Financial Distress Prediction using Linear Discriminant Analysis and Support Vector Machine

Noviyanti Santoso, Wahyu Wibowo
Department of Business Statistics, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo-Surabaya 60111, Indonesia
E-mail: noviyanti_s@statistika.its.ac.id

Abstract. A financial difficulty is the early stages before the bankruptcy. Bankruptcies caused by the financial distress can be seen from the financial statements of the company. The ability to predict financial distress became an important research topic because it can provide early warning for the company. In addition, predicting financial distress is also beneficial for investors and creditors. This research will be made the prediction model of financial distress at industrial companies in Indonesia by comparing the performance of Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) combined with variable selection technique. The result of this research is prediction model based on hybrid Stepwise-SVM obtains better balance among fitting ability, generalization ability and model stability than the other models.

1. Introduction
Economic problems in a country indirectly have an impact on the financial condition of the company. At the time of the exchange rate of the currency weakening, inflation, and soaring fuel prices will affect the performance of the company. It takes effort to monitor, evaluate, and control corporate finance in order to avoid financial distress. The condition of the financial distress that is not immediately corrected can lead to bankruptcy of the company that caused the economic problems to the company's management, investors, creditors, suppliers, and employees. Based on this reason, the financial distress prediction became an important topic of discussion in financial affairs.

A diverse range of research both in the field of accounting or finance is working to develop a model prediction of financial distress. Nowadays, the researches of prediction of financial distress are also increasingly interested in the field of statistics and machine learning. Firstly, the prediction of bankruptcy starts by [1] that makes the model using univariate analysis of financial ratios and concluded that the company's financial ratios can be used to distinguish between bankrupt and non-bankrupt companies. Then [2] identified the companies went bankrupt with combined 22 financial ratios using Multiple Discriminant Analysis (MDA). As many as five significant variables resulted in such research, i.e. working capital/total assets, retained earnings/total assets, EBIT/total assets, market value of equity/total liabilities, and sales/total assets. Ohlson [3] using a logistic regression to find out the possibilities of a company facing financial distress.

In the development of data mining in the era of 1990 was emerging variety of classification methods, such as a decision tree (DT), case-based reasoning (CBR), artificial neural networks (ANNs), evolutionary genetic algorithm (GA) and a new technique of support vector machine (SVM). The technique was applied to predict financial distress of companies [4]. Shin and Lee using SVM to predict the bankruptcy of several companies in South Korea and concluded that SVM show highest
prediction accuracy than logistic regression, MDA, and NN [5]. [6] and [7] also use SVM for predicting financial trouble in companies in China and reported that SVM outperformed other classifiers. Nisa [8] using SVM and LDA for predicting financial trouble in manufacturing company in Indonesia and obtained results that SVM give accuracy appropriately.

According to [9] variable that will be used as input will affect the accuracy. Therefore, the selection of variable financial ratios is an important step in pre-processing. Variable selection technique that is commonly used is the Principal Component Analysis (PCA), Factor Analysis, t-test, stepwise regression and correlation matrix [10]. Considering what mentioned above, the purpose of this study is to apply stepwise for selecting significant variables to use in prediction financial distress of industrial companies listed in the Indonesia Stock Exchange using LDA and SVM.

2. Materials and Methods
Classifications are statistical methods used for constructing models of predictive with the aim of separate and classify new data. All the classification methods use set features or parameters be characteristic of objects. There are two kinds of the classification methods that supervised learning and unsupervised learning. Supervised learning is an approach to where there are trained data, and there are variable targeted so the purpose of this approach has is made a data to the existing, while unsupervised learning has no data trainer, that data from, we classify this data then becomes 2 part or 3 parts and so on [11]. In this study the classification methods used are LDA and SVM where included in supervised learning.

2.1. Linear Discriminant Analysis
The discriminant analysis is a method used to distinguish two or more groups based on some independent variables. In this method, the independent variable can be either metric or non-metric but different in terms of the nature of the dependent variable. In the linear discriminant analysis, the dependent variable is the categorical / non-metric variable. If the number of categories is two, discriminant analysis is similar to binary logistic regression analysis [12]. The linear discriminant function takes form of a linear combination of coefficients of variables and their respective variables in the study as equation 1.

\[ Z = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k \]  

where, \( Z \) = discriminant score, \( \alpha \) = a constant term, \( \beta_i \) = the discriminant coefficient or weight of the variable, \( X_i \) = predictor or independent variable, \( i \) = number of predictor variables; \( i = 1, 2, 3, \ldots, k \).

Function of discriminant worthy to form if there are the differences in value the average between groups. Hence prior to the formation of discriminant function we need to done testing distinction vector the average of the groups has. Discriminant analyses having the assumption that must be fulfilled of them are normal multivariate distribution, covariance matrices to all groups is homogeneous, and every group has the average different [13]. But R.A. Fisher proposed another method for linear discriminant analysis that did not presuppose any known distribution of the training data, namely Fisher’s discriminant. The idea is finding projection of \( a^TX \) like best separating that result ratio between group sums of square within group sum of square maximal. An object will classifying into group \( l \) if \( a^TX_j \) is near to \( a^TX \), \( x \rightarrow \Pi_l \), where \( l = \arg \min_i [a^T(x - \bar{x}_l)] \).

2.2. Support Vector Machine
Support vector machine, a more recent learning algorithm that has been developed from statistical learning theory [14, 15], has a very strong mathematical foundation and has been shown to exhibit excellent performance in time series forecasting [16] and in classification [17]. The SVM technique has been effectively used to estimate parameters of multivariate function, non-linear regression problem, enhances generalization capability and uncommon representation of the solution [18]. SVM is a promising technique that follows the principle of structural risk minimization wherein the problem of over fitting by balancing the model complexity [19]. The basic of SVR is that non linearity the original dataset \( x_l \) is mapped into a high-dimensional feature space.
\((x_i, y_i)\) where \(i = 1, 2, ..., \lambda\) with \(x = \{x_1, x_2, x_3\} \ldots \subseteq \mathbb{R}^N\) as input variable vector and \(y = \{y_1, ..., y_\lambda\} \ldots \subseteq \mathbb{R}\) are output variable. The hyperplane function is as equation 2:

\[
f(x) = \mathbf{w} \times \varphi(x) + b \tag{2}
\]

where \(\varphi(x)\) denotes the high-dimensional feature space, which is non-linearity mapped from input space \(x\), \(\mathbf{w}\) and \(b\) are coefficients that estimated by minimizing the regularized risk function as follows:

\[
\min \frac{1}{2} \lVert \mathbf{w} \rVert^2 + C \sum_{i=1}^{\lambda} L_\varepsilon(y_i, f(x_i)) \tag{3}
\]

\[
L_\varepsilon(y_i, f(x_i)) = \begin{cases} 
|y_i - f(x_i)| - \varepsilon & \text{if } |y_i - f(x_i)| \geq 0 \\
0 & \text{otherwise}
\end{cases} \tag{4}
\]

Factor \(\lVert \mathbf{w} \rVert^2\) is called regularization, minimizing the \(\lVert \mathbf{w} \rVert^2\) will make the flatness function, which can control the function capacity. The parameter \(\varepsilon\) represent the distance between actual values and values calculated from regression function that measured with \(\varepsilon\)-insentive loss function. Using \(\varepsilon\)-insentive loss function must be minimizing norm from \(\mathbf{w}\) in order to get appropriate generalization for regression function. \(C\) denotes a cost function measuring empirical risk. The constant \(C > 0\) determine trade off between the flat of function and limit of deviation \(\varepsilon\). Then the constrained form can be formulated as follows:

\[
\min \frac{1}{2} \lVert \mathbf{w} \rVert^2 + C \sum_{i=1}^{\lambda} L_\varepsilon(\xi_i, \xi_i^*) \tag{5}
\]

The two slack variables \(\xi_i\) and \(\xi_i^*\) can be introduced to represent the distance from actual values to the corresponding boundary values in \(\varepsilon\)-tube. Finally, the constrained optimization problem is solved using the following form:

\[
\max R(\alpha_i, \alpha_i^*) = \sum_{j=1}^{\lambda} (\alpha_i - \alpha_i^*)y_i - \varepsilon \sum_{j=1}^{\lambda} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \tag{6}
\]

Here \(\alpha_i\) and \(\alpha_i^*\) are Lagrange multipliers. \(K(x_i, x_j)\) is kernel function. In this study, a linear is used as kernel function in SVM which denote as \(K(x_i, x_j) = x_i^T x_j\). Hence, non linear regression function is formulated as equation 7:

\[
f(x) = \sum_{i=1}^{\lambda} (\alpha_i - \alpha_i^*)K(x_i, x_j) + b \tag{7}
\]

2.3. Dataset

The dataset that used in this research are financial report of industrial company that listed on the Indonesia stock exchange. Companies that become object observations in this research are company that have a complete report financial consist of the report loser profit, the balance, equity changes report, and reports cash flow. A financial report will be used in this research was the annual financial report that has been audited in the period 2008-2014 by the number of company at any period is 31. The input variables that used refer to the previous research, adopting 20 financial ratio indexes shown in table 1 as follows:
Table 1. Financial ratio indexes.

| Variable | Definition | Variable | Definition |
|----------|------------|----------|------------|
| X_1      | EBIT/Total Asset | X_{11}   | Net Profit Ratio |
| X_2      | Sales/Total Asset | X_{12}   | EBIT/Sales |
| X_3      | Sales/Fixed Asset | X_{13}   | ROI |
| X_4      | Earning/DEBT | X_{14}   | Working Capital/Long Term DEBT |
| X_5      | Current Ratio | X_{15}   | DEBT to Equity |
| X_6      | Working Capital/Total Asset | X_{16}   | Book Equity/Total Capital |
| X_7      | ROE | X_{17}   | Market Value Equity/Total Capital |
| X_8      | Retained Earning/ Total Asset | X_{18}   | Liabilitas |
| X_9      | Gross Profit Ratio | X_{19}   | PER |
| X_{10}   | Operating Profit Ratio | X_{20}   | PBV |

The company financial distress prediction as output variables will be 0 and 1. There are 0 for financially healthy company and 1 for financial crisis company. The criteria used to determine the condition financial distress of company there are two, namely interest coverage ratio and profit negative. By due to count interest coverage ratio required value interest burden interest expense but an enterprise not always displayed in the report a finance company so by negative profit of company for a year or at year \( t \) we can determine bankruptcy.

2.4. Methodology

In order to obtain a model and prediction, the data were parted into two parts, which were training set and testing set. The data used for modelling is training set that applied in three scheme dynamic periods \( (k) \), there are \( k = 0, 1, \) and \( 2 \). Dynamic period with \( k = 0 \), means when input variables are in period \( t \) then output variable is in period \( t \) too. Additionally, \( k = 1 \) means when input variables are in period \( t \) then output variable is in period \( t-1 \) and so on. Afterward, stepwise was applied as selection variables technique to see the effectiveness. So, we have combined the scheme and approach using LDA and SVM to build model of prediction. Meanwhile, the data used for prediction is testing set that consist of one last period. The last step was to compare the prediction accuracy using area under Receiver Operating Characteristic (ROC) curve that appropriate for unbalance data [20]. That curve commonly known as AUC calculated using this following equation:

\[
AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}
\]

Table 2. Confusion Matrix

| Actual | Prediction |
|--------|------------|
| Positive = class 0 | True Positive (TP) | False Negative (FN) |
| Negative = class 1 | False Positive (FP) | True Negative (TN) |

Where \( TP_{rate} \) and \( FP_{rate} \) can be obtained by use equation 9 and equation 10 as follow:

\[
TP_{rate} = Sensitivitas = recall = \frac{TP}{TP + FN}
\]

\[
FP_{rate} = \frac{FP}{FP + TN}
\]
3. Results and Discussion
This study use 186 data obtained from financial report of industrial company in Indonesia. There are 3 points that become focus discussion. Firstly, we would find the greatest dynamic period of data. Secondly, found the effectiveness of selection variables technique and got some significant variables in prediction model and the last was find out the appropriate prediction model for financial distress. Based on Table 3 and Table 4, the highest AUC both training and testing set occur when dynamic period is \( k = 0 \), but in fact is ineffective for company, because to predict the financial distress in the end of year, they need financial report in the same year. Whereas, the evaluation of company’s performance occurs on December even early next year. But, comparing the score of AUC between \( k = 1 \) and \( k = 2 \), the first is more appropriate. It means, the company have to prepare the report of achievement on previous year to predict the financial condition in one period. So, the early warning financial can be useful for company’s financial.

| Dynamic Periods (\( k \)) | Methods | Training | Testing |
|---------------------------|---------|----------|---------|
| 0                         | LDA     | 0.89     | 0.69    |
|                           | SVM     | 1        | 0.93    |
| 1                         | LDA     | 0.71     | 0.62    |
|                           | SVM     | 0.64     | 0.63    |
| 2                         | LDA     | 0.70     | 0.54    |
|                           | SVM     | 0.61     | 0.50    |

The aim variable selection was to accomplish the important variables that affect the prediction model. In some research, selected variables can increase the accuracy of prediction. There are 5 financial ratio indexes that preferred using stepwise i.e. EBIT/Total Asset \((X_1)\), Retained Earning \((X_8)\), Operating Profit Ratio \((X_{10})\), ROI \((X_{13})\), and PER \((X_{19})\). Only one variable that relevant with the results of Herlina [21] and Wulandari [22] who have obtained PER, current ratio, ROE, leverage ratio, and EPS as significant variables in financial distress prediction. Herlina used analytic hierarchy process (AHP) as variable selection methods, while Wulandari used LP-SVM. Different variable selection method will obtain various results. Nevertheless, table 4 shown that there is enhancement score of AUC both LDA and SVM methods. It proves that variables selection is worthwhile to boost accuracy.

| Dynamic Periods (\( k \)) | Methods | Training | Testing |
|---------------------------|---------|----------|---------|
| 0                         | LDA     | 0.90     | 0.75    |
|                           | SVM     | 1        | 0.97    |
| 1                         | LDA     | 0.74     | 0.68    |
|                           | SVM     | 0.71     | 0.70    |
| 2                         | LDA     | 0.60     | 0.48    |
|                           | SVM     | 0.69     | 0.50    |

In general, financial condition of Indonesia listed companies can be better predicted using the proposed SVM model since the AUC score is 0.70 and prediction accuracy of SVM is as high as 74%. The comparison accuracy was shown completely in table 5. However, the proposed SVM model still include in “poor classification” category. Probably, it happens as a consequence the linear kernel that use in the SVM model.
| Methods | Training AUC | Accuracy % | Testing AUC | Accuracy % |
|---------|-------------|------------|-------------|------------|
| LDA     | 0.74        | 79%        | 0.68        | 70%        |
| SVM     | 0.71        | 78%        | 0.70        | 74%        |

4. Conclusion
The LDA and SVM models are successful to predict financial distress. Their predictions are able to get a nearly result. The SVM model with dynamic period $k = 1$ and hybrid with stepwise as selection variables method becomes the best model among these four combination models with maximum AUC and accuracy. It shows that the variables selection could significantly increase the prediction accuracy of both LDA and SVM models. We consider that this method could be the promising method to improve the prediction performance, especially in bankruptcy cases. However, in this study the realization accuracy of SVM models still “poor classification” category. Perhaps it caused by kernel function that used in the SVM models is linear. We consider for the further research to apply polynomial and radial basis function to improve the accuracy. Moreover, it requires optimizing the parameters of SVM using genetic algorithm.

Acknowledgments
The author would like to thanks to Indonesia Stock Exchange for provide data set and this work was supported by Grant for Beginner Research 2017 ITS Surabaya.

References
[1] Beaver W 1966 Financial ratios as Predictors of Failure *Journal of Accounting Research* 4 71-111
[2] Altman E I 1968 Financial Ratios Discriminant Analysis and The Prediction of Corporate Bankruptcy *Journal of Finance* 23 1-22
[3] Ohlson J A 1980 Financial Ratios and the Probabilistic Prediction of Bankruptcy *Journal of Accounting Research* 18 109-131
[4] Chen M Y 2011 Bankruptcy Prediction in Firms with Statistical and Intelligent Techniques and a Comparison of Evolutionary Approaches *Computers and Mathematics with Applications* 62 4514-4524
[5] Shin K S, Lee T S and Kim H J 2005 An Application of Support Vector Machine in Bankruptcy Prediction Model *Expert Systems with Applications* 28 127-135
[6] Hui X F and Sun J 2006 An Application of Support Vector Machine to Companies Financial Distress Prediction *Lecture Notes Artificial Intelligent* 3885 274-282
[7] Ding Y, Song X and Zeng Y 2008 Forecasting Financial Condition of Chinese Listed Companies based on Support Vector Machine *Expert Systems with Applications* 34 3081-3089
[8] Nisa U Z 2013 Model Prediksi Finansial Distress Pada Perusahaan Manufaktur Go Public di Indonesia *Thesis Program Magister Bidang Optimasi Sistem Industri Jurusan Teknik Industri Institut Teknologi Sepuluh Nopember*
[9] Lin F, Liang D and Chen E 2011 Financial Ratio Selection for Business Crisis Prediction *Expert Systems with Applications* 38 15094-15102
[10] Tsai C F 2009 Feature Selection in Bankruptcy Prediction *Knowledge Based Systems* 22 120-127
[11] Sthya R and Abraham A 2013 Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification International Journal of Advanced Research in Artificial Intelligence 2(2) 34-38

[12] Uddin N, Meah M S and Hossain R 2013 Discriminant Analysis as an Aid to Human Resource Selection and Human Resource Turnover Minimization Decisions International Journal of Business and Management 8 (17) 153-169

[13] Johnson R and Wichern D 2007 Applied Multivariate Statistical Analysis (New Jersey: Prentice Hall Inc.)

[14] Vapnik V N 1995 The Nature of Statistical Learning Theory (New York: Springer)

[15] Vapnik V, Golowich S and Smola A 1997 Support Vector Method for Function Approximation, Regression Estimation and Signal Processing In: Mozer M, Jordan M and Petsche T Eds Advance in Neural Information Processing System (Cambridge : MIT Press) 281-287

[16] Muller K R, Smola A J, Ratsch G, Scholkopf B, Kohlmorgen J and Vapnik V 1997 Prediction Time Series with Support Vector Machines Lecture Notes in Computer Science 1327 999-1004

[17] Yu L, Wang S and Cao J 2009 A Modified Least Squares Support Vector Machine Classifier with Application to Credit Risk Analysis International Journal of Information Technology & Decision Making 8(4) 697-710

[18] Vapnik V N 1999 An overview of statistical learning theory IEEE Transactions on Neural Networks / Publication of the IEEE Neural Networks Council 10(5) 988–99

[19] Scholkopf B and Smola A 2002 Learning with Kernels (Cambridge: The MIT Press)

[20] Wang S and Yao X 2013 Using Class Imbalance Learning for Software Defect Prediction IEEE Transactions on Reliability 434-443

[21] Herlina, Santosa B and Ciptomulyono U 2014 Pemodelan Prediksi Financial Distress Menggunakan Metode Support Vector Machine Prosiding Seminar Nasional Manajemen Teknologi XXI (ITS)

[22] Wulandari L M C, Daru A and Andrian D 2016 Aplikasi Data Mining dengan Metode Support Vector Machine (SVM) untuk Prediksi Financial Distress pada Industri Jasa Go Public yang Terdaftar di Bursa Efek Indonesia Prosiding Seminar Nasional Inovasi dan Aplikasi Teknologi di Industri (ITS)