Expert Systems With Applications 53 (2016) 87–105

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A novel corporate credit rating system based on Student's-*t* hidden Markov models

Anastasios Petropoulos[a], Sotirios P. Chatzis[b,*], Stylianos Xanthopoulos[a]

[a] Department of Mathematics, University of Aegean, Karlovasi 83200, Greece
[b] Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, Limassol 3036, Cyprus

ARTICLE INFO

Keywords:
Corporate credit rating
Hidden Markov model
Student's-*t* distribution
Expectation-maximization
Basel framework
Statistical machine learning

ABSTRACT

Corporate credit rating systems have been an integral part of expert decision making of financial institutions for the last four decades. They are embedded into the pricing function determining the interest rate of a loan contact, and play crucial role in the credit approval process. However, the currently employed intelligent systems are based on assumptions that completely ignore two key characteristics of financial data, namely their heavy-tailed actual distributions, and their time-series nature. These unrealistic assumptions definitely undermine the performance of the resulting corporate credit rating systems used to inform expert decisions. To address these shortcomings, in this work we propose a novel corporate credit rating system based on Student's-*t* hidden Markov models (SHMMs), which are a well-established method for modeling heavy-tailed time-series data: Under our approach, we use a properly selected set of financial ratios to perform credit scoring, which we model via SHMMs. We evaluate our method using a dataset pertaining to Greek corporations and SMEs; this dataset includes five-year financial data, and delinquency behavioral information. We perform extensive comparisons of the credit risk assessments obtained from our method with other models commonly used by financial institutions. As we show, our proposed system yields significantly more reliable predictions, offering a valuable new intelligent system to bank experts, to assist their decision making.

1. Introduction

In this work, we focus on the problem of*creating intelligent tools to assist credit scoring/rating of individual corporations*. In general, a credit scoring/rating system makes use of a statistical technique that combines and analyzes a series of account statement data to predict the future behavior of a company in terms of its ability to service its debt. The used account data are usually in the form of financial ratios, while the system-generated predictions are typically quantified as the likelihood of occurrence of a default event at some specific future time point. Corporate credit rating systems have been an integral part of financial institutions decision making for the last four decades. Nowadays their use is widespread, and gives the loans officers the ability to discriminate customers based on their risk profile and make informed decisions during all the stages of a loan lifecycle. Rating expert systems are embedded into the pricing function determining the interest rate of a loan contact, and play crucial role in the credit approval process.

The introduction of the Basel II framework (Basel Committee on Banking Supervision, 2005a), and its continuation in Basel III (Basel Committee on Banking Supervision, 2010), has triggered immense interest in intelligent credit rating systems research. Specifically, under these frameworks, banks have to implement appropriate rating systems for estimating the probability of a customer becoming delinquent; this is used in order to assess the borrower's creditworthiness, and estimate the capital requirements attached to the specific loan contact. Furthermore, in order to comply with the regulations set out in Basel II/Basel III, financial institutions have to satisfy the use test, i.e. use the corporate rating systems in all the phases of the customer relationship (credit granting, provisioning, pricing, collateral management, collections and write offs). Finally, under Basel II/Basel III, banks have to build their rating systems using state-of-the-art *statistical* methods, as this is expected to increase the accuracy of their risk discrimination procedures, thus guaranteeing their good financial performance in the years to come. As such, since the introduction of Basel II, banks have been consistently motivated to develop more accurate and robust intelligent systems to drive expert decisions, exploring new statistical techniques especially from the field of statistical machine learning. In the last decades, a plethora of alternative approaches have been

* Corresponding author. Tel.: +35725002041.
E-mail addresses: tasospetropoulos@gmail.com (A. Petropoulos), soteri0s@me.com (S.P. Chatzis), sxantho@aegean.gr (S. Xanthopoulos).

http://dx.doi.org/10.1016/j.eswa.2016.01.015
0957-4174/

developed to address the problem of modeling the credit quality of a company, using both quantitative information (e.g., account statements) and qualitative information (e.g., other underwriting criteria, such as obligors market and sector indicators).

A first category of approaches belongs to the family of classical regression techniques. Altman (1968) used multiple linear discriminant analysis (LDA) to build a rating system for predicting corporate bankruptcies. They estimated a linear discriminant function using liquidity, profitability, leverage, solvency, and turnover financial ratios to estimate credit quality; they dubbed their approach as the Z-score model. One of the main drawbacks of this approach is its assumption that the modeled variables are normally distributed, which is hardly ever the case in real-world scenarios. As such, this method cannot effectively capture nonlinear relationships among the modeled variables, which is crucial for the performance of the credit rating system. In a similar vein, several studies have explored the utility of probit models (e.g, (Mizen & Tsoukas, 2012)) and linear regression models (e.g., Avery, Calem, & Canner (2004)). However, these models continue to suffer from the same drawbacks that plague LDA, namely their clear inability to capture non-linear dynamics, which are prevalent in financial ratio data (Petr & Gurný, 2013).

Logistic regression is another approach broadly used for building corporate rating systems. It was first used by Ohlson (1980) to predict corporate bankruptcy based on publicly available financial data pertaining to several enterprises (e.g., financial ratios). Logistic regression models employed in this context are essentially used to classify corporations into two distinct classes characterizing their credit risk (i.e., good or bad). Typically, a sigmoid likelihood function is used for modeling purposes to allow for capturing non-linearities and relaxing the normality assumption during model estimation (Kamstra, Kennedy, & Suan, 2001).

Decision trees comprise a further category of non-parametric methods used for developing credit rating systems. Decision trees are models that consist of a set of nodes, corresponding to the modeled explanatory variables, and split conditions based on a hierarchical selection of the modeled explanatory variables. Two well-known algorithms in this field are the Chi-squared Automatic Interaction Detector (CHAID) (GV, 1978) and CART (Breiman, Friedman, Stone, & Olshen, 1984) techniques. Decision trees offer simplicity and flexibility in the employed modeling assumptions, while also allowing for easy visualization of the learned modeling strategies (obtained after training). On the negative side, the entailed variable discretization performed by these models results in potential loss of significant information, as well as overfitting proneness.

Another popular class of statistical models used for credit rating is hazard rate models. These models extend the time horizon of a rating system, by looking at the probability of default during the life cycle of the examined loan or portfolio (Chava & Jarrow, 2004; Shumway, 2001). To achieve this, hazard models explicitly model a survival function for the behavior of an examined borrower. Cox Proportional hazard models are one popular instance of this type of models (DR, 1972); it is based on the assumption that the covariates affecting the default rate are multiplicatively related to the hazard rate function (Im, Apley, Qi, & Shan, 2012).

Machine learning techniques have also been successfully applied to enhance the capabilities of conventional corporate credit scoring systems in several studies. Among such works, feedforward neural networks (FNNs) constitute the most commonly used machine learning method in the context of corporate credit rating systems (e.g., Zhao et al., 2015). Their successful application in the context of corporate credit rating is basically due to their nonlinear and non-Gaussian modeling assumptions, and their capability to capture dependencies between assets. On the negative side, the notorious proneness of FNNs to overfitting (and, thus, their limited generalization capacity), their need of tedious cross-validation to perform hyperparameter selection (e.g., network size selection), along with their black-box nature that hinders intuitive visualization of the obtained results, limit their potential appeal to the financial community. Other researchers have considered using support vector machines (SVMs) (Vapnik, 1998) to effect the credit rating task. Indeed, a significant number of studies published in the last decade have shown that SVMs outperform FNNs in credit rating scenarios (Chen & Li, 2014; Chen & Shih, 2006; Danenas & Garsva, 2015; Harris, 2015; Huang, 2009; Wang & Ma, 2012; Yi & Zhu, 2015), while reducing the possibility of overfitting, and alleviating the need of tedious cross-validation for the purpose of appropriate hyperparameter selection. On the negative side, SVMs also constitute black-box models, thus limiting their potential of offering deeper intuitions and visualizations regarding the obtained results of their modeling and inference procedure. A Bayesian inference-based analogous to SVMs, namely Gaussian processes, have also been considered by Huang (2011). A drawback of this approach is its high computational complexity, which is cubic to the number of available data points, combined with the assumption of normally distributed data, which is clearly unrealistic, as we have already explained. Finally, Random Forests (RFs) is another type of methods that has recently garnered attention by researchers working in the field of corporate credit rating. This sophisticated technique was introduced in (Breiman, 2001), while one successful application of RFs to the problem of corporate credit rating can be found in (Yeh, Lin, & Hsu, 2012).

Despite these considerable advances, all *existing work on intelligent corporate credit rating systems fails to address and to appropriately take into account two key characteristics of financial data: (i) their heavy-tailed nature; and (ii) the entailed temporal dynamics/evolution that characterize the financial behavior of corporate entities.*To resolve these issues, in this paper we develop a novel holistic corporate credit scoring system, that addresses all the parts of the modeling pipeline, from financial ratio time-series selection and preprocessing, to selection of appropriate time-series modeling techniques, and information fusion strategies used to obtain the final credit scores. At the heart of the proposed system lies a novel financial data modeling scheme, capable of modeling data with heavy-tailed nature and intricate temporal dynamics (typical in the field of finance), based on *Student's-t hidden Markov models (SHMMs)* (Chatzis, Kosmopoulos, & Varvarigou, 2009).

Our developed expert system constitutes an intricate data processing pipeline comprising a financial data preprocessing stage, developed on the basis of expert knowledge, and a core modeling stage, where SHMMs are used to capture salient temporal patterns in the modeled time-series that are associated with different credit risk scores. To perform modeling and prediction, our approach utilizes appropriate financial ratio time-series, based on the assumption that financial ratios carry all the information necessary to describe and predict the internal state of a company. Specifically, we use five-year historical data of financial ratios, that provide adequate insights on how profitable an examined company is, what the trends are, and how much risk is embedded in its business models. We fit distinct SHMMs to each one of the modeled financial ratios, and obtain separate credit scores from each one of them. Eventually, we train one final information fusion layer that combines the outputs of the individual SHMMs under a weighted linear combination scheme, to generate the final predictions obtained by our system. Parameter optimization of this final information fusion layer is performed by means of a simple yet effective genetic algorithm (GA) (Deb, Agrawal, Pratap, & Meyarivan, 2000).

SHMMs are a successful machine learning technique for modeling data with *temporal dynamics* (i.e., time-series data), that may contain a number of *outliers* and related artifacts in the available training datasets. As such, SHMMs arise as a natural selection for effecting the task of modeling financial ratio data, which