PREDICTING BANKRUPTCY USING MACHINE LEARNING ALGORITHMS

Thi Kha Nguyen[1], Thi Phuong Trang Pham[2]

[1]The University of Danang - Campus in Kontum; nguyenkha130490@gmail.com
[2]The University of Danang - University of Technology and Education; ptptrang@ute.udn.vn

Abstract - Bankruptcy prediction is of great utility for all economic stakeholders. Therefore, diverse methods have been applied for the early detection of financial risks in recent years. The objective of this paper is to propose an ensemble artificial intelligence (AI) model for effectively predicting the bankruptcy of a company. This study is designed to assess various classification algorithms over two bankruptcy datasets - Polish companies bankruptcy and Qualitative bankruptcy. The comparison results show that the bagging-ensemble model outperforms the others in predicting bankruptcy datasets. In particular, with the test data of Polish companies bankruptcy, the regression tree learner bagging (REPTree-bagging) ensemble model yields an accuracy of 100%. In predicting Qualitative bankruptcy dataset, the Random tree bagging (RTree-bagging) ensemble model has the highest accuracy with 96.2% compared to other models.

Key words - Bankruptcy prediction; single-methods; ensemble-models; artificial intelligence methods; bagging.

1. Introduction

Financial risk prediction is one of a critical topic in the domain of financial analysis because it can help companies to reduce financial distress and take appropriate actions in the future. Many financial risk prediction tasks are basically binary classification problems, which means observations are assigned to one of the two groups after data analysis [1]. This paper focuses on classifying bankruptcy problems.

Thanks to the development of computer power and data storage technologies, classification algorithms can be used to quickly and effectively predict financial data. However, the algorithm evaluation or algorithm selection play an important role in the result performance. Several classification models have been proposed for predicting financial problems in the past few decades. For example, credit risk and fraud risk prediction are given in Thomas (2000) [2] and Phua et al. (2010) [3]. Many authors have also contributed to the early warning models for classifying banks into two groups using semi parametric or nonparametric models [4]. In the study, the authors use computer-based early warning systems (EWSs) to make predictions and they concluded that nonparametric EWSs provided valuable information about the future viability of large banks. Besides, Godlewski (2006) applied a two step logit model to estimate excess credit risk and bank's default probability and they confirmed that the role of the institutional and regulatory environment as a source of excess credit risk, which increases a bank's default risk [5]. However, these approaches have been criticized a lot because of their restrictive assumptions that are not verified in reality [6] and were neglected with the emergence of the artificial intelligence (AI) techniques. AI models have greater predictive capability than conventional methods [7, 8]. Although AI-based models are convenient and effective for solving prediction problems, their accuracy is questionable. Therefore, this study uses the applicability of four single models, which are Dum stump (DStump), Random tree (RTree), a fast decision/regression tree learner (REPTree) and support vector machine (SVM) and ensembles model (bagging) to determine the situation of bankruptcy. These single AI models are the most commonly used in relevant works and some are recognized as the most effective ML models [9]. Therefore, these four models are adopted in this study to develop single AI models as well as ensembles.

Ensemble AI models were formed from the above single models, and these are ensemble bagging models. Then, we can choose the best model for forecasting the bankruptcy of a company, crucial for prediction tasks under extremely competitive and volatile business environments.

The remainder of this paper is organized as follows. Section 2 elucidates the single-AI models, ensemble-AI models, and the predictive evaluation methods. The collection and preprocess of bankruptcy datasets, and analytical results are mentioned in Section 3. Finally, conclusions are given in Section 4.

2. Methodology

2.1. Single AI Models

2.1.1. Dum stump

A DStump is one of the classification model with the simple tree structure consisting of one split, which can also be considered a one-level decision tree. The DStump [10] are often used as component base learners in machine learning ensemble techniques such as bagging and boosting.

2.1.2. Random tree

A RTree is a tree or arborescence that is formed by a stochastic process. In this study, the RTree is used as binary classifier for classification problems. Random binary tree, binary trees with a given number of nodes, formed by inserting the nodes in a random order or by selecting all possible trees uniformly at random [11].

2.1.3. Regression tree learner

The REPTree analysis is applied in WEKA. A REPTree is a classifier expressed as a recursive partition of the instance space. The REPTree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal node. All other nodes are called leaves (also known as terminal or decision nodes).

In a REPTree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values [12]. Depending on the target field, several impurity measures

can be used to locate splits for REPTree models.

2.1.4. Support vector machine

Support vector machines (SVMs) were developed by Vapnik *et al.* in 1995 [13], and these algorithms have been widely used for classification. The so-called "support vector" refers to training sample points at the edge of segment, while the "machine" refers to some concerned algorithms in the field of machine learning [14]. The SVM classifies by using an ε-insensitive loss function to map nonlinearly the input space into a high-dimensional feature space, and then constructs a linear model that implements nonlinear class boundaries in the original space.

2.2. Ensemble AI Models

The bagging method is a bootstrap method that is used to train several classifiers independently and with different training sets [15]. This is the reason why this study only uses bagging ensemble method for predicting bankruptcy problem. Bootstrapping builds *k* replicate training datasets that are used to construct k independent classifiers by random re-sampling of the original training dataset with replacement. The *k* classifiers are then aggregated through an appropriate combination method, such as a method based on the average of probabilities [9].

In this study, four individual learning techniques are combined into four homogeneous ensembles, which are an DStump-bagging ensemble, an RTree-bagging ensemble, a REPTree-bagging ensemble, and an SVM-bagging ensemble.

2.3. Evaluation methods

2.3.1. Accuracy

Accuracy can be defined as the degree of uncertainty in a measurement with respect to an absolute standard. The predictive accuracy of a classification algorithm is calculated as follows,

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (1)$$

Where true positive (*tp*) values is number of correctly recognized class examples, true negative (*tn*) values is number of correctly recognized examples that do not belong to the class that represents accurate classifications. The false positive (*fp*) value (number of examples that are either incorrectly) assigned to a class or false negative (*fn*) value (number of examples that are not assigned to a class) refers to erroneous classifications.

2.3.2. Precision

Precision is one of the extended versions of accuracy. and precision measures the reproducibility of a measurement. Precision in Eq. (2) is defined as the number of true positives as a proportion of the total number of true positives and false positives that are provided by the classifier.

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

2.3.3. Sensitivity

Sensitivity is another extended type of accuracy. It is also called recall. It measures the completeness. Sensitivity in Eq. (3) is the number of correctly classified positive examples divided by the number of positive examples in the data. In identifying positive labels, sensitivity is useful for estimating the effectiveness of a classifier.

$$Sensitivity = \frac{tp}{tp + fn} \quad (3)$$

3. Data preparation and analytical results

3.1. Data preparation

To assess the quality of the proposed methods two datasets are used, publicity available from UC Irvine Machine Learning Repository (UCI). Polish companies bankruptcy dataset contains 7027 instances with 64 predictor variables and 1 class variable. Qualitative bankruptcy dataset has 250 instances with 6 predictor variables and 1 class variable (Table 1). The model training process is conducted in a stratified 10-fold cross-validation scheme, where each model is trained/tested in parallel on the same training/testing blocks, so that the performance results are directly comparable.

3.2. Analytical results

The results of base and ensemble model using two proposed datasets are given in table 2. For each dataset, the best result of a specific performance measure is highlighted in boldface. The RTree-Bagging ensemble model achieves the best results across all measures on small size qualitative bankruptcy dataset (Accuracy=100%, Prediction=100%, Sensitivity=100%). For large dataset, such as the Polish companies bankruptcy dataset, REPTree-Bagging ensemble model produces satisfactory results on accuracy and prediction (Accuracy=96.2%, Prediction=94.5%).

Table 1. *The attributes in the datasets.*

Attribute	Polish companies bankruptcy dataset	Qualitative bankruptcy dataset
X1	net profit / total assets	industrial risk
X2	total liabilities / total assets	management risk
X3	working capital / total assets	financial flexibility
X4	current assets / short-term liabilities	credibility
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	competitiveness
X6	retained earnings / total assets	operating risk
X7	EBIT / total assets	
X8	book value of equity / total liabilities	
X9	sales / total assets	
X10	equity / total assets	
X11	(gross profit + extraordinary items + financial expenses) / total assets	
X12	gross profit / short-term liabilities	
X13	(gross profit + depreciation) / sales	
X14	(gross profit + interest) / total assets	

X15	(total liabilities * 365) / (gross profit + depreciation)	
X16	(gross profit + depreciation) / total liabilities	
X17	total assets / total liabilities	
X18	gross profit / total assets	
X19	gross profit / sales	
X20	(inventory * 365) / sales	
X21	sales (n) / sales (n-1)	
X22	profit on operating activities / total assets	
X23	net profit / sales	
X24	gross profit (in 3 years) / total assets	
X25	(equity - share capital) / total assets	
X26	(net profit + depreciation) / total liabilities	
X27	profit on operating activities / financial expenses	
X28	working capital / fixed assets	
X29	logarithm of total assets	
X30	(total liabilities - cash) / sales	
X31	(gross profit + interest) / sales	
X32	(current liabilities * 365) / cost of products sold	
X33	operating expenses / short-term liabilities	
X34	operating expenses / total liabilities	
X35	profit on sales / total assets	
X36	total sales / total assets	
X37	(current assets - inventories) / long-term liabilities	
X38	constant capital / total assets	
X39	profit on sales / sales	
X40	(current assets - inventory - receivables) / short-term liabilities	
X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))	
X42	profit on operating activities / sales	
X43	rotation receivables + inventory turnover in days	
X44	(receivables * 365) / sales	
X45	net profit / inventory	
X46	(current assets - inventory) / short-term liabilities	
X47	(inventory * 365) / cost of products sold	
X48	EBITDA (profit on operating activities - depreciation) / total assets	
X49	EBITDA (profit on operating activities - depreciation) / sales	
X50	current assets / total liabilities	
X51	short-term liabilities / total assets	
X52	(short-term liabilities * 365) / cost of products sold)	
X53	equity / fixed assets	
X54	constant capital / fixed assets	
X55	working capital	
X56	(sales - cost of products sold) / sales	
X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)	
X58	total costs /total sales	
X59	long-term liabilities / equity	
X60	sales / inventory	
X61	sales / receivables	
X62	(short-term liabilities *365) / sales	
X63	sales / short-term liabilities	
X64	sales / fixed assets	
Y	Class: { Bankruptcy, Non-Bankruptcy}	Class:{Bankruptcy, Non-Bankruptcy}

Table 2. *Classification results*

Dataset	Model	Accuracy (%)	Prediction (%)	Sensitivity (%)
Polish	DStump	96.1	94.3	0
Polish	RTree	93.0	93.1	8.8
Polish	REPTree	95.9	94.5	12.5
Polish	SVM	96.1	94.3	0
Polish	DStump-bagging ensemble	96.1	94.3	0
Polish	RTree-bagging ensemble	95.9	94.3	5.8
Polish	**REPTree-bagging ensemble**	**96.2**	**94.5**	**16.0**
Polish	SVM-bagging ensemble	96.1	94.3	0
Qualitative	DStump	98.4	98.4	96.8
Qualitative	RTree	98.8	98.8	97.6
Qualitative	REPTree	98.8	98.8	97.5
Qualitative	SVM	98.8	98.8	97.6
Qualitative	DStump-bagging ensemble	98.4	98.4	96.8
Qualitative	**RTree-bagging ensemble**	**100.0**	**100.0**	**100.0**
Qualitative	REPTree-bagging ensemble	98.4	98.4	96.7
Qualitative	SVM-bagging ensemble	99.6	99.6	99.2

4. Conclusions

As a result of the recent world-wide financial crisis and economic recession, the demand for bankruptcy prediction models have gained strong attention. Therefore, it is important to provide financial decision makers with effective predictive power to anticipate these loss scenarios. Machine learning models have been very successful in finance applications, and many studies examine their use in bankruptcy prediction.

In this work we empirically compare different base and ensemble classification models, namely, DStump, RTree, REPTree, SVM, DStump-bagging ensemble, RTree-bagging ensemble, REPTree-bagging ensemble, SVM-bagging ensemble, in a setting of real-world bankruptcy

data from the UCI.

Regarding the qualitative bankruptcy dataset, RTree-Bagging ensemble model shows to be superior in comparison with the others proposed in this study. For Polish companies bankruptcy dataset, REPTree-Bagging ensemble model achieves the best performance among the others.

Our study does not focus on feature selection. Therefore, the impact of feature selection would not be prominent in our study. Another limitation of the study is that it does not consider different classification costs. We find that, especially for prediction of bankruptcy, accuracy should not be the only performance metric, and future research should focus on adjusting classification models by considering different impacts. Future studies should also extend the analysis to bankruptcy prediction of construction companies. The methodology can be applied to banking, such as loan default prediction, fraud detection and marketing.

REFERENCES

[1] H. Frydman, E.I.A., D. Kao, Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance 40*, 1985: p. 269–291.

[2] Thomas, L.C., A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 2000. **16**(2): p. 149-172.

[3] Wang, S., A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research, in Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation - Volume 01. 2010, *IEEE Computer Society*. p. 50-53.

[4] Kolari, J., et al., Predicting large US commercial bank failures. *Journal of Economics and Business*, 2002. **54**(4): p. 361-387.

[5] Godlewski, C.J., Regulatory and Institutional Determinants of Credit Risk Taking and a Bank's Default in Emerging Market Economies. *Journal of Emerging Market Finance*, 2006. **5**(2): p. 183-206.

[6] Feki, A., A.B. Ishak, and S. Feki, Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction. *Expert Systems with Applications*, 2012. **39**(3): p. 3087-3099.

[7] Chou, J.-S. and A.-D. Pham, Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Construction and Building Materials*, 2013. **49**: p. 554-563.

[8] Chou, J.-S., N.-T. Ngo, and A.-D. Pham, Shear Strength Prediction in Reinforced Concrete Deep Beams Using Nature-Inspired Metaheuristic Support Vector Regression. *Journal of Computing in Civil Engineering*, 2016. **30**(1): p. 04015002.

[9] Chou, J.-S., N.-T. Ngo, and W.K. Chong, The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. *Engineering Applications of Artificial Intelligence*, 2016.

[10] Reyzin, L. and R.E. Schapire, How boosting the margin can also boost classifier complexity, in Proceedings of the 23rd international conference on Machine learning. 2006, ACM: Pittsburgh, Pennsylvania, USA. p. 753-760.

[11] Reed, B., The height of a random binary search tree. *J. ACM*, 2003. **50**(3): p. 306-332.

[12] Jamil, L.S., Data analysis based on data mining algorithms using Weka workbench. *International Journal of Engineering Sciences & Research Technology*, 2016. **5**(8): p. 262-267.

[13] Cortes, C. and V. Vapnik, *Support-Vector Networks. Machine Learning*, 1995. **20**(3): p. 273-297.

[14] Zhang, H., et al., *Predicting profitability of listed construction companies based on principal component analysis and support vector machine - Evidence from China*. Automation in Construction, 2015. **53**: p. 22-28.

[15] Breiman, L., *Bagging Predictors. Machine Learning*, 1996. **24**(2): p. 123-140.

(The Board of Editors received the paper on 16/5/2018, its review was completed on 17/9/2018)