PERFORMANCE EVALUATION OF CLASSIFICATION ALGORITHMS IN FINANCIAL RISKS PREDICTION

ĐÁNH GIÁ CÁC THUẬT TOÁN PHÂN LOẠI TRONG VIỆC DỰ ĐOÁN NHỮNG RỦI RO VỀ TÀI CHÍNH

Thi Phuong Trang Pham

The University of Danang, University of Technology and Education; ptptrang@ute.udn.vn

Abstract - Financial risks have always been the topic of interest of researchers as well as investors. Therefore, predicting financial risks in current economy is necessary. For a given dataset, selecting a suitable classifier or set of classifiers is an important task in financial risk forecast. The goal of this paper is to apply three popular machine-learning techniques; Support vector machine (SVM), Decision tree (DT) and Naïve Bayes (NB) to predicting financial risks based on real-life data - Qualitative Bankruptcy, Japanese bankruptcy and Australian credit card application. The results demonstrate that the SVM algorithm has the best and most reliable classification accuracy at 99.600%, 87.652% and 86.783% for Qualitative Bankruptcy, Japanese bankruptcy and Australian credit card application, respectively. However, the results of two algorithms (DT and NB) also yield good accuracy for three real datasets. This work also demonstrates the effectiveness of machine learning technique in classifying financial risks.

Key words - Financial risks; machine-learning techniques; Support vector machine; Decision tree; Naïve Bayes.

Tóm tắt - Rủi ro tài chính luôn là đề tài gây hứng thú cho các nhà nghiên cứu và những nhà đầu tư. Vì vậy, việc dự đoán những rủi ro tài chính trong nền kinh tế hiện nay là cần thiết. Và cách lựa chọn được một hay nhiều lớp phân loại là nhiệm vụ quan trọng. Mục đích bài báo này là sử dụng ba thuật toán phổ biến của phương pháp máy học; máy học vecto hỗ trợ, cây quyết định và thuật toán Naïve Bayes; để dự đoán khả năng rủi ro của ba bộ dữ liệu tài chính – sự phá sản định tính, sự phá sản tại Nhật Bản và ứng dụng thẻ tín dụng tại Úc. Kết quả cho thấy, thuật toán SVM cho kết quả phân loại tốt nhất và đáng tin cậy với độ chính xác lần lượt cho ba bộ dữ liệu sự phá sản định tính, sự phá sản tại Nhật Bản và ứng dụng thẻ tín dụng tại Úc là 99,6000%, 87,652% và 86,783%. Tuy nhiên, kết quả của hai thuật toán còn lại cho ba bộ dữ liệu trên cũng đạt kết quả tốt. Nghiên cứu này còn muốn chứng minh tính hiệu quả của phương pháp máy học trong việc phân loại rủi ro tài chính.

Từ khóa - Rủi ro tài chính; kỹ thuật học máy; máy học vecto hỗ trợ; cây quyết định; Naïve Bayes.

1. Introduction

Risk can be considered to be unfortunate, loss and danger. Risk is also a loss of property or a decrease in real profit compared to the expected profit. Financial risk is the possibility that shareholders will lose money when investing in a company that has debt if the company's cash flow cannot meet its financial tasks. Financial risks are associated with form of financing, including credit risk, business risk, investment risk, and operational risk.

Financial risk prediction plays an important role in the financial analysis. Investors can use an amount of financial risk information to assess an investment's prospects. Besides, predicting financial risks helps portfolio managers assess the amount of capital reserves to maintain, and to help guide their purchases and sales of various classes of financial assets. As a result, it is important to ensure that financial risks are identified and managed appropriately.

Many years ago, several researchers also applied traditional classification methods based on previous experience for forecasting credit and risk assessment [1]. To be honest, traditional method cannot value financial risks efficiently and effectively because of the development of business, the social needs and the increase of the size of databases. Therefore, thanks to the expansible computer power and data storage technologies, classification methods can be used to quickly forecast credit and fraud risk [1]. It is clear that the classification models provide higher prediction accuracies than traditional approach. Moreover, unfair humans cannot identify or decide investment; it based on the results of classification algorithms [1]. Additionally, classification algorithms for predicting financial risk helps to process credit applications fast, manage credit risk flexibly and require fewer humans [1].

Many classification models have been constructed for credit and fraud risk forecast in the past few decades, including statistical models, nonparametric statistical models, artificial intelligence methods, and mathematical programming methods [1]. Rosenberg and Gleit [2] surveyed the use of discriminating analysis, decision trees, and expert systems for static decisions, and dynamic programming, linear programming, and Markov chains for decision models in financial risk. Moreover, Hand and Henley [3] and Phua et al.[4] proposed some classification models in predicting credit risk and fraud risk. In the application area of building credit scoring models, Desai et al. [5] and West [6] concluded that customized neural network methods outperformed linear discriminating analysis, whereas Yobas et al. [7] reported that the predictive performance of linear discriminating analysis was superior to neural networks [1]. Hence, no one can say that any single classification algorithm could exactly achieve the best performance for all measures [1].

Machine learning is the study of algorithms and statistical models that computer systems use to perform a specific task effectively. This study applies three popular classification algorithms to classify three real datasets; Support vector machine, Decision tree and Naïve Bayes. The results show that support vector machine yields the highest accuracy for all three datasets. The SVM had 99.600% for Qualitative bank dataset, 87.652% for

Japanese bankruptcy dataset and 86.783% for Australian credit card application. However, the results between three models for three dataset are not quite different levels. Besides, the result of models has also based on the quality of dataset, if the dataset has balance between classifier in output, the results maybe better and vice versa.

The remainder of this paper is organized as follows. Section 2 elucidates the Support vector machine, Decision tree and Naïve Bayes and the predictive evaluation methods. The collection and detail of financial risk datasets, and analytical results are mentioned in Section 3. Finally, conclusions are given in Section 4.

2. Methodology

2.1. Support vector machine

Support vector machines (SVMs) were developed by Vapnik et al. in 1995 [8], and these algorithms have been widely used for classification. The SVM classifies use 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.

The formulation of an SVMs classifier can be initiated using two following assumptions.

$$\vec{w} \bullet \vec{x_+} + b \geq 1 \text{ if x = +1} \quad (1)$$

$$\vec{w} \bullet \vec{x_-} + b \leq 1 \text{ if x = -1} \quad (2)$$

where $\vec{w}$ denotes an SVMs margin vector; $\vec{x_+}$ and $\vec{x_-}$ denotes an SVMs positive class vector and an SVMs negative class vector, respectively; b denotes an SVMs bias term; y_i indicates the class to which the sample $\vec{x}$ belongs; and $\bullet$ denotes dot products. The assumptions (1) and (2) are the constraints for minimizing Eq. (3) to maximize the margins between various categories.

$$\min_{\vec{w}} = \frac{1}{2} \| \vec{w} \|^2 \quad (3)$$

The results of the Lagrange multiplier equation are used to optimize Eq. (3) as follows.

$$L(\alpha_i) = \frac{1}{2} \| \vec{w} \|^2 - \sum_{i=1}^{N} \alpha_i (y_i (\vec{w_i} \bullet \vec{x_i} + b) - 1) \quad (4)$$

where α_i denotes a Lagrange slack variable.

2.2. Decision tree

Decision tree is one of the most widely used and practical methods for inductive inference over supervised data [9]. It bases on various attributes a decision tree represents a procedure that classifies the categorical data and created a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Given training vector s$x_i \in R^n$, i=1,…,l and a label vector$y \in R^l$, a decision tree groups the sample according to the same labels.

2.3. Naïve Bayes

The NB classifier is a simple linear classifier and based on applying Bayes' theorem with strong independent assumptions between the features. A given example will be given the most likely class by the NB classifier as described by its feature vector. It assumes that the decision problem is posed in probabilistic terms and that all of the relevant probability values are known [10]. Moreover, NB is also efficient to train and use and is easy to update with new data [10].

This study applies the Weka software to run three data. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

2.4. Evaluation

There are various approaches suggested for evaluating the performance of classifiers. However, accuracy is the most popular factor to use. And in this study, the accuracy is chosen to evaluate and compare three models when applying three real datasets concerning financial risks. Moreover, all three datasets are binary-class problem, therefore accuracy is the best item to evaluate the proposed models. It can be calculated by computing four quantities: true positives (*tp*) is an outcome where the model correctly predicts the positive class, true negatives (*tn)* is an outcome where the model correctly predicts the negative class, false positive (*fp*) is an outcome where the model incorrectly predicts the positive class and false negatives (*fn)* is an outcome where the model incorrectly predicts the negative class. The predictive accuracy of a classification algorithm is calculated in Equation 5.

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (5)$$

3. Real financial risk data

3.1. Data preparation

Qualitative bankruptcy dataset obtained from UCI Machine Learning Repository (archive.ics.uci.edu/ml/datasets/ qualitative bankruptcy). This dataset includes 250 data points with 6 attributes, each corresponding to Qualitative Parameters in Bankruptcy; they are industrial risk, management risk, flexibility, credibility, competitiveness and operating risk. The class distribution is143 instances for non-bankruptcy and 107 instances for bankruptcy.

Japanese bankruptcy dataset collects bankrupt Japanese firms and non-bankrupt Japanese firms from various sources during the post-deregulation period of 1989–1999. The dataset has 14 input variables and 1 output variable (bankrupt or non-bankrupt).This study has collected the data from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening).

Australian credit card application dataset was provided by a large bank and concerns consumer credit card applications. It includes 690 instances with 15 predicator variables plus 1 class variable that an application is accepted or declined[1].Table 1 presents details about the datasets.

Table 1. *Data Description*

Dataset	Instances	Attributes
Qualitative bankruptcy	250	7
Japanese bankruptcy	656	15
Australian credit card application	690	15

3.2. Analytical results

Table 2 compares the performances of the SVM, DT and NB models using three real financial risk datasets. For all three datasets, the SVM model had the highest accuracy and outperformed other models with 99.600% for qualitative bank dataset, 87.652% for Japanese bankruptcy and 86.783% for Australian credit card application. However, from the results of table 2, other models also yielded quite high accuracy for three datasets. With Qualitative bank dataset, the accuracy of DT and NB were 98% and 99.200%, respectively. DT and NB had 86.433% and 85.518%, respectively for Japanese bankruptcy dataset. For Australian credit dataset, DT had the 85.217% and NB had 85.073% of accuracy. Therefore, three proposed machine learning techniques are quite suitable for predicting above three financial risk datasets.

Table 2. *Comparison results*

Dataset	Classification algorithms	Accuracy (%)
Qualitative bankruptcy	**SVM**	**99.600**
	Decision tree	98.000
	Naïve Bayes	99.200
Japanese bankruptcy	**SVM**	**87.652**
	Decision tree	86.433
	Naïve Bayes	85.518
Australian credit card application	**SVM**	**86.783**
	Decision tree	85.217
	Naïve Bayes	85.073

4. Conclusion

Support vector machine, Decision tree and Naïve Bayes are three relatively models applied in financial risk prediction problems. The accuracy of SVM outperforms those of the two older models (Decision tree and Naïve Bayes). In the future study, the author hopes to enhance the above models to increase the efficiency and effectiveness of models. For example, the author will optimize SVM, DT and Naïve Bayes models. For example, the author can combine SVM with Naïve Bayes, SVM with DT or integrate SVM with other optimized model to create the new model with higher accuracy in solving financial risk problems. In addition to this, in the further research, the author hopes to apply more real datasets concerning financial risks to verify the effectiveness of machine learning techniques.

REFERENCES

[1] Y. Peng, G. Wang, G. Kou, and Y. Shi, "An empirical study of classification algorithm evaluation for financial risk prediction," *Applied Soft Computing*, vol. 11, pp. 2906-2915, 2011/03/01/ 2011.

[2] A. G. E. Rosenberg, "Quantitative methods in credit management: a survey," *Operations Research 42* pp. 589–613., 1994.

[3] D. J. Hand and W. E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: a Review", *Journal of the Royal Statistical Society Series A*, vol. 160, pp. 523-541, 1997.

[4] C. Phua, V. Lee, K. Smith-Miles, and R. Gayler, *A Comprehensive Survey of Data Mining-based Fraud Detection Research (Bibliography)*, 2013.

[5] V. S. Desai, J. N. Crook, and G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment", *European Journal of Operational Research*, vol. 95, pp. 24-37, 1996/11/22/ 1996.

[6] D. West, "Neural network credit scoring models", *Comput. Oper. Res.*, vol. 27, pp. 1131-1152, 2000.

[7] Y. M. B, C. J. N, and R. P, "Credit scoring using neural and evolutionary techniques", *IMA Journal of Management Mathematics*, vol. 11, pp. 111-125, 2000.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, pp. 273-297, September 01 1995.

[9] A. R. Bhumika Gupta, Akshay Jain, Arpit Arora, Naresh Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", *International Journal of Computer Applications*, vol. 163, 2017.

[10] H. Son, C. Kim, N. Hwang, C. Kim, and Y. Kang, "Classification of major construction materials in construction environments using ensemble classifiers", *Advanced Engineering Informatics*, vol. 28, pp. 1-10, 2014/01/01/ 2014.

(BBT nhận bài: 04/11/2018, hoàn tất thủ tục phản biện: 19/01/2019)