Expert Systems With Applications 58 (2016) 93–101

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction

Maciej Zięba[a,*], Sebastian K. Tomczak[b], Jakub M. Tomczak[a]

[a] *Department of Computer Science, Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland*

[b] *Department of Operations Research, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland*

ARTICLE INFO

Article history:
Received 26 January 2016
Revised 1 April 2016
Accepted 2 April 2016
Available online 6 April 2016

Keywords:
Bankruptcy prediction
Extreme gradient boosting
Synthetic features generation
Imbalanced data

ABSTRACT

Bankruptcy prediction has been a subject of interests for almost a century and it still ranks high among hottest topics in economics. The aim of predicting financial distress is to develop a predictive model that combines various econometric measures and allows to foresee a financial condition of a firm. In this domain various methods were proposed that were based on statistical hypothesis testing, statistical modeling (e.g., generalized linear models), and recently artificial intelligence (e.g., neural networks, Support Vector Machines, decision tress). In this paper, we propose a novel approach for bankruptcy prediction that utilizes Extreme Gradient Boosting for learning an ensemble of decision trees. Additionally, in order to reflect higher-order statistics in data and impose a prior knowledge about data representation, we introduce a new concept that we refer as to synthetic features. A synthetic feature is a combination of the econometric measures using arithmetic operations (addition, subtraction, multiplication, division). Each synthetic feature can be seen as a single regression model that is developed in an evolutionary manner. We evaluate our solution using the collected data about Polish companies in five tasks corresponding to the bankruptcy prediction in the 1st, 2nd, 3rd, 4th, and 5th year. We compare our approach with the reference methods.

1. Introduction

Prediction of an enterprise bankruptcy is of great importance in economic decision making. A business condition of either small or large firm concerns local community, industry participants and investors, but also influences policy makers and global economy. Therefore, the high social and economic costs as a consequence of corporate bankruptcies have attracted attention of researchers for better understanding of bankruptcy causes and eventually prediction of business distress (Zhang, Wang, & Ji, 2013).

The purpose of the bankruptcy prediction is to assess the financial condition of a company and its future perspectives within the context of long-term operation on the market (Constand & Yazdipour, 2011). It is a vast area of finance and econometrics that combines expert knowledge about the phenomenon and historical data of prosperous and unsuccessful companies. Typically, enterprises are quantified by a numerous indicators that describe their business condition that are further used to induce a mathematical model using past observations (Altman & Hotchkiss, 2010).

There are different issues that are associated with the bankruptcy prediction. Two main problems are the following: First, the econometric indicators describing the firm's condition are proposed by domain experts. However, it is rather unclear how to combine them into a successful model. Second, the historical observations used to train a model are usually influenced by imbalanced data phenomenon, because there are typically much more successful companies than the bankrupted ones. As a consequent, the trained model tends to predict companies as successful (majority class) even when some of them are distressed firms. Both of these issues mostly influence the final predictive capability of the model.

Previous works. First attempts of the formal bankruptcy prediction trace back to the beginnings of the 20th century when first econometric indicators were proposed to describe predictive abilities of business failure (Fitzpatrick, 1932; Merwin, 1942; Winakor & Smith, 1935). The sixties of the twentieth century brought a turning point in the survey of the early recognition of the business failure symptoms. First of all, the work of Beaver (1966) initiated

* Corresponding author. Tel.: +48713204150.
E-mail addresses: maciej.zieba@pwr.edu.pl (M. Zięba), sebastian.tomczak@pwr.edu.pl (S.K. Tomczak), jakub.tomczak@pwr.edu.pl (J.M. Tomczak).

http://dx.doi.org/10.1016/j.eswa.2016.04.001

application of statistical models to the bankruptcy prediction. Following this line of thinking, Altman (1968) proposed to use multidimensional analysis to predict corporate bankruptcy that was further developed by others (Altman & Loris, 1976; Blum, 1974; Deakin, 1972; Edmister, 1972; Ketz, 1978; Koh & Killough, 1990; Laitinen, 1991; Libby, 1975; Meyer & Pifer, 1970; Pettway & Sinkey, 1980; Rujoub, Cook, & Hay, 1995; Sinkey, 1975; Wilcox, 1973). In parallel, a great interest was paid to the generalized linear models that can be used in both decision making and providing certainty of the prediction (Aziz, Emanuel, & Lawson, 1988; Grice & Dugan, 2003; Hopwood, McKeown, & Mutchler, 1994; Koh, 1991; Li & Miu, 2010; Ohlson, 1980; Platt & Platt, 1990; Platt, Platt, & Pedersen, 1994; Zavgren, 1983; Zmijewski, 1984). Additionally, the generalized linear models are of special interest because estimated weights of the linear combination of economic indicators in the model can be further used to determine importance of the economic indicators.

Since nineties of the 20th century artificial intelligence and machine learning have become a major research direction in the bankruptcy prediction. In the era of increasing volumes of data it turned out that the linear models like the logistic regression or logit (probit) models are unable to reflect non-trivial relationships among economic metrics. Moreover, the estimated weights of the linear models are rather unreliable to indicate the importance of the metrics.

In order to obtain comprehensible models with an easy to understand knowledge representation, decision rules expressed in terms of first-order logic were induced using different techniques, naming only a few, like rough sets (Dimitras, Slowinski, Susmaga, & Zopounidis, 1999) or evolutionary programming (Zhang et al., 2013). However, the classification accuracy of the decision rules are very often insufficient, therefore, more accurate methods were applied to the bankruptcy prediction. One of the most successful model was support vector machines (SVM) (Shin, Lee, & Kim, 2005). The disadvantages of SVM are that the kernel function must be carefully hand-tuned and it is impossible to obtain comprehensible model.

A different approach aims at automatic feature extraction from data, i.e., automatic non-linear combination of econometric indicators, which alleviates the problem of a specific kernel function determination in the case of SVM. This approach applies neural networks to the bankruptcy prediction (Bell, Ribar, & Verchio, 1990; Cadden, 1991; Coats & Fant, 1991; Geng, Bose, & Chen, 2015; Koster, Sondak, & Bourbia, 1991; Salchenberger, Cinar, & Lash, 1992; Serrano-Cinca, 1996; Tam, 1991; Tam & Kiang, 1992; Wilson & Sharda, 1994; Zhang, Hu, Patuwo, & Indro, 1999). The main problem of the neural networks lies in the fact that they can fail in case of multimodal data. Typically the econometric metrics need to be normalized/standardized in order to have all features of the same magnitude. This is also necessary for training neural networks so that the errors could be backpropagated properly. However, the normalization/standardization of data do not reduce the problem of the data multimodality that may drastically reduce predictive capabilities of the neural networks. That is why it has been advocated to take advantage of different learning paradigm, namely, the ensemble of classifiers (Kittler, Hatef, Duin, & Matas, 1998). The idea of the ensemble learning is to train and combine typically weak classifiers to obtain better predictive performance. First approaches but still very successful were bagging (Breiman, 1996) and boosting (Freund & Schapire, 1996; Friedman, 2001; 2002; Zięba, Tomczak, Lubicz, & Świątek, 2014). The idea of boosting was further developed to the case of unequal classification costs (Fan, Stolfo, Zhang, & Chan, 1999) and imbalanced data (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). Recently, the boosting method was modified to optimize a Taylor expansion of the loss functions, an approach known as *Extreme Gradient Boosting* (Chen & He, 2015a) that obtains state-of-the-art results in many problems on Kaggle competitions.[1] Recently, it has been shown that the ensemble classifier can be successfully applied to the bankruptcy prediction (Nanni & Lumini, 2009) and it significantly beats other methods (Alfaro, García, Gámez, & Elizondo, 2008).

Contribution. In this paper we propose a novel method for bankruptcy prediction that makes use of Extreme Gradient Boosting (Chen & He, 2015b) for developing regularized boosted trees (Chen & He, 2015a; Johnson & Zhang, 2011). Best to our knowledge, such an approach was not applied to solve the problem of predicting financial condition of the companies. However, this method is successfully applied to many classification problems (Chen & He, 2015a) and widely used in winning Kaggle competitions. The model is also insensitive to imbalanced data phenomenon because it enables to select AUC measure for evaluation and forces proper ordering of the imbalanced data. To improve the prediction of the model we use ensemble of boosted trees, where each base learner is constructed using additional *synthetic features*. The synthetic features are developed at each boosting step in an evolutionary fashion by combining features using an arithmetic operation. Each synthetic feature can be seen as a single regression model. The purpose of the synthetic features is to combine the econometric indicators proposed by the domain experts into a complex features. The synthetic features can be seen as hidden features extracted by the neural networks but the fashion they are extracted is different. At the end, we test our solution using collected data about Polish companies.

Organization of the paper. The paper is organized as follows. In Section 2 the ensemble boosted trees is introduced as the model for bankruptcy prediction. In Section 3 we present the experimental results gained on real dataset representing the financial condition of the polish companies. The paper is summarized by the conclusions in Section 4.

2. Methodology

2.1. Extreme Gradient Boosting Framework

Let us denote by $\mathbf{x} \in \mathcal{X}$ a vector of features describing an enterprise, where $\mathcal{X} \subseteq \mathbb{R}^D$ and by $y \in \{0, 1\}$ a label representing whether the enterprise is bankrupt, $y = 1$, or not, $y = 0$. Further, we utilize decision trees as discriminative models, more precisely, Classification and Regression Trees (CART). A CART tree can be represented by the weights associated with the leaves in the tree structure

$$f_k(\mathbf{x}_n) = w_{q(\mathbf{x})}, \tag{1}$$

where $q(\mathbf{x}_n)$ is the function that takes an example $\mathbf{x}$ and returns the path id in the structure of the tree, $q : \mathbb{R}^D \to \{1, \dots, T\}$, T is the number of paths (leaves). A path is ended with a leaf that contains weight w_i.

We aim at learning an ensemble of K decision trees(Chen & He, 2015a)

$$h_K(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}), \tag{2}$$

where $f_k \in \mathcal{F}$, for $k = 1, \dots, K$, and $\mathcal{F}$ is a space of all possible decision trees (CART). In order to obtain a decision for new $\mathbf{x}$ one could calculate a conditional probability of a class for h_K as follows:

$$p(y = 1|\mathbf{x}) = \sigma\left(h_K(\mathbf{x})\right), \tag{3}$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the sigmoid function.

[1] www.kaggle.com/