Decision Support Systems 52 (2012) 539–547

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

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

Multistage classification by using logistic regression and neural networks for assessment of financial condition of company

Bartosz Swiderski [a], Jarosław Kurek [a], Stanislaw Osowski [b,c,*]

[a] University of Life Sciences, Poland
[b] Warsaw University of Technology, Poland
[c] Military University of Technology, Poland

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form 5 October 2011
Accepted 19 October 2011
Available online 26 October 2011

Keywords:
Multinomial ordinary regression
Assessment of financing condition
Neural networks
Support vector machine

ABSTRACT

The paper presents the new approach to the automatic assessment of the financial condition of the company. We develop the computerized classification system applying WOE representation of data, logistic regression and Support Vector Machine (SVM) used as the final classifier. The applied method is a combination of a classical binary scoring approach and Support Vector Machine classification. The application of this method to the assessment of the financial condition of companies, classified into five classes, has shown its superiority with respect to classical approaches.

1. Introduction

The problem of assessment of the financial condition of company is very crucial to avoid customers, who may cause problems with financial liquidity or are the potential bankrupts [1, 2, 19]. To decrease the risk of transaction the assessment of financial condition of the company is necessary. Company credit ratings are very costly to obtain, since they require to invest large amount of time and human resources to perform analysis of the company risk status. Such report is based on various aspects, ranging strategic competitiveness to the operational level details [2, 10].

Although rating agencies emphasize the importance of analysts' assessment in determining credit ratings, there is a parallel way of developing the automatic methods relying on the artificial intelligence approach. Nowadays the most often used methods apply the artificial neural networks, being the universal tools able to do both classification and prediction tasks [6, 11, 13, 17]. The important point in such application is the availability of large amount of historical data of the company, regarding the financial reports and embedded valuable expertise of the agencies in evaluating companies' credit risk levels. The objective of credit rating prediction is to build the mathematical models that can extract knowledge of credit risk evaluation from past observations and to apply it to evaluate credit risk of companies with much broader scope.

However besides the prediction the modeling of the process can deliver another valuable information to the user. These studies can help the user to capture the fundamental characteristics of the most important dependencies between different financial ratings and the company risk status. Such analysis can also simplify the process of the assessment of the risk status of the company, by eliminating some parameters that are loosely associated with the level of the company risk.

The most important point in the credit report is the assessment of the financial condition of company by using many factors (not only financial). Important are also such information as changes of board of directors, status of the company, location and other registry information. Very often this information can be available online in Internet. The other source of fresh information about company is the interview via phone with subject or other companies with which our subject cooperates e.g. suppliers, sister company, parent company, affiliates etc. On the basis of this we can obtain additional information about e.g. payment history. Other sources of data are the agents distributed all over the world. This additional information can be added to the set of attributes, enriching in this way the input information taken into account at taking decision.

On the basis of all gathered information we can create the diagnostic features describing the financial state of the company, and then associate them with one of few classes, representing the level of insolvency risk. The insurance companies apply different number of classes. In this paper we assume 5 classes of insolvency risk [10]:

- excellent (without any risk)
- good

* Corresponding author at: Warsaw University of Technology, 00-661 Warsaw, Koszykowa 75, Poland. Tel.: +48 22 234 7235; fax: +48 22 234 5642.
E-mail address: sto@iem.pw.edu.pl (S. Osowski).

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.10.018

- satisfactory
- passable
- poor.

The problem that arises is to provide the unified way of representing the information as an input to the computer system performing the role of automatic extraction of knowledge and undertaking the final decision of assessment of insolvency risk of the company. In most application the numerical data is either represented directly in numerical form or converted to some classes, while the other (non-numerical) data is somehow coded in a binary way. In this paper we apply the unified way of representing data by using the weight of evidence (WOE) concept [17] associated with each feature. The data represented by WOE will be classified by us in two step procedure. In the first step we apply the binary logistic regression associating the data with seven models of 2-class classification. The results in the form of probability of membership to these 7 models are applied as the input attributes for the second stage of classification recognizing the final class (one of 5 already defined). As the classifier we apply the support vector machine (SVM), generally regarded as the most efficient classification tool [6, 15]. The novelty of the paper may be characterized in the following points.

- Development of continuous representation of financial data, very efficient in practical application and leading to the improvement of the quality of the classification system.
- Proposing the 2-step classification of financial data by applying the binary classification systems in both stages. We will show that such solution leads to the significant improvement of the accuracy of classification.

2. Binary logistic regression and WOE approach

In the first step of proposed approach we transform the multiclass task into few simple binary models of classification (similar to decision tree) and then in the second stage apply the additional classifier responsible for undertaking the final decision of the classification. We will show that such dissolution of classification task into two steps is profitable and leads to the increase of the accuracy of an automatic system of assessment of financial condition of company. In this work we substitute the entries of vector $\mathbf{x}$ representing any real financial data by their numerical codes in the form of weights of evidence (WOE). Weights of evidence is a quantitative method for combining evidence in support of a hypothesis [4, 7].

Suppose we have only two classes labeled by either zero or one. Let us assume that the input features (represented by WOE) defined for the data under classification are organized in the form of vector $\mathbf{x}$ of the dimension N, where N denotes the number of input features. We assume that the actual target y has the binomial distribution i.e. $y_i \sim B(n_i, pd_i)$ where n_i denotes the number of trials and pd_i probability of success. This distribution is dependent on the set of input variables (features) x_i. Our goal is to estimate the conditional probability

$$pd(y_i = 1) = E\left(\frac{y_i}{n_i}\middle| x_i\right) \tag{1}$$

The model of logistic regression is now defined in the form [7, 14]

$$\ln\left(\frac{pd(y_i = 1)}{1 - pd(y_i = 1)}\right) = \mathbf{x}_i\boldsymbol{\alpha} \tag{2}$$

where $\mathbf{x}_i$ denotes the vector of input variables. From this model by solving the set of linear equations written for the learning data we can estimate the unknown vector $\boldsymbol{\alpha}$. To the most popular approaches to this estimation belongs the maximum likelihood method [2, 7]. Alternatively, one can express the previous formula in a probability category as

$$pd(y_i) = pd(y_i = 1) = \frac{1}{1 + \exp(-\boldsymbol{\alpha}\mathbf{x}_i)} \tag{3}$$

Then the output $pd(y_i)$ can be interpreted as the probability of belonging the input vector $\mathbf{x_i}$ to the class labeled by 1.

In the proposed approach the input vector $\mathbf{x}$ is represented by weights of evidence. Weights of evidence is a quantitative method for combining evidence in support of a hypothesis [2, 4]. Such representation enables to apply this model to the evidential themes with binary (presence/absence) classes as well as to multi-class maps. Binary evidence is relatively straightforward to interpret and we limit its definition for two classes. Let us assume for example that we deal with the input data called 'type of consolidation' of the company, recognizing three types of it: consolidated, non-consolidated and group consolidated. In each group there is some quantity of companies belonging to class 1 and some quantity of data classified as class 0 (binary representation of classes). Let us denote by $odds_i$ for ith attribute of the feature the ratio of the number of all companies belonging to class 0 and to the class 1, that is

$$odds_i = \frac{rate_i(y = 1)}{rate_i(y = 0)} \tag{4}$$

where the $rate_i(y = a)$ is calculated as the number of examples of representatives of class a ($a = 0$ or $a = 1$) related to the total population of the members of this class for all attributes of the particular feature (in the case of type of consolidation it will be three groups counted together: consolidated, non-consolidated and group consolidated). The weights of evidence for this category variable is defined for each ith attribute in the form

$$WOE_i = \ln(odds_i) \tag{5}$$

We illustrate the procedure of calculation of WOE for the input feature 'type of consolidation' on the exemplary data presented in Table 1.

After transformation, the input variable 'type of consolidation' is represented by its attribute's WOE. For example $WOE_1 = 1.427$, $WOE_2 = -0.088$ or $WOE_3 = -1.073$ will represent the consolidated, non-consolidated or group consolidated data, according to the particular type of the considered company.

We can use this approach to order any type of variables, including the category as well as numerical data. In the case of category data if some input variable has less than 20 categories we treat it as a standard one and code directly each category by WOE. In the case of higher number of categories we may merge the groups characterized by similar values of WOE, keeping the number of categories not higher than 20. In the case when the class contains less than 5% of data we may merge it with the class closest in respect to WOE and then we calculate the new value of WOE for the merged group.

Table 1
The example of determination of the WOE for the feature 'type of consolidation'.

i	Type of consolidation	Class 0	Class 1	Rate (y = 0)	Rate (y = 1)	Odds	WOE
1	Consolidated	30	5	0.375	0.090	4.166	1.427
2	Non-consolidated	40	30	0.500	0.546	0.915	−0.088
3	Group consolidated	10	20	0.125	0.364	0.342	−1.073
Sum		80	55	1	1		