.7	72.5	78.6	76.7	61.9	74.9	71.4	70.8	75.6
	50	74.9	69.1	73.7	78.6	76.7	61.9	75.5	69.1	70.8	78.1
	75	75.5	69.1	73.1	76.2	75.5	61.9	74.3	71.4	73.8	80.5
	100	74.9	69.1	75.5	71.4	76.1	61.9	71.4	74.3	73.8	80.5
50	1	73.7	66.7	73.1	76.2	74.9	57.1	73.1	69.1	71.4	75.6
	25	73.1	66.7	72.5	78.6	76.1	61.9	75.5	69.1	73.8	80.5
	50	74.9	66.7	73.7	78.6	75.5	61.9	75.5	71.4	73.8	80.5
	75	78.9	66.7	73.1	76.2	76.1	61.9	76.7	71.4	74.4	78.1
	100	75.5	69.1	75.5	71.4	76.1	59.5	76.7	71.4	74.4	78.1
75	1	73.7	66.7	73.1	76.2	75.5	57.1	73.1	69.1	72.0	75.6
	25	73.1	66.7	73.1	76.2	76.1	54.8	73.7	69.1	69.6	75.6
	50	73.7	64.3	71.9	78.6	76.1	61.9	74.9	69.1	69.6	75.6
	75	74.9	66.7	72.5	78.6	76.7	61.9	75.5	69.1	70.8	75.6
	100	74.9	66.7	73.7	78.6	76.1	61.9	72.5	69.1	70.8	75.6
100	1	73.7	66.7	72.5	76.2	75.5	57.1	73.7	69.1	69.6	75.6
	25	73.7	66.7	73.1	76.2	74.9	57.1	73.7	69.1	69.6	75.6
	50	73.7	66.7	73.1	76.2	76.1	57.1	71.9	69.1	69.6	75.6
	75	73.1	64.3	72.5	78.6	76.7	61.9	75.5	69.1	69.6	75.6
	100	74.9	69.1	72.5	78.6	76.7	61.9	74.9	71.4	70.8	75.6

Note : *Optimal parameters in the set.

for credit loans are rated by the expertise of each financial institution. So the credit rating process is a complex and important task adjusted by judgmental information in each financial institution.

<Table 3> The Results of logit, ANN, and SVM

(Unit : %)

	Logit		ANN		SVM	
	Train	Test	Train	Test	Train	Test
Set1	75.45	76.19	76.05	71.43	79.64	71.43
Set2	73.05	73.80	80.24	71.43	82.04	73.81
Set3	78.44	64.29	80.24	69.05	80.84	71.43
Set4	74.85	71.43	78.44	78.57	80.74	78.57
Set5	75.00	80.49	72.62	80.49	79.17	78.05
Avg.	75.36	73.24	77.52	74.19	80.49	74.66

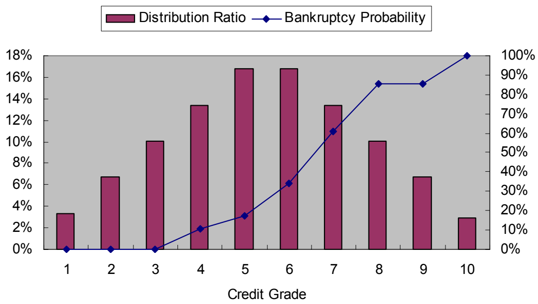
However, the loss function of the AdaBoost algorithm-based SVM model proposed in this study transforms the margin of the SVM into the probability that randomly chosen predicted group disagrees with the correct group (bankrupt or non-bankrupt) for each company. We generated a normal distribution shape of estimated bankruptcy probability using the loss function and then classified 10-grade in credit ratings according to the distribution and estimated the bankruptcy probabilities. This procedure is more effective in minimizing the misclassification problems on condition that each credit grade for its credit loan borrowers has its own credit risk, i.e. bankruptcy probability.

<Table 4> and <Figure 2> show one example of the credit grade ratings for 207 companies in IT-based Industry. For example, the grade 1~3 shows the lowest bankruptcy probabilities(0%) but the grade 10 has the highest bankruptcy probability(100%). Therefore, the higher the grade number becomes, the higher the bankruptcy probability becomes. The number of credit grades in our study is fixed as 10. However, the number is flexibly changeable according to the corporate credit rating characteristics. Actually, in <Table 4>, credit grades 1~3 shows the same bankruptcy probabilities(0%) even though the distribution ra-

<Table 4> Credit Ratings Using Estimated Probability to Bankruptcy from SVM

Credit Grade	Bankrupt	Non-Bankrupt	Total	Bankruptcy Probability	Distribution Ratio
1	0	7	7	0.00%	3.35%
2	0	14	14	0.00%	6.70%
3	0	21	21	0.00%	10.05%
4	3	25	28	10.71%	13.40%
5	6	29	35	17.14%	16.75%
6	12	23	35	34.29%	16.75%
7	17	11	28	60.71%	13.40%
8	18	3	21	85.71%	10.05%
9	12	2	14	85.71%	6.70%
10	6		6	100.00%	2.87%
Total	74	135	209	35.41%	100.00%

tio is reasonable in terms of a normal distribution shape. It is difficult to distinguish among these grades(1~3) according to bankruptcy probability. One of the reasons is partially due to the small sample size. So these grades can be combined as the same grade in this study. However, if the sample size is larger, it will be more possible that the grades(1~3) have different bankruptcy probabilities and their own credit characteristics.



<Figure 2> The Distribution Ratio of Credit Grade and the Bankruptcy Probability of Firms

In general, there are two errors in a generic bankruptcy prediction model. One is a false positive error, which occurs when good company is classified as bad or bankrupt. The other is a false negative error, which occurs when bad company is classified as good. In IT-based industry, the companies have the tendency of the sudden bankruptcy. Therefore, the false negative error is always larger than the false positive error on IT-based industry. Finally our results have an implication that our proposed method using a variant of SVM is very useful for credit rating in IT-based industry by adjusting estimated bankruptcy probability. That is, our approach can minimize the misclassification costs, in particular, related with the false negative error by adjusting the credit grade interval ranges.

5. Conclusions

This study showed a method for multi-class credit ratings by generating estimated bankruptcy probability from an AdaBoost algorithm-based SVM. This method is briefly summarized as follows. First, we selected the variables using logit model with stepwise method. After feature selection, the bankruptcy prediction techniques were applied with real data. To validate the performance of SVM, we applied our empirical cases to SVM, ANN, and logit model through 5-fold cross validation. We compared the performance of SVM with that of logit and NN by the 5-fold cross validation. Then, a margin of SVM was transformed into the loss function using a variant of AdaBoost and credit grades were generated using a normal distribution shape of estimated bankruptcy probability from the loss function. Finally, the credit grades for the cases applied in the study were assigned using the credit grade criteria.

The main benefit of our credit rating approach is to show more detailed explanatory powers by transforming a binary bankruptcy prediction problem into multi-class credit rating analysis. This method can also minimize the misclassification problems by adjusting the credit grade interval ranges on condition that each credit grade for its credit loan borrowers has its own credit risk, i.e. bankruptcy probability. However, this study has a few limitations. First, a few grades among all the credit grades partially have the same bankruptcy probabilities. So we need to classify more sophisticated credit grade interval ranges and find its criteria. Second, the sample size of this study is a little small. So our method has a limitation in terms of model generalization. Finally in the future research, the asymmetric error costs in the credit rating will be able to be considered in more detail.

References

Allwein, E. K., R. E. Schapire and Y. Singer, "Reducing multiclass to binary : a unifying approach for margin classifiers", *Journal of Machine Learning Research*, Vol.1(2001), 113~141.

Altman, E., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *The Journal of Finance*, Vol.23 (1968), 589~609.

Amendolia, S. R., G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala, and G. M. Mura, "A Comparative Study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia Screening", *Chemometrics and Intelligent Laboratory Systems*, Vol.69(2003), 13~20.

Baran, A., J. Lakonishok, and A. R. Ofer, "The Value of General Price Level Adjusted Data to Bond Rating", *Journal of Business Finance and Accounting*, Vol.7, No.1(1980), 135~149.

Barniv, R., A. Agarwal, and R. Leach, "Predicting the Outcome Following Bankruptcy Filing : A Three-state Classification Using Neural Networks", *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 177~194.

Batchelor, R. and P. Dua, "Forecast Diversity and the Benefits of Combining Forecasts", *Management Sciences*, Vol.41(1995), 68~75.

Belkaoui, A., "Industrial Bond Ratings : A New Look", *Financial Management*, Vol.9, No.3 (1980), 44~51.

Bell, T., "Neural Nets or the Logit Model? A Comparison of Each Model's Ability to Predict Commercial Bank Failures", *Intelligent Systems in Accounting*, *Finance and Management*, Vol.6(1997), 249~264.

Boritz, J. E. and Kennedy, D. B., "Effectiveness of Neural Network Types for Prediction of Business Failure", *Expert Systems with Applications*, Vol.9, No.4(1995), 503~512.

Breiman, L., "Arcing the edge. Tech. rep. 486, Statistics Department", University of California at Berkeley, 1997a.

Breiman, L., "Prediction games and arcing classifiers, Tech, rep. 504, Statistics Department", University of California at Berkeley, 1997b.

Buta, P., "Mining for Financial Knowledge with CBR", *AI Expert*, Vol.9, No.2(1994), 34~41.

Chen, S., P. Mangiameli, and D. West, "The Comparative Ability of Self-organizing Neural Networks to Define Cluster Structure", *Omega,* Vol.23, No.3(1995), 271~279.

Chung, H. and K. Tam, "A Comparative Analysis of Inductive Learning Algorithm", *Intelligent Systems in Accounting*, *Finance and Management*, Vol.2(1992), 3~18.

Collins, M., R. E. Schapire, and Y. Singer, "Logistic regression, AdaBoost and Bregman distances", *Proceedings of the Thirteenth Annual Conference on Computational Learn-*

ing Theory, 2000.

Drish. J., "Obtaining Calibrated Probability Estimates from Support Vector Machines", *Seminar on Learning Algorithms*, San Diego, Spring, 2001.

Ederington, H. L., "Classification Models and Bond Ratings", *Financial Review*, Vol.20, No.4(1985), 237~262.

Etheridge, H. and R. Sriram, "A Comparison of the Relative Costs of Financial Distress Models : Artificial Neural Networks, Logit and Multivariate Discriminant Analysis", *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 235~248.

Fan, A. and M. Palaniswami, "Selecting Bankruptcy Predictors Using a Support Vector Machine Approach", in *Proceedings of the International Joint Conference on Neural Networks*, 2000.

Fan, A. and M. Palaniswami, "A New Approach to Corporate Loan Default Prediction from Financial Statements", *Proceedings of the computational finance/forecasting financial markets conference*, London (CD), UK, 2000.

Fletcher, D. and E. Goss, "Forecasting with Neural Networks : An Application Using Bankruptcy Data", *Information and Management*, Vol.24, No.3(1993), 159~167.

Freund, Y. and R. E. Schapire, "A decision theoretic generalization of online learning and an application to boosting", *Journal of Computer and System Sciences*, Vol.55, No.1(1997), 119~139.

Friedman, J., T. Hastie, and R. Tibshirani, "Additive logistic regression : a statistical view of boosting", *The Annals of Statistics*, Vol.38, No.2(2000), 337~374.

Härdle, W., R. A. Moro, and D. Schäfer, "Predicting corporate bankruptcy with support vector machines", *Working Slide*, Humboldt University and the German Institute for Economic Research available on, 2003.

Horrigan, J. O., "The Determination of Long Term Credit Standing with Financial Ratios", *Journal of Accounting Research*, Supplement, (1966), 44~62.

Huang, W., Y. Nakamori, and S. Wang, "Forecasting Stock Market Movement Direction with Support Vector Machine", *Computers and Operations Research*, Vol.32(2005), 2513~2522.

Huang, Z., H. Chen, C. Hsu, W. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks : a market comparative study", *Decision Support Systems*, Vol.37, No.4(September 2004), 543~558.

Jo, H., I. Han, and H. Lee, "Bankruptcy Prediction Using Case-based Reasoning, Neural Networks, and Discriminant Analysis, *Expert Systems with Applications*, Vol.13, No.2 (1997), 97~108.

Kaplan, R. S. and G. Urwitz, "Statistical Models of Bond Ratings : A Methodological Inquiry", *Journal of Business*, Vol.52, No.2(1979), 231~262.

Kim, B. O. and S. M. Lee, "A Bond Rating Expert System for Industrial Companies", *Expert Systems with Applications*, Vol.9, No.1(1995), 63~70.

Kim, K. J., "Financial Time Series Forecasting Using Support Vector Machines", *Neurocomputing,* Vol.55, No.1-2(2003), 307~319.

Kingdom, J. and K. Feldman, "Genetic Algorithms for Bankruptcy Prediction", *Search Space Research Report*, No.01-95, Search Space Ltd. London, 1995.

Kwon, Y. S., I. Han, and K. C. Lee, "Ordinal Pairwise Partitioning (OPP) Approach to Neural Networks Training in Bond Rating", *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 23~40.

Maher, J. J. and T. K. Sen, "Predicting Bond Ratings Using Neural Networks : A Comparison with Logistic Regression", *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 59~72.

Markham, I. and C. Ragsdale, "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis", *Decision Sciences*, Vol.26, No.2 (1995), 229~242.

Mason, L., J. Baxter, P. Bartlett, and M. Frean, "Functional gradient techniques for combining hypotheses. In Smola, A. J., Bartlett, P. J., Sch¨olkopf, B., and Schuurmans, D. (Eds.)", *Advances in Large Margin Classifiers*, MIT Press, 1999.

Messier, W. and J. Hansen, "Inducing Rules for Expert System Development : and Example Using Default and Bankruptcy Data", *Management Science*, Vol.34, No.12(1988), 1403 ~1415.

Min, J. H. and Y. Lee, "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters", *Expert Systems with Applications*, Vol.28 (2005), 603~614.

Odom, M. and R. Sharda, "A Neural Networks Model for Bankruptcy Prediction", *Proceedings of the IEEE International Conference in Neural Networks*, San Diego, CA., Vol.Ⅱ (1990), 163~168.

Ohlson, J., "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, Vol.18, No.1(1980), 109~131.

Park, J., "Bankruptcy Prediction of Banks and Insurance Companies : An Approach Using Inductive Methods", *Ph.D. Dissertation*, University of Texas at Austin, 1993.

Pinches, G. E. and K. A. Mingo, "A Multivariate Analysis of Industrial Bond Ratings", *Journal of Finance*, Vol.28, No.1(1973), 1~18.

Platt, J., "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", In *Advances in Large Margin Classifiers*, 1999.

Pogue, T. F. and R. M. Soldofsky, "What's in a Bond Rating?", *Journal of Financial and Quantitative Analysis*, Vol.4, No.2(1969), 201~228.

Salchenberger, L., E. Cinar, and N. Lash, "Neural Networks : A New Tool for Predicting Thrift Failures", *Decision Sciences*, Vol.23(1992), 899~916.

Schapire, R. E. and Y. Singer, "Improved boosting algorithms using confidence-rated predictions", *Machine Learning*, Vol.37, No.3 (1999), 297~336.

Shin, K. S., T. S. Shin, and I. Han, "Using Induction Technique to Support Case-based Reasoning : A Case of Corporate Bond Rating", *Proceedings of MS/OR Society Conference*, Seoul, Korea, (1997), 199~202.

Shin, K., T. S. Lee, and H. Kim, "An Application of Support Vector Machines in Bankruptcy Prediction Model", *Expert Systems with Applications*, Vol.28, No.1(January 2005), 127~135.

Singleton, J. C. and A. J. Surkan, "Bond Rating with Neural Networks", In Refenes, A. ed., *Neural Networks in the Capital Markets*, John Wiley, Chichester, 1995.

Sollich, P., "Bayesian Methods for Support Vector Machines : Evidence and Predictive Class Probabilities", *Machine Learning*, Vol.46, No.1-3(2002), 21~52.

Tam, K. and M. Kiang, "Managerial Applications of Neural Networks : The Case of Bank Failure Predictions", *Management Science*, Vol.38, No.7(1992), 926~947.

Tay, F. E. H. and L. Cao, "Application of Support Vector Machine in Financial Time Series Forecasting", *Omega*, Vol.29, No.4(2001), 309~317.

Vapnik, V., *The Nature of Statistical Learning Theory*, New York, Springer-Verlag, 1995.

West, R. R., "An Alternative Approach to Predicting Corporate Bond Ratings", *Journal of Accounting Research*, Vol.8, No.1(1970), 118~125.

Wilson, R. and R. Sharda, "Bankruptcy Prediction Using Neural Networks", *Decision Support Systems*, Vol.11, No.5(1994), 545~557.

Zadrozny, B. and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers", in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

Zmijewski, M. E., "Methodological Issues Related to the Estimated of Financial Distress Prediction Models", *Journal of Accounting Research*, Vol.22, No.1(1984), 59~82.

Abstract

AdaBoost 알고리즘기반 SVM을 이용한 부실 확률분포 기반의 기업신용평가

신택수* · 홍태호**

최근 몇 년간 SVM(support vector machines)기법은 패턴인식 또는 분류의사결정문제를 위한 분석기법으로서 기존의 데이터마이닝 기법과 비교할 때, 매우 높은 성과를 갖는 것으로 인식되어 왔다. 더 나아나 많은 연구자들은 SVM기법이 1980년대 이후 대표적인 예측 및 분류모형으로 인정받은 인공신경망기법(ANNs : Artificial Neural Networks)에 비해 더 성과가 좋다는 사실을 실증적으로 입증해 왔다(Amendolia et al., 2003; Huang et al., 2004, Huang et al., 2005; Tay and Cao, 2001; Min and Lee, 2005; Shin et al., 2005; Kim, 2003).

일반적으로 이와 같이 다양한 데이터마이닝 기법에 의해 분석되는 이진분류 또는 다분류 의사결정문제들은 특히 금융분야 등에 있어서 오분류비용에 민감하며, 이로 인한 오분류의 경제적 손실도 상대적으로 매우 크다고 할 수 있다. 따라서 기업부도예측모형과 같은 이진분류모형의 결과값을, 부도확률에 기초하여 정교하게 계산된 사후확률의 개념으로서 다분류의 신용등급평가의 문제로 변환할 필요가 있다. 그러나, SVM 모형의 결과값은 기본적으로 그와 같은 부도확률분포를 보여주지 않는다. 따라서, 그러한 확률분포를 정교하게 보여줄 방법을 제시할 필요가 있다(Platt, 1999; Drish, 2001).

본 연구는 AdaBoost 알고리즘기반의 SVM 모형을 이용하여, 이진분류모형으로서 IT 기업의 부실예측모형에 적용한 후, 이 SVM 모형의 예측결과를 SVM의 손실함수에 적용하여 계산된 값을 사후부도확률의 정규분포 특성에 따라 이를 구간화하여 IT기업에 대한 다분류 신용등급 평가의 문제로 전환시키는 방법을 제시하였다. 그리고 본 연구에서 제안하는 방법은 이러한 AdaBoost 알고리즘기반 SVM 모형이 각 기업이 고유한 신용위험(부도확률)을 갖고 있다는 조건하에서, 신용등급부여를 위한 부도확률분포 구간을 정교하게 조정함으로써 오분류 문제를 좀 더 줄일 수 있음을 제시하였다.

Keywords : 기업신용평가, Support Vector Machine, AdaBoost 알고리즘, 부도확률, IT기업

* 연세대학교 정경대학 경영학부
** 부산대학교 경영대학 경영학과

저 자 소 개

신택수
현재 연세대학교 정경대학 경영학부 부교수로 재직하고 있다. 연세대학교 경영학과에서 학사 및 석사학위를 받고, 한국과학기술원에서 경영정보시스템으로 경영공학 박사학위를 받았다. 주요 관심분야는 데이터마이닝, 전략적 성과관리, 지식경영, 고객관계관리 등이다.

홍태호
현재 부산대학교 경영대학 부교수로 재직하고 있다. KAIST에서 산업공학사를 취득하였고 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 딜로이트 컨설팅에서 컨설턴트로 재직했으며, 주요 관심분야는 데이터마이닝, CRM, 지식경영, Social Networks 등이다