The Comparison of Machine Learning Model to Predict Bankruptcy: Indonesian Stock Exchange Data

E Rainarli
Department of Informatics Engineering, Universitas Komputer Indonesia, Jl Dipatiukur 112 -116
Bandung, Indonesia

Email: ednawati.rainarli@email.unikom.ac.id

Abstract. This study aims to determine the Machine Learning Model used to predict bankruptcy. The data was conducted from the financial statements of two public companies reported by the Indonesia Stock Exchange from 2009 to 2015. This research method uses an analysis feature in which the accounting ratios are used in statistical analysis of financial statements that handle missing values, choose the correlation feature related to class, and dealing with unbalanced datasets. This problem was resolved at the beginning of the pre-processing phase. The training process uses pre-processing results to fit the data with the prediction model. Accuracy is used to measure the performance of the model in predicting bankruptcy. The result is Sequential Minimal Optimization (SMO) with linear kernel function that works best to predict 1 year before bankruptcy with an accuracy of 91.57% and SMO with Radial Basis Function (RBF) works well to predict 2 years before bankruptcy; the accuracy is 93.8%. This study shows the effect of feature selection and normalization process in making correct predictions using the SMO method.

1. Introduction
The bankruptcy of a company is a condition where a company is unable to repay a debt that has matured, or a company declared bankrupt by the court. It is important for investors, creditors or financial institutions to analyze conditions of the public company. The result helps them to determine the policies. For this reason, every year issuers must submit their financial reports to the public. Based on the financial statements, it calculates the financial ratios and those ratios are analyzed using statistical analysis. Business conditions becoming more complex, dynamic and difficult to fulfill the assumptions of normal, homogenous and independent. Therefore, Machine Learning can be required to predict bankruptcy.

According to the prediction of bankruptcy by Altman [1], he used the Multivariate Discriminant Analysis (MDA) method to determine the company's bankruptcy. The problems arise when using MDA are a prerequisite for normality of data, inequality of dispersion matrices from all groups and unfulfilled non-random-sampling requirements of bankrupt or non-bankrupt companies [2]. Barboza et al. used machine learning to predict corporate bankruptcy [3], they used 133 bankrupt companies and 13300 solvent company data. The results showed that the Random Forest method works properly between Linear Discriminant Analysis, Logistic Regression, Neural Network, Support Vector Machine, Boosting, and Bagging. Different results were shown by Klepac and Gogas [4-5] which found that SVM using RBF function worked well for predicting bankruptcy. Klepac used data from retail companies while Gogas used SVM to predict bankruptcy at the Bank. Antunes [6] indicated that the Gaussian Process worked better than the SVM or Logistic Regression to the data from the French market. A comparative study carried out by Tsai [7] to compare several methods in machine learning to predict bankruptcy. The difference in stock exchange data characteristics of each country lead to...
bankruptcy prediction needs in the Indonesian stock exchange needs to be reviewed by comparing several methods in machine learning. Moreover, the Veganzones [8] research explained the need to pay attention to the imbalanced data set conditions that exist in data of bankrupt companies. Le [9] described the oversampling technique to overcome unbalanced data problems. The existence of a missing value problem is something that can solved in preprocessing to ensure the success of predictions using machine-learning methods. The needed for financial ratios selection also influences the bankruptcy model. Wang, Liang, Dellepiane [10-12] have shown the effect of using feature selection on prediction results.

This study aims to determine the Machine Learning Model used to predict bankruptcy. Therefore, in this study we will examine several supervised learning methods to get the best model that can predict one before and two previous bankruptcies. The model considers various problems such as unbalanced datasets, selection of appropriate financial ratios, and handling incomplete datasets. In addition, this study will also determine ratios that affect the determination of company bankruptcy based on Indonesian stock exchange data. The stages carried out in this research are preprocessing, feature selection, training, and testing.

2. Method
The experimental method used to get a predictive model and suitable machine learning methods. Following are the steps of the research carried out:

2.1. Collecting data
The processing data are financial reports taken from public companies of the Indonesian Stock Exchange from 2008 - 2014. It consists of 100 solvent companies and 20 bankrupt companies. Table 1 is 21 parameters used to predict the condition of the company. The parameters used are the value of financial ratios. Financial ratios are commonly used by accountants to evaluate the condition of the company based on audited financial statements [3,13].

2.2. Analysis of Process
Data processed through the preprocessing stage. The following problems considered in preprocessing are the appearance of the missing value data on the observed feature values, the need to normalize and the need to standardize the data. Training and testing are carried out to get a prediction model. It uses the 5-fold cross-validation to test the model. The method used is Baseline Classification (ZeroR), Naive Bayes (NB), Support Vector Machine (SVM) using a linear kernel, Sequential Minimal Optimization (SMO), Logistic Regression (Logistic), K-nearest Neighbor (K-NN), Decision Tree (C45). It is giving a greater weighting to the cross-validation table for classes with fewer data to handle unbalanced data.

| Symbol | Feature                     | Symbol | Feature                           |
|--------|-----------------------------|--------|-----------------------------------|
| $x_1$  | Market Capitalization      | $x_{12}$ | Return on capital employed (ROCE) |
| $x_2$  | Price Earnings Ratio       | $x_{13}$ | Dividend Yield                   |
| $x_3$  | Price Per Sales            | $x_{14}$ | Dividend Payout Ratio             |
| $x_4$  | Price Per Book Value       | $x_{15}$ | Equity to Assets                  |
| $x_5$  | Price Per Free Cash Flow   | $x_{16}$ | Current Ratio                     |
| $x_6$  | Gross Profit Margin        | $x_{17}$ | Quick Ratio                       |
| $x_7$  | Operating Profit Margin    | $x_{18}$ | Cash Ratio                        |
| $x_8$  | Net Profit Margin          | $x_{19}$ | Inventory Turnover Ratio          |
| $x_9$  | Basic Earning Power        | $x_{20}$ | Leverage                          |
| $x_{10}$ | Return on Equity (ROE)     | $x_{21}$ | Assets Turnover                   |

Table 1. Financial Ratio
2.3. Learning and Testing
After obtaining the model and selecting the method, then the model accuracy will be tested. The data are splitting 60% of training and 40% of testing data. The testing measures the performance of the model.

2.4. Discussion
In this section, it is compared the research with the others. Furthermore, it is analyzed the features chosen from the selection results by comparing it with the theory of financial ratios commonly used in determining bankruptcy.

3. Results and Discussion
Weka 3.8 application used to measure the performance of the model. Meanwhile, Brownly [14] explains the steps of using Weka especially for binary classification on his website. Model testing uses all data and accuracy testing uses 5-fold cross-validation. Table 2 is the accuracy result of using seven machine-learning methods. Selection methods using statistical testing. The test results show that using ZeroR the value gains 83.33. This value is the baseline of the issuer's data used. From this table 2 can see the test result, SMO works superior compared to the other methods.

Table 2. The Testing Result of Several Machine Learning Methods

| Cases                | ZeroR | NB    | SVM    | SMO    | Logistic | K-NN | C45 |
|----------------------|-------|-------|--------|--------|----------|------|-----|
| Accuracy in Percentage (%) |
| 1 prior               | 83.33 | 82.83 | 83.33  | 83.67  | 82.08    | 82.75| 79.50|
| 1 prior normalization | 83.33 | 82.67 | 85.08  | 84.08  | 82.00    | 82.75| 79.33|
| 1 prior standardization | 83.33 | 83.17 | 82.42  | 83.83  | 82.08    | 82.75| 79.50|
| 1 prior missing value | 83.33 | 71.67 | 83.33  | 84.00  | 80.58    | 83.33| 75.58*|
| 2 priors              | 83.33 | 79.75 | 83.33  | 87.17  | 79.00*   | 84.00| 77.00*|
| 2 prior normalization | 83.33 | 79.67 | 85.83  | 87.25  | 79.00*   | 84.00| 76.25*|
| 2 prior standardization | 83.33 | 80.33 | 82.58  | 87.00  | 79.00*   | 84.00| 77.00*|
| 2 prior missing value | 83.33 | 76.42 | 83.33  | 83.83  | 77.00    | 83.33| 75.08*|

*The statistical test results state that the method is no better than SMO

The accuracy of 83.33%, the Baseline Classification method (ZeroR), is a reference to describe the other methods worked in classification [14]. Table 2, there are four conditions tested, process data without preprocessing, normalized data, standardized data, and processing of missing value data. Each model examined to predict bankruptcy one year and two years before. Naïve Bayes got the highest accuracy, 83.17%, in the 1-year before bankruptcy prediction and 80.33%, in the 2-years before bankruptcy prediction. This result corresponds to the theory which states Naïve Bayes requires the assumption that each feature must follow a standard normal distribution [15]. SVM using Radial Basis Function (RBF) as a kernel function, gamma = 1 and Sequential Minimal Optimization using Linear Kernel Functions show consistency in predicting bankruptcy with an accuracy value above 80%. Multiple Logistic Regression (Logistic) and Decision Tree (C45) produces an accuracy value that is no better than the ZeroR method. Using neighbors K = 3, The K-NN method shows good results only in predicting bankruptcy 2 years prior.

Based on Table 2, the results of the two-tails statistical test using the t-distribution that compares each method. The SMO method shows that the Logistics Method and C45 produce significant differences with SMO. The statistical hypothesis results, in Table 2, show that the accuracy of C45 (77.00%) is different significantly with SMO (87.00%) in predicting bankruptcy 2-years prior. Furthermore, for the feature selection process, it uses Chi-Squared, Pearson Correlation (PC), Gain Ratio (GR), and Information Gain (IG) methods. All data through the normalization process. Each feature ranked based on the calculated value of Chi-Squared, PC, GR, and IG. Seven of the highest
value features of each selection method are selected as features and will be used to obtain prediction models. Especially for the Pearson Correlation, all tested variables have a correlation value. Therefore, it was selected 7 of the variables with the highest values as features for the Pearson Correlation. Table 3 shows the details of the results. Although the ranking features in Chi-Squared, GR and IG are not the same, the seven features that appear are the same, except for the Pearson Correlation. Gross Profit Margin feature appears as a relevant feature.

Based on Havier and Qiyu’s research [16-17], ROA always appears as a feature to predict bankruptcy. Operating Profit Margin, Equity to Asset, Net Profit Margin are features that are also used by Havier, Qiyu and Tian’s research [16-18]. From table 3, it concludes that feature selection using Pearson Correlation is the best choice to use as a feature selection method.

### Table 3. The Result of Feature Selection

| Chi-squared | Pearson Correlation | Gain Ratio | Information Gain |
|-------------|---------------------|------------|------------------|
| x₁₀ ROE     | x₁₁ ROA             | x₁₁ ROA   | x₁₁ ROA          |
| x₁₁ ROA     | x₉ Basic E. Power   | x₆ Basic E. Power | x₂ PER |
| x₅ Basic E. Power | x₇ Operating P. M. | x₅ Operating P. M. | x₈ Net Profit Margin |
| x₇ Operating P. M. | x₁₅ Equity to Assets | x₁₅ Equity to Assets | x₁₁ ROA |
| x₂ PER      | x₁₀ ROE             | x₇ Operating P. M. | x₅ Basic E. Power |
| x₁₅ Equity to Assets | x₆ Gross Profit M. | x₈ Net Profit Margin | x₁₀ ROE |
| x₈ Net Profit Margin | x₅ Gross Profit M. | x₂ PER |
| x₁₅ Equity to Assets | x₆ Gross Profit M. | x₈ Net Profit Margin | x₁₁ ROA |

Furthermore, the performance of the models is measured using 21 features, eight features, and seven features. For performance testing, it uses the SVM method. The weighting process is carried out in the cross-validation table to overcome unbalanced data sets. The first testing uses seven selected features from the Chi-Squared, GR and IG methods. While eight features testing adds the X6 variable (Gross Profit Margin). Table 4 and table 5 use splitting datasets, 60%, and 40%. The tested data consisted of 48 company data, 40 non-bankrupt companies, 8 bankrupt companies. Table 4 and table 5 use splitting datasets, 60%, and 40%. The tested data consisted of 48 company data, 40 non-bankrupt companies, 8 bankrupt companies. In table 4 and table 5, the solvent TP represents the percentage of solvent companies and those detected correctly not bankrupt (True Positive). True Negative (TN) of bankruptcy is a percentage of bankrupt companies that successfully detected.

The use of feature selection in the prediction of 1-year bankruptcy can increase the accuracy produced. However, for predicting the accuracy of bankruptcy accuracy of 2 years before, the highest accuracy value obtains 93.8%, and it happens using all existing features. The linear kernel function is suitable for predicting bankruptcy one year before while for predicting bankruptcy two years prior uses the Radial Basis Function (RBF) as a kernel function. Although the unbalanced data set tried to overcome by weighting the cross-validation table, but this method only succeeded in detecting bankrupt issuers at a maximum of 75%. Table 4 shows that the best result is achieved using the SMO method, 7 features. The best method to predicted 2 years before bankruptcy is the SMO method, 21 features. This result is consistent with Alaka’s research [19]; which states that the SMO as a method suitable for use in case of data sets with small amounts. Additionally, SMO can be implemented along with the existing filtering feature selection method. Although Lin [20] recommends using the RBF function on SMO, this result only applies to predictions of bankruptcy 2- years before, while predicting 1-year bankruptcy it simply uses linear kernel functions. The possibility of this happens because for the prediction of 1-year bankruptcy has a shorter time so that the factors that determine bankruptcy are fewer. Furthermore, for the prediction of one year of bankruptcy, the characteristics of the company predicted to have a tendency can separated linearly. As can see at Table 4 and 5.
Table 4. The Accuracy of Bankruptcy Prediction 1 Year Prior

| 1 prior | 21 Feature | 8 Feature | 7 Feature |
|---------|------------|-----------|-----------|
|         | SVM(RBF)   | SMO(Lin)  | SVM(RBF)| SMO(Lin) | SVM(RBF) | SMO(Lin) |
| TP Solvent (%) | 100 | 97.5 | 100 | 100 | 100 | 100 |
| TP Bankruptcy (%) | 50 | 37.5 | 37.5 | 50 | 37.5 | 50 |
| Accuracy (%) | 91.57 | 87.5 | 89.6 | 91.57 | 89.6 | 91.57 |

Table 5. The Accuracy of Bankruptcy Prediction 2 Year Prior

| 2 priors | 21 Feature | 8 Feature | 7 Feature |
|----------|------------|-----------|-----------|
|          | SVM(RBF)   | SMO(Lin)  | SVM(RBF)| SMO(Lin) | SVM(RBF) | SMO(Lin) |
| TP Solvent (%) | 100 | 97.5 | 100 | 90 | 90 | 95 |
| TP Bankruptcy (%) | 50 | 75 | 62.5 | 50 | 62.5 | 37.5 |
| Accuracy (%) | 91.57 | 93.8 | 85.4 | 83.33 | 85.4 | 85.4 |

4. Conclusion

This research has shown that the best method used in bankruptcy prediction cases for company data from the Indonesian Stock Exchange is Sequential Minimal Optimization (SMO) method. The issuer's data used needs to normalize first. The optimal prediction of 1-year before uses feature selection and linear kernel functions. The best way to predict 2-years before is not using feature selection and using the RBF function. There are 7 ratios influence the determination effect of 1-year before bankruptcy. Those are PER, Operating Profit, Net Profit Margin, Basic Earning Power, ROE, ROA, Equity Assets. For the problem of unbalanced data sets, it still needs to be reviewed using different techniques; one of them is by using methods to detect outliers by assuming that the issuer is bankrupt as an outlier.

References

[1] Altman E I 1968 Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy The J. of Finance 23 589-609
[2] Hadad M D, Santos W and Rulina I 2003 Indikator Kepailitan di Indonesia: An Additional Early Warning Tools Pada Stabilitas Sistem Keuangan (Jakarta: Bank Indonesia)
[3] Barboza F, Kimura H and Altman E 2017 Machine Learning Models and Bankruptcy Prediction Expert Systems with Application 83 405-17
[4] Klepáč V and Hampel D 2016 Prediction of Bankruptcy with SVM Classifiers among Retail Business Companies in EU Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 64 627-34
[5] Gogas P, Papadimitriou T and Agrapetidou A 2018 Forecasting Bank Failures and Stress Testing: A Machine Learning Approach Int. J. of Forecasting 34 440-55
[6] Antunes F, Ribeiro B and Pereira F 2017 Probabilistic Modeling and Visualization for Bankruptcy Prediction Application Soft Computing 60 831-43
[7] Tsai C F, Hsu Y F and Yen D C 2014 A Comparative Study of Classifier Ensembles for Bankruptcy Prediction Application Soft Computing 24 977-84
[8] Veganzones D and Séverin E 2018 An Investigation of Bankruptcy Prediction in Imbalanced Datasets Decision Support System 112 111-124
[9] Le T, Lee M and Park J, Baik S 2018 Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset Symmetry 10 79
[10] Wang G, Ma J and Yang S 2014 An Improved Boosting based on Feature Selection for Corporate Bankruptcy Prediction Expert System with Application 41 2353-61
[11] Liang D, Tsai C F and Wu H T 2015 The Effect of Feature Selection on Financial Distress Prediction Knowledge-Based System 73 289-97
[12] Dellepiane U, Di Marcantonio M, Laghi E and Renzi S 2015 Bankruptcy Prediction using Support Vector Machines and Feature Selection during the Recent Financial Crisis *Int. J. of Economics and Finance* **7** 182-96

[13] Song Y G, Cao Q L and Zhang C 2018 Towards a New Approach to Predict Business Performance using Machine Learning *Cognitive System Research* **52** 1004-12

[14] Brownlee J 2014 Machine Learning Mastery URL: http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it.

[15] Rish I 2001 An Empirical Study of The Naive Bayes Classifier. *IJCAI Workshop on empirical Methods in Artificial Intelligence* **3** pp 41-46

[16] de Andrés J, Landajo M and Lorca P 2012 Bankruptcy Prediction Models based on Multinorm Analysis: An Alternative to Accounting Ratios *Knowledge-Based Systems* **30** 67-77

[17] Yu Q, Miche Y, Séverin E and Lendasse A, 2014 Bankruptcy Prediction using Extreme Learning Machine and Financial Expertise *Neurocomputing* **128** 296-302

[18] Tian S and Yu Y 2017 Financial Ratios and Bankruptcy Predictions: An International Evidence *Int. Review of Economics and Finance* **51** 510-26

[19] Alaka H A, Oyedele L O, Owolabi H A, Kumar V, Ajayi S O, Akinade O O and Bilal M 2018 Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection *Expert Systems with Applications* **94** 164-84

[20] Lin H and Lin C 2003 A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods *Neural Computation* **2** 1-32