Research Paper Management

Corporate Default Prediction with Data Mining Techniques

Suresh Ramakrishnan	Department of Management, Universiti Teknologi Malaysia
Maryam Mirzaei	Department of Management, Universiti Teknologi Malaysia
Hishan Shanker	Department of Management, Universiti Teknologi Malaysia

ABSTRACT

Default has recently upraised as an excessive concern due to the recent world crisis. Early forecasting of firms default provides decision-support information for financial and regulatory institutions. 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 different multiple classifier systems are assessed in terms of their capability to appropriately classify default and non-default Iranian firms listed in Tehran Stock Exchange (TSE). On the other hand, TSE have had very high return which provided more than 140 percent return in last year. For this reason, TSE could be more attractive for investors. Most multi-stage combination classifiers provided significant improvements over the single classifiers. In addition, Adaboost afford enhancement in performance over the single classifiers.

KEYWORDS Data Mining, Default Prediction, Classifier Combination

Introduction

Due to the significant consequences which default imposes on different groups of society and the problems faced by firms during the Global Financial Crisis the importance of measuring and providing credit risk has increased. Since the mid-1990s, this has been a growing concern in emerging and developing economies among researchers. One of the least studied emerging markets is the Tehran Stock Exchange (TSE), however a study of the TSE would contribute to the literature on emerging and developing market's finance especially in the Middle East. The value of Tehran Stock Exchange return was increased by 140 percent at the end of 2013. Regarding the growth in financial services, there has been an increase in the number of 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 the applicants are likely to default with their repayments. Review of literature on the subject confirmed that some studies were conducted in the last four decades. However in spite of these studies, the recent credit crisis indicated that there are certain areas of the study that needs researchers' attention. Moreover, after the regulatory changes such as Basel III accord have emphasized the need for more precise and comprehensive risk management procedures. This justifies the need for research in the area of credit risk modeling and banking supervision. The requirements 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 Related works

There has been significant advancement in the past few decades in terms of methodologies used for default prediction. Beaver (1966) introduced the Naïve Bayes approach using a single variable and Altman in 1968 suggested the use of Linear Discriminant Analysis (LDA). Since then several contributions have been made to improve the Altman's results, using different techniques. The use of data mining techniques such as Artificial Neural Networks (ANN), decision trees, and Support Vector Machine (SVM) for bankruptcy prediction started in the late 1980s (Pompe & Feelders, 1997; Shin, lee, & Kim, 2005).

Frydman et al. (1985) used Decision Trees first time for default prediction. Using this model, they classified firms to failed and non-failed based on firm-level and country-level factors. According to their results, this technique allows for an easy identification of the most significant characteristics in default prediction. In another study, Quinlan (1986) noted that decision trees method can deal with noise or non-systematic errors in the values of features. There are some other studies which predicted default using this method such as, (Messier & Hansen, 1988; Pompe & Feelders, 1997). Detailed examination of corporate default prediction by Lin and McClean (2001) showed a better performance of the hybrid model. They used four different techniques to predict corporate default, which two of the methods were statistical and the outstanding two models were machine learning techniques. In different but related work, Shin and Lee (2002) suggested a model using genetic algorithms technique. Some other related studies have employed Artificial Neural Networks to predict default.

Artificial Neural Networks was first demonstrated experimentally by Hertz, Krogh, and Palmer (1991) to analyze bankrupt

companies. Since then the method became a common accuracy amongst. Recently, some of the main commercial loan default prediction products applied ANN technique. For example, Moody's public firm risk model ANN and many banks and financial institutions have developed this method for default prediction (Atiya, 2001). More recently, the support vector machine was commenced for default risk investigation. This technique which is based on statistical learning theory compared with the traditional methods is more accurate in predicting default likelihood (Härdle, Moro, & Schäfer, 2005). In a major study on default prediction Gestel et al. (2005) employed SVM and logistic regression. The results based on combination of both techniques showed more stability in prediction power which is necessary for rating banks.

The limited research undertaken into the application of classifier combination to default prediction problems has arguably generated better results. In this regard, Myers and Forgy (1963) implemented a multi-stage methodology in which they employed a two stage discriminant analysis model. The second stage model was constructed using the lowest scoring of the development sample used in the first stage. They reported that the second stage model identified 70% more bad cases than the first stage model. In another study, Lin (2002) conveyed up to 3% improvement when employing a logistic model, followed by a neural network. There has been relatively little research effort to compare different classification methodologies within the credit risk area. Only in the study by West et al. (2005) was more than a single combination strategy given consideration, and in this case only one type of classifier which is neural network have been employed. In another study, Abellán and Masegosa (2012) showed that using Bagging ensembles on a special type of decision trees, called credal decision trees (CDTs), provides an appealing tool for the classification task.

3 Methods

3.1 Framework of ensemble method

i. Adaboost

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:

$$T_n = \{(X_1,Y_1), (X_2,Y_2),..,(X_n,Y_n)\}$$

Where y takes values of {-1,1}. The weight $\omega_b(i)$ 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 $C_b(Xi)$ is built on this new training set, Tb, and is applied to each training sample. The error of this classifier is represented by ξ_b and is calculated as:

$$\xi_{b=} \sum \omega_b(i)\ \xi_b(i)$$

Where

$$\xi_b(i) = \begin{cases} 0 & C_b(X_i) = y_i \\ 1 & C_b(X_i) \neq y_i \end{cases}$$

The new weight for the (b+1)-th iteration will be,

$$\omega_{b+1}(i) = \omega_b(i) \cdot \exp(\alpha_b \xi_b(i))$$

where α_t is a constant calculated from the error of the classifier in the b-th iteration. This process is repeated in every step for b=1,2,3,..,B. Finally, the ensemble classifier is built as a linear combination of the single classifiers weighted by the corresponding constant α_b. The framework of Adaboost algorithm, weak learning algorithm and combination mechanism for default prediction is shown in figure 1.

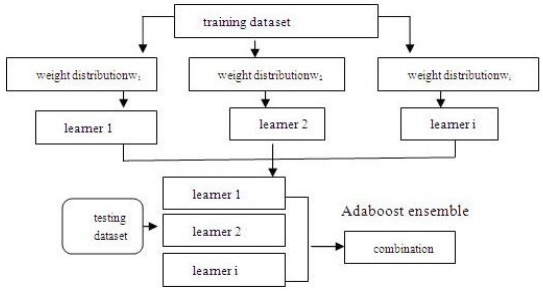

Figure 1: The framework of Adaboost algorithm

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 Supervised learners

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 Networks

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).

3.4 Support Vector Machines

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.

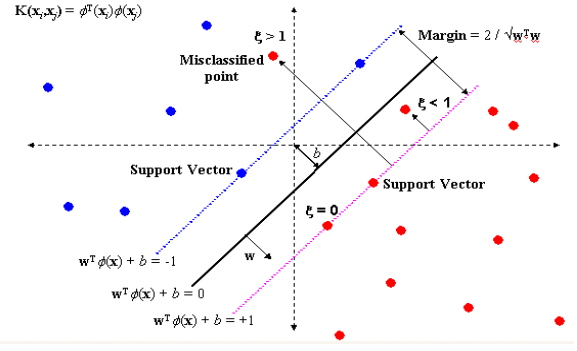

Figure 2: The SVM learns a hyperplane which best separates the two classes.

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.

4. Empirical experiment

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 217 observations of Iranian companies. All of them were or still are listed on the Tehran Stock Exchange (TSE). Of the 217cases for training 100 belong to the default case under paragraph 141 of Iran Trade Law and the other 117to non-default case.

The 21 significant variables in this study were selected by using a two stages predictive variable selection process. At the first stage, default prediction literature was reviewed and 65 variables from more than 230 financial ratios were selected as predictive variables. These financial ratios were chosen based on their popularity in the literature. In the second stage, 21 variables were selected based on the availability of the necessary data. 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 21 indicators for the model which could best discriminate the default firms from the non-default firms. These selected financial ratios include: EBIT to total assets (X1), current assets to total assets (X5), net profit to liability (X11), working capital to total assets (X6) and net profit to sale (X16).

4.2. Experimental Results

The results are presented in two parts. First part of this section displays the percent of misclassified cases for each classifier system. Then, the enhancement over the baselines has been shown for ensemble classifiers. Also, using ROC curve the accuracy of each classifier is assessed. Table 2 shows the percent of model accuracy and misclassified cases 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 difference between SVM and the next best model is small but statistically significant. 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. SVM with a high generalization capacity seems to be a capable technique for default prediction in Iran as a developing economy. Also, table 2 shows the performance accuracy of multi-stage classifiers in compare with baselines.

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		Definition of variable	Means of non-default companies	Means of default companies	Test of equality of group means
1	EBIT/TA	0.155647	-0.02608	0	12	NP/E	-0.08432	-1.0931	0.079
2	Ca/TA	0.086677	0.031281	0	13	S/TA	2.611479	2.424024	0.499
3	Ca/CL	1.854502	1.178482	0	14	S/CA	4.410169	4.135828	0.511

	Definition of variable	Means of non-default companies	Means of default companies	Test of equality of group means		Definition of variable	Means of non-default companies	Means of default companies	Test of equality of group means
4	CA/CL	0.618186	0.584403	0.271	15	R/L	0.807465	0.561762	0
5	CA/TA	0.151896	-0.05514	0	16	NP/S	0.039157	-0.04186	0
6	WC/TA	0.098689	-0.87107	0.174	17	L/TA	0.522716	0.704322	0
7	WC/S	9.504097	16.74657	0.147	18	L/E	16.12618	7.795392	0.483
8	S/R	0.095607	-0.06198	0	19	LL/E	6.580743	0.176757	0.315
9	NP/TA	0.095607	-0.06197	0	20	CL/E	9.545434	7.618636	0.738
10	GP/S	0.602299	-0.616	0	21	CA/S	0.933685	0.426074	0.441
11	NP/L	0.347821	0.080305	0					

EBIT: Earning Before Interests and Taxes **GP: Gross Profit**
TA: Total assets **L: Liabilities**
Ca: Cash **TI: Total Income**
CL: Current Liabilities **LL: Long Term Liabilities**
CA: Current Assets **E: Equity**
WC: Working Capital **R: Receivables**
S: Sale **NP: Net Profit**

The multi-stage classifiers considerably outperform the baseline. However, the improvement by the bagging is not significant, with only ensembles using logistic regression showing the major improvement. Also, simple majority vote of the baseline classifiers is relatively poor. By the results, all multi-stage systems outperform the baseline including Adaboost with Logistic regression, and neural network, decision tree and support vector machines. 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 1, shows the ROC curve for baseline and multi-stage classifiers.

Table 2
Performance of classifier systems

	Classifier system	% Accuracy	% Misclassified	% ROC Area
Baseline models	LR	84.79	15.21	84.9
	NN	83.41	17.38	91.3
	DT	84.79	17.12	83.8
	SVM	84.79	14.42	92.5
Multi-stage	LR (adaboost)	86.17	17.57	90.29
	NN (adabboost)	86.17	13.42	91.45
	DT (adaboost)	86.17	14.03	91.5
	SVM (adaboost)	89.07	14.38	91.9
	LR (bagging)	86.17	15.29	91.4
	NN (bagging)	82.94	18.11	92.8
	DT (bagging)	83.87	19.44	91.6
	SVM (bagging)	86.83	14.33	94.3

5. 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 from researchers. Appropriate identification of firms 'approaching default is undeniably required. Different 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 have made 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 multi-stage classifiers for Iranian firms as a developing economy. The accuracy of five classifiers was assessed to find out whether it is conceivable to forecast default of Iranian firms based on financial ratios. Empirical results highlighted that the financial ratios of EBIT/TA, CA/TA, NP/L, WC/TA and NP/S is recognized as highly predictive indicators. Most multi-stage combination classifiers provided significant improvements over the single classifiers. In addition, Adaboost afford enhancement in performance over the single classifiers. Bagging with support vector machines and logistic regression performed well.

According to the literature in various scientific and engineering fields, it has been found that ensemble of classifiers improves the prediction accuracy of single classifiers. The results of this study showed the improvement in prediction accuracy of classifiers. In addition, this paper employed ensemble classifiers for default risk prediction to afford a more accurate model.

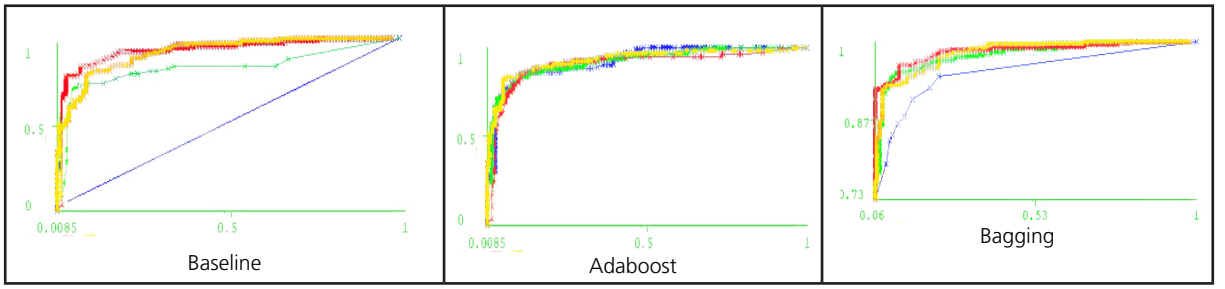

Fig. 2. Performance of Adaboost and Bagging

REFERENCES

Abellán, J., & Masegosa, A. (2012). Bagging schemes on the presence of noise in classification. Expert Systems with Applications, 39(8), 6827–6837. | Allison, P.D. (2001). Logistic Regression Using the SAS System: Theory and Application. Cary, NC: SAS Publishing, BBU Press. | Altman, Edward I. (1968). Financial Ratios, Discriminant Analaysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), 589-609. | Altman, Edward I. (1973). Predicting Railroad Bankruptcies in America. Bell Journal of Economics & Management Science, 4(1), 184. | Atiya, A. (2001). Bankruptcy Prediction for Credit Risk using Neural Networks: A Survey and New Results. IEEE Transactions on Neural Networks, 12, 929-935. | Beaver, William H. (1966). Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4(3), 71-111. | Brown, G., Wyatt, J.L. Harris, R. Yao, X. (2005). Diversity creation methods: a survey and categorization, Information Fusion, 6 (1), 5–20. | Carter, C., & Catlett, J. (1987). Assessing credit card applications using machine learning. IEEE Expert, 2, 71-79. | Dimitras, A. I., Zanakis, S. H., & Zopounidis, C.(1996). A survey ofbusiness failure with anemphasis on prediction methods andindustrialapplication. European Journal of Operational Research, 90, 487–513. | Freund Y. and Schapire R.E.A decision-theoreticgeneralisation of on-line learning and anapplicationto boosting.J. of Computer andSystemScience, 55(1):119–139, 1997. | Frydman, H., Altman, E., & Kao, D. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. Journal of Finance, 269-291. | 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 | 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. | Härdle, Wolfgang, Moro, Rouslan, & Schäfer, Dorothea. (2005). Predicting Bankruptcy with Support Vector Machines. Statistical Tools in Finance & Insurance, 225-248. | Hertz, J., Krogh, A., & Palmer, R.G. (1991). The Theory of Neural Network Computation. Addison Welsey: Redwood, CA. | Hosmer, D.W., & Lemeshow, S. (2000). Applied Logistic Regression. New York: Wiley. | Jackendoff, N. (1962). A Study of Published Industry Financial and Operating Ratios. Philadelphia: Temple University, Bureau of Economic and Business Research. | Libby, Robert. (1975). Accounting Ratios and the Prediction of Failure: Some Behavioral Evidence. Journal of Accounting Research, 13(1), 150-161. | Lin, F., and McClean, S. (2001). A Data Mining Approach to the Prediction of Corporate Failure- Knowledge-Based Systems, 14(3-4), 189-195. | Lin, Y., 2002. Improvement on behavioural scores by dual-model scoring system.International Journal of Information Technology andDecision Making 1, 153–165. | Martin, D. (1977). Early warnings of bank failure: A logit regression approach. Journal of Banking and Finance, 1, 249-276. | Messier, JR. W., & Hansen, J. (1988). Inducing rules for expert system development: an example using default and bankruptcy data. Management Science, 34, 1403-1415. | Morris, Richard. (1997). Early warning indicators of corporate failure : a critical review of previous research and further mpirical evidence. Ashgate. | Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. Journal of the American Statistical Association, 58(303), 799-806. | Ohlson, James A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1). | Pompe, P., & Feelders, A. (1997). Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy. Microcomputers in Civil Engineering 12, 267-276. | Polikar, R. (2006). Ensemble based systems in decision making, IEEE Circuits and Systems Magazine, 6 (3), 21–45. | Quinlan, J.R. (1986). Induction of Decision Trees, Machine Learning, 1, 81-106. | 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. | Rokach, L. (2010). Ensemble Methods in Supervised Learning. Data Mining and Knowledge Discovery Handbook, 959-979. | Shin, K.S., and Lee, Y.J. (2002). A Genetic Algorithm Application in Bankruptcy Prediction Modeling. Expert Systems with Applications, 9, 503-512. | 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. | Shumway, Tyler. (2001). Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business, 74(1), 101-124. | Smith, C.W., & Stulz, R.M. (1985). The Determinants of Firms' Hedging Policies. Journal of Financial and Quantitative Analysis, 20(4), 391-405. | Smith, R., & Winakor, A. (1935). Changes in financial structure of unsuccessful industrial corporations", Bureau of Business Research. Bulletin Urbna University of Illinois Press, 51. | West, D., Dellana, S., Qian, J., 2005. Neural network ensemble strategies for financial decision applications. Computers andOperations Research 32, 2543–2559. | Zhu, H., Beling, P.A., Overstreet, G., 2001. A study in the combination of twoconsumer credit scores. Journal of the Operational ResearchSociety 52, 974–980. | Zmijewski, M.E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research 22, 59–86. |