Knowledge-Based Systems 33 (2012) 166–172

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A hybrid KMV model, random forests and rough set theory approach for credit rating

Ching-Chiang Yeh [a,*], Fengyi Lin [b], Chih-Yu Hsu [c]

[a] Department of Business Administration, National Taipei College of Business, No. 321, Sec. 1, Ji-Nan Rd., Zhongzheng District, Taipei 10051, Taiwan
[b] Department of Business Management, National Taipei University of Technology, No. 1, Sec. 3, Chung-hsiao E. Rd., Taipei 10608, Taiwan
[c] Graduate Institute of Commerce Automation and Management, National Taipei University of Technology, No. 1, Sec. 3, Chung-hsiao E. Rd., Taipei 10608, Taiwan

ARTICLE INFO

Article history:
Received 22 October 2011
Received in revised form 13 February 2012
Accepted 2 April 2012
Available online 12 April 2012

Keywords:
Credit rating
KMV model
Rough set theory
Random forests
Distance to default

ABSTRACT

In current credit ratings models, various accounting-based information are usually selected as prediction variables, based on historical information rather than the market's assessment for future. In the study, we propose credit rating prediction model using market-based information as a predictive variable. In the proposed method, Moody's KMV (KMV) is employed as a tool to evaluate the market-based information of each corporation. To verify the proposed method, using the hybrid model, which combine random forests (RF) and rough set theory (RST) to extract useful information for credit rating. The results show that market-based information does provide valuable information in credit rating predictions. Moreover, the proposed approach provides better classification results and generates meaningful rules for credit ratings.

1. Introduction

Credit ratings are determined by rating agencies' assessments of the probability distribution of future cash flows to bondholders, which in turn, depends on the future cash flows to the firm. A firm's credit worthiness is determined by assessing the likelihood that its future cash flows will be sufficient to cover debt service costs and principal payments. As the mean of the firm's future cash flow distribution shifts downward or the variance of its future cash flows increases, the likelihood of default increases and the firm's credit rating will decline. Because credit ratings play a significant role in financing and investment decisions, there are many studies attempting to predict credit ratings by the academics and rating agencies.

Many researchers have attempted to construct automatic classification systems to solve credit rating problems using traditional statistical and artificial intelligence technique. The former include linear regression [26,47,54], linear multivariate discriminant analysis (MDA) [2,46], probit regression [28,32], logit analysis [31] and multidimensional scaling [40]. The latter consists of neural networks [15,22,35], case-base reasoning [34,50] and support vector machines (SVM) [3,8,27,33]. However, whilst these are well-established credit rating prediction techniques, some main problems arise.

Firstly, the above approaches are mainly based on accounting-based variables have some drawbacks. The most obvious drawbacks accounting information is inherently backward looking, based on historical information rather than the market's assessment for future [24]. In an attempt of more accurately credit rating analysis, much effort has been invested in the development of new models of default probability that uses financial theory and market information. Many of these models are so-called structural and reduced form models. The most well known application is the Moody's KMV (hereafter KMV) model. The models are proposed by Black and Scholes [3] and Merton [41]. It is claimed that market prices reflect future expected cash flows, and thus should be more useful in credit rating prediction. Market-based measures are further examined by a number of studies, including [6,25,48,52], in assessing default probability.

Moreover, the above-related works show that different researchers took different independent (predictive) variables as input for credit rating prediction. Based on these predictive variables, few of them took predictive variables as a module of their credit rating prediction. They also did not pay much attention to finding and selecting important predictive variables based on their importance. Moreover, fewer studies are able to identify interesting conceptual patterns or structural relationships among these variables and are difficult to use in making generalizations for credit rating.

In order to find the relative importance of the potential predictive variables, random forests (RF) is used in this study. RF is a new ensemble method that combines trees grown on bootstrap samples of data and random subset bagging of predictive variables [5]. During randomization of features, RF can provide an importance index of predictive variables by calculating accuracy. Furthermore, the importance index has captured the variable importance index

* Corresponding author. Tel.: +886 2 2322 6161; fax: +886 2 2322 6323.
E-mail addresses: ychinc@webmail.ntcb.edu.tw (C.-C. Yeh), fengyi@ntut.edu.tw (F. Lin), t7488049@ntut.edu.tw (C.-Y. Hsu).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.04.004

based on random forests [19]. In terms of robustness to outliers and noise, and calculation time, RF is superior to other machine learning methods such as bagging or boosting [37]. RF has been successfully applied to various problems in, e.g., genetic epidemiology and microbiology in general within the last 5 years. Within a very short period, RF has become a major data analysis tool that performs well in comparison with many standard methods [13]. Recently, many different applications for the RF have already been studied in Verikas et al. [53].

Moreover, rough set theory (RST), first proposed by Pawlak [45] in 1982, employed mathematical modeling to deal with classification problems, and then turned out to be a very useful tool for decision support systems, especially when hybrid data, vague concepts and uncertain data were involved in the decision process. Compared to other methods used in financial area, conventional statistical methods for example, the RST has several advantages. First, RST often implement classification by giving "if-then" like rules and are sometimes more welcomed by decision makers because they are easily understood. Second, it does not make any assumption about the distribution of data. In addition, it is a tool suitable for analyzing quantitative and qualitative attributes. RST has been successfully applied in many different fields, such as controlling industrial processes [9,29], diagnosis analysis [42,51], image processing [43], market decision-making [1,16,49], environmental problem detection [11,39], knowledge acquisition [17,56,57], web and text categorization [12,18,30] and early warning [7,55]. Unfortunately, it is rarely applied to the prediction problem of credit rating [10].

In this study, we use market-based information as the predictive variable. In the proposed method, KMV is employed as a tool to evaluate the market-based information of each corporation. To verify the proposed method, using a hybrid model, this combines RF and RST to improve accuracy rate for credit rating. Firstly, we incorporate financial variables and KMV as the potential predictive variables for credit rating prediction. Secondly, RF is used to perform variable selection because of its reliability in obtaining the relative importance of the predictive variables. Next, we use the obtained the important predictive variables from RF as inputs of RST models. The obtained results can then be compared to see whether the one including KMV will give better classification accuracy or not. Last, we generate results in the form of *if-then* rules that are transparent and easily understood for decision makers.

The remainder of this paper is organized as follows. Section 2 describes the methods used in the paper: the KMV model, RF and RST, respectively. Section 3 outlines the hybrid RF with RST strategies and experiment framework used in this study. Section 4 presents the experimental results of the proposed method. Finally, the conclusions are contained in Section 5.

2. Methodology

2.1. The KMV model

The KMV model was developed by the KMV Corporation in the late 1980s, being successfully marketed by KMV until KMV was acquired by Moody's in April 2002. The KMV model estimates the market value of debt by applying the Merton [41] bond-pricing model. In this model, a company's equity is viewed as an option related to the assets of the company. The total asset value of a firm is not only assumed to be tradable, but to also follow a geometric Brownian motion, such that

$$dV_A = \mu V_A\,dt + \sigma_A V_A\,dW \tag{1}$$

where V_A is the total value of the firm, μ is the expected continuously compounded return on V_A, σ_A is the volatility of firm returns and dW is a standard Weiner process. The firm is assumed to have one discount bond maturity in T periods. Under these assumptions, the equity of the firm E is a call option on the underlying value of the firm with an exercise price equal to the face value of debt F and a time to maturity of T. The value of equity is then modeled using the Black–Scholes–Merton formula,

$$E = V_A \cdot N(d_1) - F \cdot e^{-rT} \cdot N(d_2) \tag{2}$$

where E is the market value of the firm's equity, F is the face value of the firm's debt, r is the instantaneous risk-free rate, $N(\cdot)$ is the cumulative normal distribution function with d_1 and d_2 given by:

$$d_1 = \frac{\ln\frac{V_A}{F} + \left(r + \frac{\sigma_A^2}{2}\right)\cdot T}{\sigma_A \cdot \sqrt{T}}, \quad d_2 = d_1 - \sigma_A \cdot T. \tag{3}$$

In order to estimate the DD measure from equity values, we follow a similar procedure adopted by KMV and Vassalou and Xing [52] to calculate the volatility of a firm's assets (σ_A) and the market value of the firm's assets (V_A). At the end of each month, using the past 12 months of daily equity returns, we estimate volatility of equity returns (E_σ) and use it as an initial value of the volatility of returns on the firm's assets (σ_A). Therefore, for each day in the past 12-month period, the market value of the firm's assets (V_A) can be computed from Eq. (2). The standard deviation of returns on firm assets (σ_A) is then re-estimated and used for the new iteration. The procedure is repeated until the values of the volatility of returns on firm assets (σ_A) from two consecutive iterations converge. By keeping the estimation window equal to 12 months, the estimation of the volatility of returns on firm assets (σ_A) is repeated at the end of each month. The estimates of monthly volatility of the firm's assets and the market value of the firm's assets can be obtained.

We then estimate the drift term (μ) by calculating the mean of the changes of the natural logarithm of the firm's assets. Finally, the theoretical DD can be obtained as follows:

$$\mathrm{DD} = \frac{\ln\left(\frac{V_A}{F}\right) + \left(\mu - \frac{\sigma_A^2}{2}\right)T}{\sigma_A T} \tag{4}$$

DD, in short, is a volatility-adjusted measure of leverage. Hence, the larger the DD, the greater is the distance of a company from the default point, and the lower is the probability of default.

2.2. Random forests

Random forests (RF) are an algorithm for classification developed by Breiman [5] that uses an ensemble of classification trees [4,23]. Let us briefly recall the statistical framework by considering a learning set $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ made of n i.i.d. observations of a random vector (X, Y). Vector $X = (X^1, \ldots, X^p)$ contains predictors or explanatory variables, say $X \in R^P$, and $Y \in \overline{Y}$ where $\overline{Y}$ is either a class label or a numerical response. For classification problems, a classifier t is a mapping $t : R^p \to \overline{Y}$ while for regression problems, we suppose that $Y = s(X) + \varepsilon$ with $[\varepsilon|X] = 0$ and s the so-called regression function (for more background on statistical learning, see e.g. Hastie et al. [23]). RF is a model building strategy providing estimators of either the Bayes classifier, which is the mapping minimizing the classification error $P(Y \neq t(X))$, or the regression function.

The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the learning sample L and choosing randomly at each node a subset of explanatory variables X. More precisely, with respect to the well-known classification and regression trees (CART) model building strategy [4] performing a growing step followed by a pruning one, two differences can be noted. First, at each node, a given number (denoted by m_{try}) of input variables are randomly