Knowledge-Based Systems 73 (2015) 289–297

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

The effect of feature selection on financial distress prediction

Deron Liang [a], Chih-Fong Tsai [b,*], Hsin-Ting Wu [a]

[a] *Department of Computer Science and Information Engineering, National Central University, Taiwan*
[b] *Department of Information Management, National Central University, Taiwan*

ARTICLE INFO

Article history:
Received 17 December 2013
Received in revised form 7 September 2014
Accepted 11 October 2014
Available online 17 October 2014

Keywords:
Financial distress prediction
Bankruptcy prediction
Credit scoring
Feature selection
Data mining

ABSTRACT

Financial distress prediction is always important for financial institutions in order for them to assess the financial health of enterprises and individuals. Bankruptcy prediction and credit scoring are two important issues in financial distress prediction where various statistical and machine learning techniques have been employed to develop financial prediction models. Since there are no generally agreed upon financial ratios as input features for model development, many studies consider feature selection as a pre-processing step in data mining before constructing the models. However, most works only focused on applying specific feature selection methods over either bankruptcy prediction or credit scoring problem domains. In this work, a comprehensive study is conducted to examine the effect of performing filter and wrapper based feature selection methods on financial distress prediction. In addition, the effect of feature selection on the prediction models obtained using various classification techniques is also investigated. In the experiments, two bankruptcy and two credit datasets are used. In addition, three filter and two wrapper based feature selection methods combined with six different prediction models are studied. Our experimental results show that there is no the best combination of the feature selection method and the classification technique over the four datasets. Moreover, depending on the chosen techniques, performing feature selection does not always improve the prediction performance. However, on average performing the genetic algorithm and logistic regression for feature selection can provide prediction improvements over the credit and bankruptcy datasets respectively.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Financial distress prediction is very critical in enterprise risk management, especially for financial institutions. In particular, financial institutions have to develop various risk management models, such as bankruptcy prediction and credit scoring models [37,43]. For bankruptcy prediction, financial institutions need effective prediction models in order to make appropriate lending decisions. On the other hand, credit scoring models are used for the management of large loan portfolios and/or credit admission evaluation.

Specifically, bankruptcy prediction and credit scoring are two binary classification problems in financial distress prediction, which aim at assigning new observations to two pre-defined decision classes (e.g., 'good' and 'bad' risk classes) [40]. For example, bankruptcy prediction models are used to predict the likelihood that the loan customers will go bankrupt whereas credit scoring models are used to determine whether the loan applicants should be classified into a high risk or low risk group. In the literature, many supervised machine learning (or classification) techniques have been used for financial distress prediction [2,10,24,29].

Though many novel sophisticated techniques have been proposed for effective prediction, very few have examined the effect of feature selection on financial distress prediction. Feature selection is an important data pre-processing step of knowledge discovery in databases (KDD). The aim is to filter out unrepresentative features from a given dataset [11,17]. As there are no generally agreed financial ratios for bankruptcy prediction and credit scoring, collected variables must be examined for their representativeness, i.e., importance and explanatory power, in the chosen dataset [29]. Therefore, the performance of classifiers after performing feature selection could be enhanced over that of classifiers without feature selection.

Generally speaking, feature selection can be broadly divided into the filter, wrapper, and hybrid approaches [3,31]. The filter based method (usually based on some statistical techniques) evaluates and selects feature subsets by the general characteristics of the given dataset. The wrapper based method is based on a pre-determined mining algorithm and its performance is used as

[*] Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 4254604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

http://dx.doi.org/10.1016/j.knosys.2014.10.010
0950-7051/© 2014 Elsevier B.V. All rights reserved.

the evaluation criterion to select feature subsets. Specifically, it aims at searching for features that are better suited to the mining algorithm to improve the mining performance. The hybrid method is based on combining these two methods by exploiting different evaluation criteria in different search stages.

In recent studies, the filter and wrapper based feature selection methods have been widely used for bankruptcy prediction [41,13,14,27,25,26,8,4] and credit scoring [18,7,30,16,42,5]. Most studies apply either filter or wrapper based methods for single domain problems, i.e., bankruptcy prediction and credit scoring. One reason for the lack of using hybrid based feature selection methods is because currently there is no standard and representative method. In addition, there are no guidelines for which filter and wrapper based methods should be combined to select the best features for the later prediction performance.

One major limitation of current studies is that each work only considers one specific feature selection method for either bankruptcy prediction or credit scoring problems. In other words, there is no study focusing on comparing both types of feature selection methods for both bankruptcy prediction and credit scoring problems (c.f. Section 2.2). Therefore, the aim of this paper is to examine the effect of the filter and wrapper based feature selection methods on both bankruptcy prediction and credit scoring problems. Moreover, the effect of performing feature selection on different classification techniques will also be investigated.

The contributions of this paper are twofold. First, we provide a comprehensive study of comparing different filter and wrapper based feature selection methods in terms of two financial distress problems, which are bankruptcy prediction and credit scoring. In particular, the most suitable methods for these two specific problems are identified. As a result, the identified methods can also be regarded as the baseline feature selection methods for future related researches. Second, the research findings also allow us to understand which classification technique(s) are more sensitive to feature selection. Therefore, this can provide a guideline for future studies to choose suitable techniques for their prediction models.

The rest of this paper is organized as follows. Section 2 overviews related literature about filter and wrapper based feature selection methods. Moreover, related works are compared in terms of the feature selection methods employed, prediction methods constructed, etc. Sections 3 and 4 present the experimental setup and results, respectively. Finally, Section 5 concludes the paper.

2. Literature review

2.1. Feature selection

As there are no generally agreed factors (i.e., variables) for bankruptcy prediction and credit scoring, some of the collected variables as features may contain noise that could affect the prediction result. On the other hand, if too many features were used for data analysis, it can cause high dimensionality problems [36]. In data mining, feature selection or dimensionality reduction can be approached to reduce irrelevant or redundant features. This is an important data pre-processing technique in data mining, which aims at selecting more representative features having more discriminatory power over a given dataset [11,17].

Feature selection can be defined as the process of choosing a minimum subset of m features from the original dataset of n features ($m < n$), so that the feature space (i.e. the dimensionality) is optimally reduced according to four steps, which are subset generation, subset evaluation, stopping criteria, and result validation [11,31].

In general, subset generation is a search procedure which generates subsets of features for evaluation. Each subset generated is evaluated by a specific evaluation criterion and compared with the previous best one with respect to this criterion. If a new subset is found to be better, then the previous best subset is replaced by the new subset.

2.2. Filter based feature selection

The filter based feature selection methods usually contain the following procedures. Given a dataset, the method based on a particular search strategy initially searches from a given subset, which may be an empty set, a full set, or any randomly selected subset. Then, each generated subset is evaluated by a specific measure and compared with the previous best one. This search process iterates until the pre-defined stopping criterion is met. Consequently, the final output of this method is the last current best subset.

More specifically, the search strategy and evaluation measure can be different depending on the algorithms used. In addition, filter based methods do not involve any mining algorithm during the search and evaluation steps, they are computationally efficient. Some examples of filter based methods that are used in financial distress prediction are based on statistical techniques, such as *t*-testing, principal component analysis, discriminant analysis, and regression [4,8,37,26,41,45].

2.2.1. Discriminant Analysis

Linear Discriminant analysis (LDA) is used to find a linear combination of features which characterizes or separates two or more classes of objects. The resulting combination can be used for dimensionality reduction. LDA can also be used to express one dependent variable as a linear combination of other features. In other words, LDA looks for the linear combination of features which best explains the given data [34].

LDA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is

$$D = v_1X_1 + v_2X_2 + v_3X_3 + \cdots v_iX_i + a \tag{1}$$

where D is the discriminant function, v is the discriminant coefficient or weight for that feature, X is the respondent's score for that feature, a is a constant, and i is the number of predictor features.

2.2.2. t-Test

The *t*-test method is used to determine whether there is a significant difference between two group's means. It helps to answer the underlying question: Do the two groups come from the same population, and only appear differently because of chance errors, or is there some significant difference between these two groups? Three basic factors help determine whether an apparent difference between two groups is a true difference or just an error due to chance [35]:

1. The larger the sample, the less likely that the difference is due to sampling errors or chance.
2. The larger the difference between the two means, the less likely that the difference is due to sampling errors.
3. The smaller the variance among the participants, the less likely that the difference is created by sampling errors.

2.2.3. Logistic regression

Logistic regression (LR) is a type of probabilistic statistical classification model. LR measures the relationship between a categorical dependent variable and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variables. LR allows us to look at the fit of the model as well as at the

Download English Version:

https://daneshyari.com/en/article/404985

Download Persian Version:

https://daneshyari.com/article/404985

Daneshyari.com