Prediction of bankruptcy on industry classification

Masanobu MATSUMARU\textsuperscript{a}, Takaaki KAWANAKA\textsuperscript{b}, Hideki KATAGIRI\textsuperscript{a}, Shoichi KANEKO\textsuperscript{c}

\textsuperscript{a}Department of Industrial Engineering and Management, Faculty of Engineering, Kanagawa University, 3-27-1 Rokkakubashi, Kanagawa-ku, Yokohama, Kanagawa, 221-8686 Japan, 
\textsuperscript{b}Institute for Innovation in International Engineering Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan, 
\textsuperscript{c}Faculty of Management Information, Yamanashi Gakuin University, 2-4-5 Sakaori, Kofu city, Yamanashi, 400-8575 Japan

Abstract

This study predicts bankruptcy by making predictions after separating the entire industry into individual industries via the decision tree method and support vector machines (SVMs), and then it compares two methods, decision tree and SVM. This research predicts bankruptcy in two steps. The first step is to predict using decision tree method and the second step is to predict using support vector machine for 11 industries. The financial statements of the listed companies in the Tokyo Stock Exchange, the Osaka Securities Exchange, and other stock exchanges were used as data in this study. The data of 244 bankrupt companies that went bankrupt between 1991 and 2014 are used. On the other hand, data of 64708 non-bankrupt companies that did not go bankrupt between 1991 and 2014 for 24 years are used. The data is acquired from Nikkei NEEDS database. In the decision tree, the multi-step branching process is stratified, and the bankruptcy prediction is performed in the tree diagram. In SVM, prediction of bankruptcy is almost perfectly conducted to discriminate the companies for each industry.

Key words: bankruptcy, decision tree, support vector machines (SVMs)

1. Introduction

Japanese companies are improving their corporate performance and financial structure because of the yen depreciation and gradual recovery of the economy in recent years. According to Teikoku Databank, Ltd., the number of bankruptcies in 2016 was 8164, less than the previous year for the seventh consecutive time. Though the number of bankruptcies has decreased, even today it is a highly important management subject to determine whether a company will become bankrupt. The causes of bankruptcy are diverse, and they can be qualitative and quantitative factors. In this research, financial indicators are the quantitative factors considered to explain the reason for bankruptcy in the industry. The significance of this research is that clients, investors, and financiers can prevent loss before finalizing a deal based on the information of financial indicators from bankruptcy prediction. On the other hand, managers can improve their business by paying attention to these financial indicators. This study predicts bankruptcy by making predictions after separating the entire industry into individual industries via the decision tree method and support vector machines (SVMs), and then compares the methods. This research makes the predictions in two steps.

The first step predictions in two steps. The first step predicts using the decision tree method, and the second step predicts by applying the SVM method on eleven industries. A bankruptcy discrimination research was first conducted by Beaver\cite{1}.

Altman\cite{2} classified companies into groups by using multivariate discriminant analysis (MDA) to distinguish between bankrupt and non-bankrupt companies, and further developed the research by Beaver. Logistic regression was first employed by Ohlson\cite{3}. Subsequently, studies of bankruptcy discrimination have been using statistical analysis such as MDA, multiple regression equation, logistic model, and probit model. However, such statistical methods have some restrictions and preconditions under which the input variables are assumed to exhibit linearity, normality, and independence. Deakin\cite{4} pointed out that most of the assumptions are not valid because the financial data are independent variables. Therefore, it can be inferred that statistical methods have limitations in achieving effectiveness and validity. In recent years, techniques for overcoming such limits have been developed, and studies incorporating artificial intelligence, such as k-nearest neighbors, decision trees, neural networks\cite{5}\cite{6}\cite{7}\cite{8}\cite{9}, genetic algorithm\cite{10}\cite{11}, and SVMs\cite{12}\cite{13}\cite{14}\cite{15}, are

\*Corresponding author: mxk00065@nifty.com

Received: September 21, 2017
Accepted: October 26, 2018
increasing. Decision trees and SVMs are used in this study as the methods for predicting bankruptcy. In a decision tree, the multilevel branch processing is hierarchized, and the bankruptcy prediction is performed with a tree diagram. This research predicted the bankruptcy of all the industries by using the decision tree method without separating the entire industry into individual industries at first. However, it was impossible to predict bankruptcy accurately. Therefore, we decided to classify the entire industry into eleven industries. However, only for some industries the bankruptcy could be accurately predicted. Consequently, we considered this as the limitation of the decision tree method, and subsequently chose to use SVMs to predict bankruptcy and compared it with the decision tree method. In this study, we predict the classification of the industry by using the decision tree method and SVMs and then compare the two methods.

2. Theoretical framework

2.1 Decision tree

In a decision tree, each node is assigned some text attribute (explanatory variable) according to whose value the samples are divided into several groups. Subsequently, the child nodes are generated sequentially from each group, and a parent and child node are connected by a link. For example, at the top node (1) in Fig. 1, non-bankruptcy is used as a test attribute, short-term debt rotation \( x_2 \geq 0.64 \), and it is categorized into "yes" and "no" groups to form two child nodes and terminal nodes. The node at the top of the tree is called the root node. A node that does not have child nodes is called a leaf node or terminal node, and it indicates the final classification result or the value of the objective variable.

To be specific, 34 bankrupt companies and 680 non-bankrupt companies are separated by short-term debt rotation \( x_2 \geq 0.64 \) at the top node. 34 bankrupt companies are further separated by debt capacity ratio \( x_1 \geq 0.026 \). As a result, 34 bankrupt companies are judged to be 29

![Decision Tree Diagram]

companies as bankrupt companies, and the remaining 5 bankrupt companies are judged as non-bankrupt companies. 680 non-bankrupt companies are all judged as non-bankrupt companies. The advantage of the decision tree method is that each node represents a conditional branch, and it is possible to automatically reproduce rules that are easy to understand from the decision tree.

Based on Fig. 1, the following rules can be generated:

- IF \( (x_2 < -0.64) \) THEN (non-bankruptcy)
- IF \( (x_2 \geq -0.64) \ AND \ (x_1 < -0.026) \) THEN (non-bankruptcy)
- IF \( (x_2 \geq -0.64) \ AND \ (x_1 \geq -0.026) \) THEN (bankruptcy)

2.1.1 Division criterion

(1) Generation and purity of decision trees

When generating decision trees, we require attribute selection criteria or division evaluation criteria to decide which attribute to assign to each node. Algorithms developed in machine learning focus on the purity by splitting. Algorithms in the statistical field are considered with a statistically significant difference in the distributions between the child nodes, and decision trees are generated based on purity maximization. Generally, there are \( s \) samples (individuals) in data set \( S \) to be classified by parent node \( t \). The objective variable of this \( s \) number of specimens accepts \( m \) different values. Assuming that \( s_i \) is the number of samples belonging to class \( C_i \), \( S = \sum_{i=1}^{m} s_i \). At this point, the impurity at node \( t \) is defined based on the number of samples belonging to each class and is expressed as follows.

\[
I(t) = \emptyset(s_1, s_2, \ldots, s_m)
\]  

(2.1) where \( \emptyset(x) \) represents a function that defines the impurity. This function can select various functions according to the employed measure of impurity. In this research, we use the Gini diversity index and gain of information quantity based on entropy.

(2) Gini diversity index

The Gini diversity index indicates that in a classification problem such as data mining, the specimen is pure and cannot be divided further. If all specimens belong to one group, the Gini diversity index = 0. However, if all the samples belong to more than one group with a balanced and equal probability, it takes a value close to the Gini diversity index = -1. Therefore, it is possible to generate a good decision tree by branching the node while selecting the test attribute so that the Gini diversity index approaches 0 after starting from 1. Specifically, the Gini diversity index is defined as the sum of the variances of each class as follows:

\[
\text{Gini}(s_1, s_2, \ldots, s_n) = 1 - \sum_{i=1}^{n} \left( \frac{s_i}{S} \right)^2
\]  

(2.2)

The decision tree method is one of the tree-based models involving nonlinear regression analysis and nonlinear decision analysis. When a decision tree is used and the classification rules are represented by a tree structure, the data to be classified are the objective variables (dependent variables) and the data used to classify it are called explanatory variables (independent
variables). The decision tree consists of a node indicated by an ellipse and a link or branch represented by a line.

(3) Gain of information amount based on entropy

The entropy decreases as the order increases, and becomes large in a disordered state. Similar to the Gini diversity index, if all the specimens belong to one group, it becomes entropy $P = 0$. However, if all the specimens are equally divided into multiple groups, the entropy is 1. The entropy is expressed as follows:

$$\text{Ent}(s_1, s_2, \ldots, s_n) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2 \left( \frac{s_i}{s} \right) \quad (2.3)$$

### 2.1.2 R function in decision tree

Recalling, a decision tree is a tree-based model involving nonlinear regression and decision analyses. In the open source statistical software R, tree and rpart are the related packages of decision trees. We explain the rpart function here as it was used in this study. It is a function for generating a decision tree in the package rpart and uses the following format:

```r
rpart(formula, data, method, parms)
```

Here, the argument "formula" specifies the label of the group, such as the model expression " $y = x_1 + x_2 + \cdots $", or objective variable $y$ and explanatory variables $x_1, x_2$ etc. Specifically, the variable to be used is labelled with the symbol "+" shown as follows in the case of seven financial variables.

$$y(\text{bunrui}) = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

The argument "data" describes a data frame containing data of the objective variable and explanatory variable specified in the "formula". The argument "method" specifies the type of objective variable; "method = anova" is for general quantitative variables, and "method = class" is for qualitative variables. In this research, we use "method = class". The argument "parms = " specifies the parameters required to act as the partitioning criteria at each node. In case of "parms = list (split = gini)", the Gini diversity index is used, and for "parms = list (split = information)", the gain of the information amount based on the entropy is used as the division criterion. In this study, decision tree analysis is performed using rpart function. This is shown below in an example. The financial indicator used includes 23 indicators that are presented later in the Appendix.

```r
> (res<-rpart(bunrui~x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}+x_{11}+x_{12}+x_{13}+x_{14}+x_{15}+x_{16}+x_{17}+x_{18}+x_{19}+x_{20}+x_{21}+x_{22}+x_{23},data =X, method = "class", parms=list (split="information"), cp=cp))
```

### 2.2 Support vector machines

An SVM, presented as machine learning by Vapnik [12], is a data analysis method that mainly deals with classification and regression problems. An SVM is a high dimensional hypothesis space that can be linearly separated, and can be understood as follows as it is a method that follows a linear approach. We define the training data sets as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x = (x_1, x_2, \ldots, x_p)^T$ is the feature vector of the individual and $y$ is the objective variable that is a numerical value in the regression problem and it is the label of the class in the classification problem. The data consists of a pair of feature vectors $x_i \in \mathbb{R}^n, i = 1, 2, \ldots, l$ and class $y_i \in \{-1, 1\}$. $x_i$ is the financial indicator of company $i$, such that $y_i = -1$ represents bankruptcy and $y_i = 1$ indicates non-bankruptcy. For linear regression and linear discrimination problems, the following linear model is used:

$$y = w^T x + b \quad (2.4)$$

Positive and negative samples are separated by hyperplane $H_0: w^T x + b = 0$.

SVM determines the coefficient that maximizes the margin, and performs the discrimination as follows:

$$y = \begin{cases} 1, & \text{if } w^T x + b \geq 1 \\ -1, & \text{if } w^T x + b \leq -1 \end{cases} \quad (2.5)$$

The interval between the straight lines $w^T x + b = 1$ and $w^T x + b = -1$ is the margin; maximizing margin $M$ is a problem of maximizing $M = \min \left( \frac{1}{w^T w} \right) = \frac{2}{\|w\|^2}$.

![Fig. 2 Identification boundary of SVM](image-url)

The individual on the dotted line in Fig. 2 is called a support vector, and only a support vector affects the margin. The maximization of the margin is equal to the minimization of $\|w\|$; therefore, the SVM problem for obtaining hyperplane $H_0$ that maximizes the margin is...
formulated as the following quadratic programming problem:

\begin{align}
\text{Minimize} & \quad \frac{1}{2} w^T w \\
\text{Subject to} & \quad y_i (w^T x_i + b) \geq 1; \ i = 1, 2, \ldots, n
\end{align}

In the basic concept and method of SVM, as shown in Fig. 2, it is assumed that the sample data can be linearly separated; however, in a realistic problem, there is a high probability that the boundary of the two groups is a complex superposition. In this case, if it is possible to convert the sample data that cannot be linearly separated into another space that enables linear separation, then we can use an SVM so that the boundary between the groups becomes a hyperplane in the space of the conversion destination. Therefore, the space in which the original sample data exists is called the input space, and the space of the destination is called the feature space \( \Phi(x_i) \). After input sample \( x_i \) is transformed into the sample in the feature space, the SVM can be applied to the sample because the sample is linearly separated in the feature space. Corresponding to Eq. (2.6), the discriminant hyperplane in feature space \( \Phi(x_i) \) becomes the optimal solution for the following minimization problem:

\begin{align}
\text{Minimize} & \quad \frac{1}{2} w^T w \\
\text{Subject to} & \quad y_i (\Phi(x_i) + b) \geq 1; \ i = 1, 2, \ldots, n
\end{align}

### 2.2.1 Soft margin SVM and SVM for regression analysis

In the basic concept and method of SVM, it is assumed that the sample data can be linearly separated, but in a real problem, there may be no hyperplane that can perfectly separate the sample data in the feature space. Therefore, relaxation variable \( \varepsilon_i \geq 0 (i = 1, 2, \ldots, n) \) is introduced so that sample data that does not satisfy the constraint condition in Eq. (2.7) may exist. The constraint condition of Eq. (2.7) is relaxed as follows:

\begin{align}
y_i (w^T \Phi(x_i) + b) \geq 1 - \varepsilon_i
\end{align}

Under the relaxed constraints of Eq. (2.8), soft margin SVM is formulated as follows:

\begin{align}
\text{Minimize} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \\
\text{Subject to} & \quad y_i (w^T \Phi(x_i) + b) \geq 1 - \varepsilon_i \\
& \quad \varepsilon_i \geq 0; \ i = 1, 2, \ldots, n
\end{align}

Where, \( \sum_{i=1}^{n} \varepsilon_i \) is the upper limit value of sample data misclassified, and \( C (C \geq 0) \) is a penalty[16][17].

### 2.2.2 Kernel functions and ksvm in kernlab

An SVM has several improved forms, one of which is SVM by the kernel method. An SVM by the kernel method is a nonlinear classifier that is represented by a linear function using a kernel function. The expression is optimized to maximize the margin between the classes in the projection space.

The kernel functions return the inner product between the two points in a suitable feature space, defining a notion of similarity. Kernels are used in the kernel methods, and the package kernlab features various kernel-based methods and includes the SVM method based on the optimizers.

It aims to deliver a flexible and an extensible SVM implementation. The purpose of package kernlab [18] is to provide an R user with a basic kernel functionality in the kernel-based methods. Package kernlab is available from CRAN (http://CRAN.R-project.org/) under the GPL license.

ksvm() function, the kernlab implementation of the SVMs, provides a standard formula for the interface as a matrix interface. The SVM implementations available in ksvm() include the C-SVM classification algorithm and v-SVM classification. ksvm() function in kernlab is a flexible SVM implementation that includes most of the SVM formulations and kernels.

The general format of the function K SVM is described below:

ksvm(formula, data, type="C-svc", kernel = "rbfdot", kpar=list(sigma=0.1), cross=k)

The argument "formula" specifies the label of the group, such as the model expression " \( y \sim x_1 + x_2 + \cdots \)", or the objective variable \( y \) and explanatory variables \( x_1, x_2 \) etc. The variable to be used is labelled with the symbol "+" shown as follows in the case of seven financial variables:

\( y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \)

The argument "data" describes a data frame containing the data of the objective and explanatory variables specified in "formula".

The argument "type" specifies the type of classification and regression analysis model. In addition to the basic model proposed by Vapnik, several revisions of the SVM method have been introduced. As a specific example, a classification method based on the formulation of Eq. (2.8) is designated as "type = C-svc". This is the soft margin SVM classification (C-support vector classification) proposed by Cortes and Vapnik [10].

The argument "kernel" specifies the kernel function. In this research, the following three types of kernel functions were used.

- kernel="rbfdot"
- kernel="polydot"
- kernel="laplacedot"

In addition to the above arguments, penalty \( C \) in Eq. (2.8) is specified as the argument of ksvm function, e.g., \( C = 10.0 \). Furthermore, the error upper limit value of the v-SVM model or \( v = 0.2 \) is specified, which is the lower limit value of the number of support vectors.

The argument "kpar" sets the parameters related to the kernel function specified by the argument "kernel". As a specific example, set the \( \sigma \) value of Gaussian RBF or Laplace RBF kernel as \( \sigma = s \), such as "kpar = list (sigma = s)". " s " is a specific numerical value, and in this study,
s = 0.1 was set. The argument "cross" is set to cross = k when performing a k multiple intersection verification, where k is an integer \( k \geq 2 \).

The classification method (Gaussian RBF kernel) based on the v-SVM model used in this study is shown below. The financial indicator used included 23 indicators.

\[
> \text{ksvm(bunrui}\sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} + x_{22} + x_{23}, \text{data} = X, \text{type} = "C-svc", \text{kernel} = \text{"laplacedot"}, \text{C}=10, \text{cross}=5)
\]

As the default kernel, "kernel = rbf" and "kernel = poly" are used in this study. These functional formulas are defined in references [19] [20] [21][22].

2. 3 Fitness function

The fitness function used herein is expressed in Eq. (2.9) [23]. The fitness function is defined as the number of successes in the correct classification of bankrupt and non-bankrupt companies. This function utilizes the financial indicators of both correct and incorrect predictions.

\[
f = \frac{T_p + F_n}{T_p + T_n + F_p + F_n}
\]

(2.9)

Where true positive (\( T_p \)) is the number of bankrupt companies that the rule states are bankrupt, false positive (\( F_p \)) is the number of non-bankrupt companies that the rule predicts to be bankrupt, true negative (\( T_n \)) is number of bankrupt companies that the rule predicts to be non-bankrupt, and false negative (\( F_n \)) is number of non-bankrupt companies that the rule predicts to be non-bankrupt.

3. Simulation design

3.1 Sample size

We generated a bankruptcy prediction model by the decision tree and SVM methods. The data of the company to be used are the financial statements of the listed companies in the Tokyo Stock Exchange, the Osaka Securities Exchange, and other stock exchanges. Bankruptcy was defined as a company that applied the civil rehabilitation law, bankruptcy suspension, bankruptcy, business activity stoppage, and the data of 244 bankrupt companies that went bankrupt between 1991 and 2014 are used. On the other hand, data of 3260 non-bankrupt companies that did not become bankrupt between 1991 and 2014 for 24 years are used. In total, 64708 non-bankrupt companies data has been used. The data is acquired from Nikkei NEEDS database. It is possible to consider bankruptcy prediction for all the industries, but considering special circumstances according to the industry, it is believed that the bankruptcy prediction can be achieved with a high accuracy. The data of 244 bankrupt companies are classified into 11 industries: construction industry (43 companies), real-estate industry (34 companies), services industry (20 companies), retail industry (19 companies), electrical equipment industry (17 companies), machinery industry (16 companies), wholesale industry (15 companies), other financial services industry (13 companies), textiles and apparels industry (9 companies), information and telecommunications industry (8 companies), and other industries (31 companies).

3.2 Financial indicators and Software

We selected some financial indicators with high discrimination power from the 23 financial indicators presented in the Appendix. We first investigated the 23 financial indicators and categorized the companies based on profitability, safety, efficiency, ability to pay, fund recovery capacity, funding capacity, and cash flow.

The software used to perform the simulation, predict, and conduct the overall study is the open source statistical software R.

4. Empirical analysis result

4.1 All industries classification

This research predicted the bankruptcy of all the industries by using the decision tree method without separating the entire industry into individual industries at first. However, it was impossible to predict bankruptcy accurately. To be specific, the number of bankrupt companies that the decision tree methods predict precisely was 77 companies out of bankrupt companies. The ratio is 34.38% as follows.

| Table 1 | All industries by decision tree | Total | Total precision(%) |
|--------|--------------------------------|-------|--------------------|
|        | bankruptcy status | Total  | 0 | 1 | 224 | 99.72% |
| Number |                        |        | 0 | 77 | 147 | 224 |
| %      |                        |        | 0 | 34.38% | 65.63% | 100.00% |
| 1      |                        |        | 0 | 0.05% | 99.95% | 100.00% |

In the table, 0 represents bankruptcy and 1 represents non-bankruptcy. In addition, the figure is the number of companies.

4.2 Industry classification

It was impossible to predict precisely the bankruptcy of all the industries without separating the entire industry. So we decided to classify the entire industry into eleven industries. The results of the construction industry based on decision tree are shown in Table 2 and Fig.3. In Fig. 3, 0 represents bankruptcy and 1 represents non-bankruptcy. Similarly in the following table, the numbers represent the
bankruptcy and non-bankruptcy. The model utilizing rpart classified the companies in the construction industry with 99.17% accuracy into bankrupt and non-bankrupt groups and obtained 3942 correct predictions from 3975 available companies. In Table 2, 99.71% of precision is the value of fitness function / in Eq. (2.9). Hereinafter, the numbers in the precision column on the other industries are similarly used. The results show that perfectly the model classifies the bankrupt companies with 25.58% accuracy and obtains 11 correct predictions from 43 bankrupt companies. The model also predicts 3931 non-bankrupt companies from 3932 non-bankrupt companies with 99.70% accuracy. On the other hand, the results of the construction industry based on SVM are shown in Table 3. The model utilizing SVM classifies the companies in the construction industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 714 correct predictions from 714 companies. The model also predicts 680 non-bankrupt companies from 680 non-bankrupt companies with 100% accuracy. On the other hand, the results of the real-estate industry based on SVM are shown in Table 5. The model utilizing SVM classifies the companies in the real-estate industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 714 correct predictions from 714 companies.

The results of the real-estate industry based on decision tree are shown in Table 4 and Fig. 4. The model utilizing rpart classifies the companies in the real-estate industry with 99.30% accuracy into bankrupt and non-bankrupt groups and obtains 3932 correct predictions from 3932 available companies. The results show that the model classifies bankrupt companies with 85.29% accuracy and obtains 29 correct predictions from 34 bankrupt companies. The results of the service industry based on decision tree are shown in Table 6 and Fig. 5. The model utilizing rpart classifies the companies in the service industry with 99.72% accuracy into bankrupt and non-bankrupt groups and obtains 4257 correct predictions from 4269 available companies. The results show that the model classifies these industries with high accuracy into bankrupt and non-bankrupt groups and obtains good results from 20 companies. The model also predicts 4247 non-bankrupt companies from 4249 non-bankrupt companies with 99.95% accuracy. On the other hand, the results of the service industry based on SVM are shown in Table 7. The model utilizing SVM classified the companies in the service industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 4269 correct predictions from 4269 companies.
Table 6 Service industry by decision tree

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 10    | 10                | 99.72% |
| 1                 | 4247  | 4249              |       |
| %                 | 0.00% | 50.00% 100.00%   |       |

Table 7 Service industry by SVM

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 20    | 20                | 100.00% |
| 1                 | 4249  | 4249              |       |
| %                 | 100.00% | 0.00% 100.00% |       |

Table 8 Retail industry by decision tree

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 9     | 10                | 99.70% |
| 1                 | 5022  | 5027              |       |
| %                 | 0.10% | 99.90% 100.00%   |       |

Table 9 Retail industry by SVM

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 18    | 18                | 99.98% |
| 1                 | 5027  | 5028              |       |
| %                 | 100.00% | 0.00% 100.00% |       |

The results of the retail industry based on decision tree are shown in Table 8 and Fig. 6. The model utilizing rpart classifies the companies in the retail industry with 99.70% accuracy into bankrupt and non-bankrupt groups and obtains 5031 correct predictions from 5046 available companies. The results show that the model classifies bankrupt companies with 47.37% accuracy and obtains 9 correct predictions from 19 bankrupt companies. The model also predicts 5022 non-bankrupt companies from 5027 non-bankrupt companies with 99.90% accuracy.

On the other hand, the results of the retail industry based on SVM are shown in Table 9. The model utilizing SVM classifies the companies in the retail industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 5046 correct predictions from 5046 companies.

Table 10 Electric appliances industry by decision tree

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 7     | 10                | 99.79% |
| 1                 | 6286  | 6289              |       |
| %                 | 0.10% | 99.90% 100.00%   |       |

Table 11 Electric industry by SVM

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|-------------------|
| 0                 | 17    | 17                | 100.00% |
| 1                 | 6289  | 6289              |       |
| %                 | 0.00% | 100.00% 100.00%  |       |

The results of the electric industry based on decision tree are shown in Table 10 and Fig. 7. The model utilizing rpart classifies the companies in the electric industry with 99.79% accuracy into bankrupt and non-bankrupt groups and obtains 6293 correct predictions from 6306 available companies. The results show that the model classifies bankrupt companies with 41.18% accuracy and obtains 7 correct predictions from 17 bankrupt companies. The model also predicts 6286 non-bankrupt companies from 6289 non-bankrupt companies with 99.95% accuracy. On the other hand, the results of the electric industry based on SVM are shown in Table 11. The model utilizing SVM classifies the companies in the electric industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 6393 correct predictions from 6293 companies.
classifies bankrupt companies with 0.00% accuracy and obtains 0 correct predictions from 16 bankrupt companies. The model also predicts 5264 non-bankrupt companies from 5264 non-bankrupt companies with 100.00% accuracy. On the other hand, the results for the machinery industry based on SVM are shown in Table 13. The model utilizing SVM classifies the companies in the machinery industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 5280 correct predictions from 5280 companies.

The results of the wholesale industry based on decision tree are shown in Table 14 and Fig. 9. The model utilizing rpart classifies the companies in other financial industries with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 524 correct predictions from 524 available companies. The results show that the model classifies bankrupt companies with 100.00% accuracy and obtains 13 correct predictions from 13 bankrupt companies. The model also predicts 511 non-bankrupt companies from 511 non-bankrupt companies with 100.00% accuracy. On the other hand, the results of the machinery industry based on SVM are shown in Table 15. The model utilizing SVM classifies the companies in other financial industries with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 5951 correct predictions from 5951 companies.
The results of the textile products industry based on decision tree are shown in Table 18 and Fig. 11. The model utilizing rpart classifies the companies in the textile products industry with 99.74% accuracy into bankrupt and non-bankrupt groups and obtains 2336 correct predictions from 2342 available companies. The results show that the model classifies bankrupt companies with 55.56% accuracy and obtains 5 correct predictions from 9 bankrupt companies. The model also predicts 2331 non-bankrupt companies from 2333 non-bankrupt companies with 99.91% accuracy. On the other hand, the results of the textile products industry based on SVM are shown in Table 19. The model utilizing SVM classifies the companies in the textile products industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 2342 correct predictions from 2342 companies.

The results of the information and communication industry based on decision tree are shown in Table 20 and Fig. 12. The model utilizing rpart classifies the companies in the information and communication industry 99.84% accuracy into bankrupt and non-bankrupt groups and obtains 4437 correct predictions from 4444 available companies. The results show that the model classifies companies with 75.00% accuracy and obtains 6 correct predictions from 8 bankrupt companies. The model also predicts 4431 non-bankrupt companies from 4436 non-bankrupt companies with 99.89% accuracy. On the other hand, the results of the information and telecommunication industry based on SVM are shown in Table 21. The model utilizing SVM classifies the companies in the information and communication industry with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 4444 correct predictions from 4444 companies.

The results of the other industries based on decision tree are shown in Table 22 and Fig. 13. The model utilizing rpart classifies the companies in the other industries with 99.90% accuracy into bankrupt and non-bankrupt groups and obtains 26043 correct predictions from 26068 available companies. The results show that the model classifies bankrupt companies with 22.58% accuracy and obtains 7 correct predictions from 31 bankrupt companies. The model also predicts 26036 non-bankrupt companies from 26037 non-bankrupt companies with 99.996% accuracy. On the other hand, the results of the other industries based on SVMs are shown in Table 23. The model utilizing SVM classifies the companies in the other industries with 100.00% accuracy into bankrupt and non-bankrupt groups and obtains 26068 correct predictions from 26068 companies.
Table 22 Other industry by decision tree

| Number | 0 | 1 | Total | Precision (%) |
|--------|---|---|-------|--------------|
| %      | 7 | 24 | 31    | 99.90%       |

Table 23 Other industry by SVM

| Number | 0 | 1 | Total | Precision (%) |
|--------|---|---|-------|--------------|
| %      | 31 | 26036 | 26037 | 100.00%       |

Table 24 Key financial indicators for dividing root node and child node

|        | Short-term debt rotation period | Net sales margin ratio | Debt ratio for total assets | Capital share | Total capital net income margin | Quick ratio | Current liability turnover | Debt turnover period |
|--------|--------------------------------|------------------------|----------------------------|---------------|---------------------------------|-------------|----------------------------|---------------------|
| Industry |                                |                        |                            |               |                                 |             |                            |                     |
| Construction industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Real estate industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Service industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Retail industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Electric appliances industry | ○                   |                        |                            |               |                                 |             |                            |                     |
| Machinery industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Wholesale industry | ○                           |                        |                            |               |                                 |             |                            |                     |
| Other financial industry | ○                |                        |                            |               |                                 |             |                            |                     |
| Textile apparels industry | ○               |                        |                            |               |                                 |             |                            |                     |
| Information and telecommunication industry | ○          |                        |                            |               |                                 |             |                            |                     |
| Other industry | ○                           |                        |                            |               |                                 |             |                            |                     |

Let us examine the relationship between the financial indicators and industries in Table 24. The key financial indicators predicting the bankruptcy were found to be different for each industry. Specifically, it is as follows. The key financial indicators were safety indicator (short-term debt rotation period, debt ratio for total assets, quick ratio, current liability turnover, debt turnover period) in 7 industries. These industries are the construction industry, real estate industry, retail industry, machinery industry, other financial industry, information and telecommunication industry, and other industry. The key financial indicator was the profitability indicator (net sales margin ratio, total capital net income margin) in 3 industries. These industries are the service industry, electric appliances industry, and wholesale industry. The key financial indicator of textile apparels industry was the productivity index (Capital share). On the other hand, the SVM prediction of bankruptcy was perfectly conducted to discriminate the industries. As the default of Kernel, "kernel = rbf(dot)" and "kernel = poly(dot)" are used other than "kernel = laplacedot". To be specific, kernel = "laplacedot" of kernel was able to predict bankruptcy and non-bankrupt companies almost perfectly, so "4. Empirical analysis result" shows only the result of kernel = "laplacedot". Here, the effectiveness of kernel = "laplacedot" is shown below for the construction industry as an example.

The following is an example of kernel = "rbfdot", cross = 5).
Prediction of bankruptcy on industry classification

<ksvm(bunrui−a1+a2+a3+a4+a5+a6+a7+a8+a9+a10+a11+a12+a13+a14+a15+a16+a17+a18+a19+a20+a21+a22+a23,data=X,type="Csvc",kernel="rbfdot",cros
s=5))

Table 25 Kernel="rbfdot",cross=5

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|--------------------|
| Number            |       |                    |
| 0                 | 13    | 30                 |
| 1                 | 0     | 3932               |
| %                 | 30.23%| 100.00%            |

The following is an example of kernel="laplacedot", C=10, cross=5.
<ksvm(bunrui−x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+data=X,type="Csvc",kernel="laplacedot",C=10,cross=
5)).

Table 26 Kernel="rbfdot",C=10,cross=5

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|--------------------|
| Number            |       |                    |
| 0                 | 37    | 6                  |
| 1                 | 0     | 3932               |
| %                 | 86.05%| 100.00%            |

The following is an example of kernel= "laplacedot", C=10,cross=5.
<ksvm(bunrui−x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+data=X,type="Csvc",kernel="laplacedot",C=10,cross=
5)).

Table 27 Kernel="laplacedot",C=10,cross=5

| bankruptcy status | Total | Total precision(%) |
|-------------------|-------|--------------------|
| Number            |       |                    |
| 0                 | 43    | 0                  |
| 1                 | 0     | 3932               |
| %                 | 100.00%| 100.00%           |

As shown in 「4. Empirical analysis result」, "kernel = "laplacedot", C = 10, cross = 5" was predictable almost perfectly even in industries other than the construction industry. The result of this research clarified the relationship between bankruptcy in the industry and financial indicators for forecasting bankruptcy. This research proposed financial indicators that can predict bankruptcy precisely for each industry, so this prediction bankruptcy system will assist companies to improve their financial situation.

6. Conclusion

This study predicted the bankruptcy by separating all the companies into individual industries by using the decision tree method and SVMs, and compared the methods. This research was predicted two steps. The first step was to predict using the decision tree method and the second step was to predict using SVM for eleven industries. In the decision tree analysis, only for some industries bankruptcy prediction could be made accurately. On the other hand, SVM could predict bankruptcy in companies almost perfectly for each industry. It can be derived the following conclusions from the results of this study. Customers, investors and financiers can prevent losses by focusing on the information of these financial indicators before finalizing the transaction. On the other hand, managers can improve their business by paying attention to these financial indicators.

References

[1] Beaver, W. H.: Financial indicators as predictors of Failure, Journal of Accounting Research, Empirical Research in Accounting: Selected Studies: Supplement, 4 (1966), 71–111.
[2] Altman E. I.: Financial indicators, discriminant analysis and the prediction of corporate bankruptcy, The Journal of Finance, 23(4) (1968), 589–609.
[3] Ohlson J. A.: Financial indicators and the probabilistic prediction of bankruptcy, Journal of Accounting Research, 18 (1980), 109-131.
[4] Deakin, E.: A discriminant analysis of predictors of business failure, Journal of Accounting Research, Spring (1972), 167–179.
[5] Barniv R., Agarwal, A., and Leach, R.: Predicting the outcome following bankruptcy filing-A three-state classification using neural networks, Intelligent Systems in Accounting, Finance and Management, 6 (1997),177–194.
[6] 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, 6 (1997), 249–264.
[7] Zhang, G., Hu, Y. M., Patuwo, E. B., and Indro, C. D.: Artificial neural networks in bankruptcy prediction, General framework and cross validation analysis, European Journal of Operational Research, 116 (1999), 16–32.
[8] Odom, M., and Sharda, R.: A neural networks model for bankruptcy prediction. In Proceedings of the IEEE international conference on neural network, Vol.2 (1990), 163–168.
Appendix

| Number | Ratio |
|--------|-------|
| 1      | Deb capacity ratio |
| 2      | Short-term debt rotation period |
| 3      | Net sales margin ratio |
| 4      | Debt ratio for total assets |
| 5      | Loan To Value |
| 6      | Ratio of profit before tax to sales |
| 7      | Capital share |
| 8      | Total capital net income margin |
| 9      | Quick ratio |
| 10     | Current liability turnover |
| 11     | Total capital ordinary income ratio |
| 12     | Interest-bearing debt monthly sales ratio |
| 13     | Debt monthly sales magnification |
| 14     | Debt turnover period |
| 15     | Ordinary income ratio |
| 16     | Interest coverage ratio |
| 17     | Operating profit per employee |
| 18     | Operating profit per employee |
| 19     | Sales financial balance ratio |
| 20     | Long-term debt repayment years |
| 21     | Finance cost ratio |
| 22     | Ratio of operating profit to operating capital |
| 23     | Return on asset |

Masanobu MATSUMARU, Takaaki KAWANAKA, Hideki KATAGIRI, Shoichi KANEKO

International Journal of Japan Association for Management Systems

- 12 -