Jurnal Teknologi

Full paper

Corporate Default Prediction with AdaBoost and Bagging Classifiers

Suresh Ramakrishnan[a], Maryam Mirzaei[a*], Mahmoud Bekri[b]

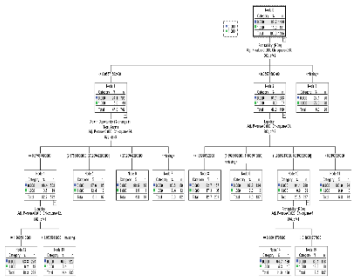
[a]*Finance and Accounting Department, Faculty of Management, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[b]*Economic and Statistic Institute, Karlsruhe Institute of Technology, Germany*

*Corresponding author: mmirzai72@yahoo.ca

Article history

Received :10 December 2014
Received in revised form :
1 February 2015
Accepted :12 February 2015

Graphical abstract



Abstract

This study aims to show a substitute technique to corporate default prediction. Data mining techniques have been extensively applied for this task, due to its ability to notice non-linear relationships and show a good performance in presence of noisy information, as it usually happens in corporate default prediction problems. In spite of several progressive methods that have widely been proposed, this area of research is not out dated and still needs further examination. In this paper, the performance of ensemble classifier systems is assessed in terms of their capability to appropriately classify default and non-default Malaysian firms listed in Bursa Malaysia. AdaBoost and Bagging are novel ensemble learning algorithms that construct the base classifiers in sequence using different versions of the training data set. In this paper, we compare the prediction accuracy of both techniques and single classifiers on a set of Malaysian firms, considering the usual predicting variables such as financial ratios. We show that our approach decreases the generalization error by about thirty percent with respect to the error produced with a single classifier.

Keywords: Default prediction; Adaboost; bagging; data mining

1.0 INTRODUCTION

Prediction of corporate default is an important science problem and its main aim is to differentiate those firms with a high probability of default from healthy firms. Due to the significant consequences which default imposes on different groups of society as well the noteworthy troubles qualified by firms during the Global Financial Crisis, the crucial importance of measuring and providing for credit risk have highlighted. Since the mid-1990s, there has been growing concern in emerging and developing economies among researchers. Regarding the growth in financial services, there have been swelling sufferers from off ending loans. Therefore, default risk forecasting is a critical part of a financial institution's loan approval decision processes. Default risk prediction is a procedure that determines how likely applicants are to default with their repayments. Review of literature on the subject confirmed hand full of studies conducted in the last four decades. Despite of these studies, the recent credit crisis indicated that yet there are areas of the study that needs researchers' attention. Moreover, emerging of the regulatory changes such as Basel III accord and the need for more precise and comprehensive risk management procedures justifies need of research in area of credit risk modeling and banking supervision. This requirement like these pushes companies especially banks and insurance companies to have a very robust and transparent risk management system.

As a valuable implement for scientific decision making, corporate default prediction takes an imperative role in the prevention of corporate default. From this point of view, the accuracy of default prediction model is an essential issue, and many researchers have focused on how to build efficient models. In supervised classification tasks, the mixture or ensemble of classifiers represent a remarkable method of merging information that can present a superior accuracy than each individual method. To improve model accuracy, classifier ensemble is a capable technique for default prediction. In fact, the high classification accuracy performance of these combined techniques makes them appropriate in terms of real world applications, such as default prediction. However, research on ensemble methods for default prediction just begins recently, and warrants to be considered comprehensively.

Former researches on ensemble classifier for default prediction used DT or NN as base learner, and were both compared to single NN classifier. This paper further explores AdaBoost and Bagging ensemble for default prediction to compare with various baseline classifiers including learning logistic regression (LR), decision tree (DT), artificial neural networks (NN) and support vector machine (SVM) as base learner.

2.0 LITERATURE REVIEW

The majority of the discussion related to default prediction develops around the decisive works of Altman (1968), Ohlson (1980), Zmijewski (1984) and recently Shumway (2001). It was Altman (1968) who applied Multivariate Discriminant Analysis (MDA) for the first time to classify failed and non-failed U.S firms. Researchers still use his model as a benchmark to predict firm default. Altman's Z-score model is a linear analysis of five ratios and this score is a basis for firm classification. Besides this, Blum (1974) employed the same MDA technique for default prediction some years prior to failure. Similarly, to assess the predictive accuracy of accounting ratios, Libby (1975) measured the prediction achievement of a selected set of accounting ratios for U.S firms. In the past literature, there are numerous studies which applied this method as a benchmark for default prediction in U.S. firms such as, (Deakin, 1972; Altman, 1973; Benishay, 1973; Blum, 1977; Norton and Smith, 1979; Rose *et al*., 1982; Hennawy and Morris, 1983; Taffler, 1984; Gilbert, Menon, and Schwartz, 1990; Goudie and Meeks, 1991; Hellegiest, 2004).

Consistent with above spat of discussion, the literature on credit risk mainly tended to focus on developed countries. For instance, Goudie and Meeks, 1991; Vassalou and Yuhang, 2004; Liou and Smith, 2007;Chen, *et al*., 2011; Shiyi, *et al*., 2011. Despite the emerging need of credit risk modeling in the purview of global and regional economic shocks and financial turmoil, this area relatively remained less explored in emerging and developing markets. Though a few studies highlighted the connotation of default prediction. For example, Sandin and Porporato, 2007; Yap, *et al*., 2010; Yildiz and Akkoc, 2010. However, the significance of credit risk modeling remained untapped in emerging and developing markets. In purview of this strand of arguments, the subsequent section provides a comparative overview of emerging and developing economies.

Keeping in view the level of economic and industrial development, the economists have classified countries around the world as developed, emerging and developing markets (Economy Watch, 2010. According to World Bank (2010), the emerging economy was a term coined by economist Agtmael (1981) in reference to nations undergoing rapid economic development and industrialization. Moreover, the emerging economies involve policy and structure reforms and capital market development. On the other hand, the developing economy is a market with underdeveloped industrial base, less developed banking sector and capital market, and has low Gross National Income (GNI) per capita relative to emerging markets (World Bank Development Indicators, 2012). The fundamental difference is that emerging economies are growing rapidly and becoming more important on the world economic stage, while developing markets struggle in comparison and still need help from trade partners around the world. Furthermore, the emerging economies differ from developing countries in that they have made impressive gains in infrastructure and industrial growth, and are experiencing increasing incomes and quick economic growth (Economy Watch, 2010).

The lack of a unified theory on corporate default has meant that most studies dealing with default prediction have focused on increasing the accuracy of the model. This is clearly important in default prediction as firm must make appropriate decisions. The literature shows that no studies have been made in order to default prediction using ensemble classifiers in Malaysia. Therefore, this study attempts to improve default prediction model accuracy using Adaboost and Bagging classifiers.

3.0 METHODOLOGY

3.1 Framework of Ensemble Method

i. Adabbost

The key idea of multiple classifier systems is to employ ensemble of classifiers and combine them in various approaches. Theoretically, in an ensemble of N independent classifiers with uncorrelated error areas, the error of an overall classifier obtained by simply averaging/voting their output can be reduced by a factor of N. Boosting is a meta-learning algorithm and the most broadly used ensemble method and one of the most powerful learning ideas introduced in the last twenty years. The original boosting algorithm has been proposed by Robert Schapire (a recursive majority gate formulation and Yoav Freund (boost by majority) in 1990. In this type, each new classifier is trained on a data set in which samples misclassified by the previous model are given more weight while samples that are classified correctly are given less weight. Classifiers are weighted according to their accuracy and outputs are combined using a voting representation. The most popular boosting algorithm is Adaboost (Freund and Schapire, 1997). Adaboost applies the classification system repeatedly to the training data, but at each application, the learning attention is focused on different examples of this set using adaptive weights (ωb(i)). Once the training procedure has completed, the single classifiers are combined to a final, highly accurate classifier based on the training set. A training set is given by:

Tn= {(X1,Y1), (X2,Y2),..,(Xn,Yn)}

Where y takes values of {-1,1}. The weight ωb(i) is allocated to each observation Xi and is initially set to 1/n. This value will be updated after each step. A basic classifier denoted Cb(Xi) is built on this new training set, Tb, and is applied to each training sample. The framework of Adaboost algorithm, weak learning algorithm and combination mechanism for default prediction is shown in Figure 1.

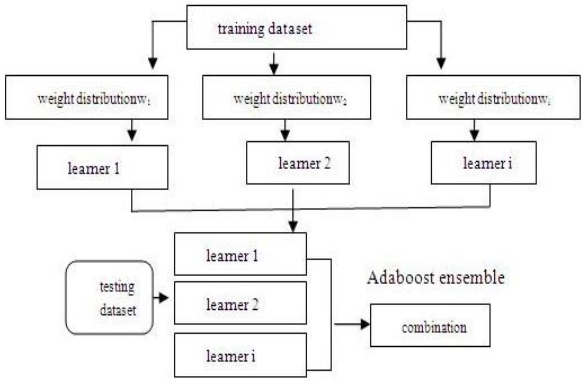


Figure 1 The framework of Adaboost algorithm

ii. Bagging

Bagging is an also meta algorithm that pool decisions from multiple classifiers. In bagging we train k models on different sample (data splits) and average their predictions. Then, we predict the test set by averaging the results of k models. The bagging algorithm can be described as follow:

- Training

In each iteration t, t=1,...T

• Randomly sample with replacement N samples from the training set

• Train a chosen "base model" (e.g. neural network, decision tree) on the samples.

- Test

For each test example

• Start all trained base models

• Predict by combining results of all T trained models:

– Regression: averaging

- Classification: a majority vote.

3.2 Single Classifiers

3.2.1 Logistic Regression

Logistic regression is a type of regression methods (Allison, 2001; Hosmer & Lemeshow, 2000) where the dependent variable is discrete or categorical, for instance, default (1) and non-default (0). Logistic regression examines the effect of multiple independent variables to forecast the association between them and dependent variable categories. According to Morris (1997), Martin (1977) was the first researcher who used logistic technique in corporate default perspective. He employed this technique to examine failures in the U.S. banking sector. Subsequently, Ohlson (1980) applied logistic regression more generally to a sample of 105 bankrupt firm and 2,000 non-bankrupt companies. His model did not discriminate between failed and non-failed companies as well as the multiple discriminant analysis (MDA) models reported in previous studies. According to Dimitras, *et al*. (1996), logistic regression is in the second place, after MDA, in default prediction models.

3.2.2 Decision Tree

Decision trees are the most popular and powerful techniques for classification and prediction. The foremost cause behind their recognition is their simplicity and transparency, and consequently relative improvement in terms of interpretability. Decision tree is a non-parametric and introductory technique, which is capable to learn from examples by a procedure of simplification. Frydman, Altman, and Kao (1985) first time employed decision trees to forecast default. Soon after, some researchers applied this technique to predict default and bankruptcy including (Carter & Catlett, 1987; Gepp, Kumar, & Bhattacharya, 2010; Messier & Hansen, 1988; Pompe & Feelders, 1997).

3.2.3 Neural Network

Neural networks (NNs), usually non-parametric techniques have been used for a variety of classification and regression problems. They are characterized by associates among a very large number of simple computing processors or elements (neurons). Corporate default have predicted using neural networks in early 1990s and since then more researchers have used this model to predict default. As a result, there are some main profitable loan default prediction products which are based on neural network models. Also, there are different evidence from many banks which have already expanded or in the procedure of developing default prediction models using neural network (Atiya, 2001). This technique is flexible to the data characteristics and can deal with different non-linear functions and parameters also compound prototypes. Therefore, neural networks have the ability to deal with missing or incomplete data (Smith & Stulz, 1985).

3.2.4 Support Vector Machine

Among different classification techniques, Support Vector Machines are considered as the best classification tools accessible nowadays. There are a number of empirical results attained on a diversity of classification (and regression) tasks complement the highly appreciated theoretical properties of SVMs. A support vector machine (SVM) produces a binary classifier, the so-called optimal separating hyper planes, through extremely nonlinear mapping the input vectors into the high-dimensional feature space. SVM constructs linear model to estimate the decision function using non-linear class boundaries based on support vectors. Support vector machine is based on a linear model with a kernel function to implement non-linear class boundaries by mapping input vectors non-linearly into a high-dimensional feature space. The basic idea of the SVM classification is to find such a separating hyperplane that corresponds to the largest possible margin between the points of different classes (Figure 2).

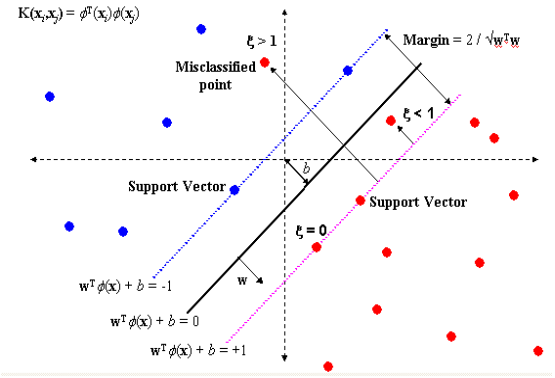


Figure 2 The SVM learns a hyperplane which best separates the two classes

■4.0 Empirical Results

4.1 Data Description

The dataset was used to classify a set of firms into those that would default and those that would not default on loan payments. It consists of 285 observations of Malaysian companies. Of the 285 cases for training, 121 belong to the default case under the requirements of PN4, PN17 and Amended PN17 respectively and the other 164 to non-default case. Consulting an extensive review of existing literature on corporate default models, the most common financial ratios that are examined by various studies were identified. The variable selection procedure should be largely based on the existing theory. The field of default prediction, however, suffers from a lack of agreement as for which variables should be used. The first step in this empirical search for the best model is therefore the correlation analysis. If high correlation is detected, the most commonly used and best performing ratios in the literature are prioritized. Therefore, the

choice of variables entering the models is made by looking at the significance of ratios.

The components of the financial ratios which are estimated from data are explained below and Table 1 shows the summary statistics for selected variables for default and non-default firms. To select the variables, two approaches including linear regression and decision tree analysis were used. The most significant variables based on two methods were identified. These variables selected from the significant indicators for the model which could best discriminate the default firms from the non-default firms. These selected financial ratios include: Profitability ratios, liquidity and growth opportunity (Figure 3).

Table 1 The summary statistics for selected variables for default and non-default firms

	Definition of variable	Means of non-default companies	Means of default companies	Test of equality of group means
1	EBIT /Total Assets	0.155647	-0.02608	0
2	Cash/Total Assets	0.046677	-0.191281	0.137
3	Current asset/Current Liabilities	1.854502	1.178482	0
4	size	6.18186	5.84403	0.271
5	Total Current Liabilities to Total Assets	0.151896	-0.05514	0
6	Net profit/Net sale	0.098689	-0.87107	0.174
7	Growth Opportunity	0.095607	-0.06198	0
8	Net Profit/Total Assets	0.095607	-0.06197	0

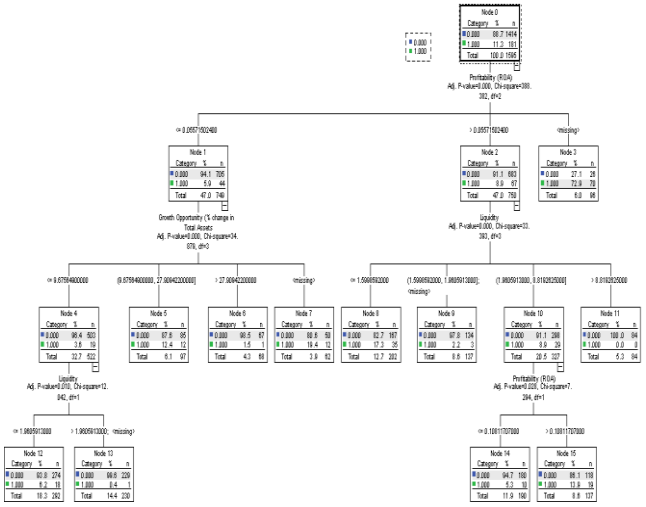


Figure 3 Decision tree

4.2 Results and Discussion

In this experiment study, the main goal is to compare ensemble classifiers. To obtain comparable experimental results, the same default prediction problem is solved by four different classification methods, i.e. Adaboost ensemble with logistic regression (represented as LA), AdaBoost ensemble with decision tree (represented as DTA), single classifier on DT and single classifier on SVM.

The results are presented in two parts. First part of this section displays the percent of accuracy rate for each classifier system. Then, the enhancement over the baselines has been shown for ensemble classifiers. Table 2 shows the percent of model accuracy and the area under ROC curve for each classifier system. Comparison of forecasting accuracy reveals that the SVM has a lower model risk than other models. According to the results, SVM is the best. The performance of Neural Network is significantly worse than other approaches. Generally, the findings for the baseline classifiers are not predominantly unexpected and are well-matched with previous empirical researches of classifier performance for default risk data sets especially in case of SVM classifier. SVM with a high generalization capacity seems to be a capable technique for default prediction in Malaysia as an emerging economy. Also, Table 2 shows the performance accuracy of ensemble classifiers in compare with baselines. The ensemble classifiers considerably outperform the baseline. By the results, all ensemble systems outperform the baseline including Adaboost with logistic regression, and neural network, decision tree and support vector machines. The results also state the improvement by the bagging is significant, which ensembles using neural network showing the major improvement. Roc curve plots the type II error against one minus the type I error. In the case of default prediction in this study, it describes the percentage of non-defaulting firms that must be inadvertently denied credit (Type II) in order to avoid lending to a specific percentage of defaulting firms (1- Type I) when using a specific model. Figure 4, shows the ROC curve for baseline, Adaboost and Bagging classifiers.

5.0 CONCLUSION

Due to the importance of default and its impact on different parties of the companies, the issue of default prediction has received extensive attention of researchers. Appropriate identification of firms 'approaching default is undeniably required. By this time, various methods have been used for predicting default. The use of ensemble classifiers has become common in many fields in the last few years. According to various studies, diverse individual classifiers make errors on different instances (Polikar, 2006; Rokach, 2010). The variety is supposed to improve classification accuracy. According to (Brown, Wyatt, Harris, & Yao, 2005; Rokach, 2010), diversity creation can be obtained in several ways, and the approaches to classify them vary. The selection of a particular technique can have important consequences on the data analysis and subsequent interpretation of findings in models of credit risk prediction, especially if the quality of data is not good. This paper focused on corporate default prediction; the approach is differentiated one in accordance with employing Ensemble classifiers for Malaysian firms as a developing economy. The accuracy of five classifiers was assessed to figure out whether it is conceivable to forecast default of Malaysian firms based on financial ratios. Empirical results highlighted out the financial ratio in three groups including profitability, liquidity and growth opportunity.

In this study, two ensemble classifiers have been compared, showing the improvement in accuracy that Adaboost and Bagging achieve against single classifiers. As has been seen, AdaBoost is based on building consecutive classifiers on modified versions of the training set which are generated according to the error rate of the previous classifier. The practical application has worked with two classes, where failed companies have been distinguished from healthy companies. The results show that Adaboost and Bagging achieve reduction in the test error compared with the individual classifiers.

Table 2 Performance of classifier systems

	Classifier system	% Accuracy	% ROC Area
Baseline models	LR	74.39	70.2
	NN	58.2	54.8
	DT	69.12	61.2
	SVM	83.41	71.9
Ensemble	LR (adaboost)	86.17	71.2
	NN (adabboost)	70.5	71.9
	DT (adaboost)	82.94	83.4
	SVM (adaboost)	86.83	82.9
	LR (bagging)	86.17	91.4
	NN (bagging)	82.94	92.8
	DT (bagging)	83.87	91.6
	SVM (bagging)	86.83	94.3

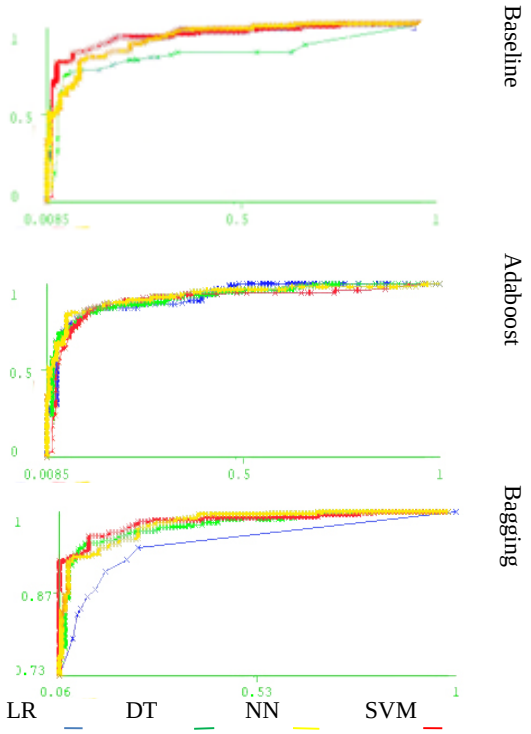


Figure 4 Performance of adaboost and bagging

References

[1] J. Abellán, J., & Masegosa, A. 2012. Bagging Schemes on the Presence of Noise in Classification. *Expert Systems with Applications*. 39(8): 6827–6837.

[2] Allison, P. D. 2001. *Logistic Regression Using the SAS System: Theory and Application*. Cary, NC: SAS Publishing, BBU Press.

[3] Altman, Edward I. 1968. Financial Ratios, Discriminant Analaysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*. 23(4): 589–609.

[4] Altman, Edward I. 1973. Predicting Railroad Bankruptcies in America. *Bell Journal of Economics & Management Science*. 4(1): 184.

[5] Atiya, A. 2001. Bankruptcy Prediction for Credit Risk using Neural Networks: A Survey and New Results. *IEEE Transactions on Neural Networks*. 12: 929–935.

[6] Beaver, William H. 1966. Financial Ratios as Predictors of Failure. *Journal of Accounting Research*. 4(3): 71–111.

[7] Brown, G., Wyatt, J.L. Harris, R. Yao, X. 2005. Diversity Creation Methods: A Survey and Categorization, Information Fusion. 6 (1): 5–20.

[8] Carter, C., & Catlett, J. 1987. Assessing Credit Card Applications Using Machine Learning. *IEEE Expert*. 2: 71–79.

[9] Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. 1996. A Survey of Business Failure with Anemphasis on Prediction Methods andindustrialapplication. *European Journal of Operational Research*. 90: 487–513.

[10] Freund Y. and Schapire R. E. 1997. A Decision-theoretic Generalisation of On-Line Learning and an Application to Boosting. *J. of Computer and System Science*. 55(1): 119–139.

[11] Frydman, H., Altman, E., & Kao, D. 1985. Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. *Journal of Finance*. 269–291.

[12] Gepp, Adrian, Kumar, Kuldeep, & Bhattacharya, Sukanto. 2010. Business Failure Prediction Using Decision Trees. *Journal of Forecasting*. 29(6): 536–555. doi: 10.1002/for.1153.

[13] Gestel, T. V., Baesens, B., Dijcke, P. V., Suykens, J., Garcia, J., & Alderweireld, T. 2005. Linear and Nonlinear Credit Scoring by Combining Logistic Regression and Support Vector Machines. *Journal of Credit Risk*. 1(4): 31–60.

[14] Härdle, Wolfgang, Moro, Rouslan, & Schäfer, Dorothea. 2005. Predicting Bankruptcy with Support Vector Machines. *Statistical Tools in Finance & Insurance*. 225–248.

[15] Hertz, J., Krogh, A., & Palmer, R.G. 1991. *The Theory of Neural Network Computation*. Addison Welsey: Redwood, CA.

[16] Hosmer, D. W., & Lemeshow, S. 2000. *Applied Logistic Regression*. New York: Wiley.

[17] Jackendoff, N. 1962. A Study of Published Industry Financial and Operating Ratios. Philadelphia: Temple University, Bureau of Economic and Business Research.

[18] Libby, Robert. 1975. Accounting Ratios and the Prediction of Failure: Some Behavioral Evidence. *Journal of Accounting Research*. 13(1): 150–161.

[19] Lin, F., and McClean, S. 2001. A Data Mining Approach to the Prediction of Corporate Failure-Knowledge-Based Systems. 14(3–4): 189–195.

[20] Lin, Y., 2002. Improvement on Behavioural Scores by Dual-model Scoring System. *International Journal of Information Technology and Decision Making*. 1: 153–165.

[21] Martin, D. 1977. Early Warnings of Bank Failure: A Logit Regression Approach. *Journal of Banking and Finance*. 1: 249–276.

[22] Messier, JR. W., & Hansen, J. 1988. Inducing Rules for Expert System Development: An Example Using Default and Bankruptcy Data. Management Science. 34: 1403–1415.

[23] Morris, Richard. 1997. Early Warning Indicators of Corporate Failure: A Critical Review of Previous Research and Further Mpirical Evidence. Ashgate.

[24] Myers, J. H., & Forgy, E. W. 1963. The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association*. 58(303): 799–806.

[25] Ohlson, James A. 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*. 18(1).

[26] Pompe, P., & Feelders, A. 1997. Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy. *Microcomputers in Civil Engineering*. 12: 267–276.

[27] Polikar, R. 2006. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*. 6(3): 21–45.

[28] Quinlan, J. R. 1986. *Induction of Decision Trees, Machine Learning*. 1: 81–106.

[29] Ravi Kumar, P., & Ravi, V. 2007. Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Techniques-A Review. *European Journal of Operational Research*. 180(1): 1–28.

[30] Rokach, L. 2010. *Ensemble Methods in Supervised Learning. Data Mining and Knowledge Discovery Handbook*. 959–979.

[31] Shin, K. S., and Lee, Y. J. 2002. A Genetic Algorithm Application in Bankruptcy Prediction Modeling. Expert Systems with Applications. 9: 503–512.

[32] Shin, K. S., lee, T. S., & Kim, H. J. 2005. An Application of Support Vector Machines in Bankruptcy Prediction Model. *Expert system Application*. 28(1): 127–135.

[33] Shumway, Tyler. 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *Journal of Business*. 74(1): 101–124.
[34] Smith, C. W., & Stulz, R. M. 1985. The Determinants of Firms' Hedging Policies. *Journal of Financial and Quantitative Analysis*. 20(4): 391–405.
[35] Smith, R., & Winakor, A. 1935. Changes in Financial Structure of Unsuccessful Industrial Corporations. Bureau of Business Research. Bulletin Urbna University of Illinois Press. 51.
[36] West, D., Dellana, S., Qian, J. 2005. Neural Network Ensemble Strategies for Financial Decision Applications. *Computers and Operations Research*. 32: 2543–2559.
[37] Zhu, H., Beling, P.A., Overstreet, G. 2001. A Study in the Combination of Twoconsumer Credit Scores. *Journal of the Operational Research Society*. 52: 974–980.
[38] Zmijewski, M. E. 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*. 22: 59–86.