Knowledge-Based Systems 27 (2012) 333–342

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Simple instance selection for bankruptcy prediction

Chih-Fong Tsai *, Kai-Chun Cheng

Department of Information Management, National Central University, Taiwan

ARTICLE INFO

Article history:
Received 11 April 2011
Received in revised form 22 August 2011
Accepted 24 September 2011
Available online 1 October 2011

Keywords:
Instance selection
Outlier detection
Data mining
Bankruptcy prediction
Clustering

ABSTRACT

Instance selection or outlier detection is an important task during data mining, which focuses on filtering out bad data from a given dataset. However, there is no rigid mathematical definition of what constitutes an outlier and an outlier is not a binary property. Therefore, different volumes of outliers may be detected depending on the setting of the threshold for what constitutes an outlier, e.g., the distance in distance-based outlier detection. In this study, we examine bankruptcy prediction performance achieved after removal of different outlier volumes from four widely used datasets, namely the Australian, German, Japanese, and UC Competition datasets. Specifically, a simple distance-based clustering outlier detection method is used. In addition, four popular classification techniques are compared, artificial neural networks, decision trees, logistic regression, and support vector machines. Experiments are conducted to examine (1) the prediction performance of the bankruptcy prediction models with and without instance selection, (2) the stability of bankruptcy prediction models after the removal of outliers from the testing set, and (3) the characteristics of these four different datasets. The results show that with the German dataset it is much more difficult for the prediction models to provide high rates of accuracy after outlier removal, while it is easier with the UC Competition dataset. Removing 50% of the outliers can lead to optimal performance of these four models. In addition, using the removed outliers to test the prediction accuracy of these models, we find that it is support vector machines (SVM) that provide the highest rate of prediction accuracy and perform with much more stability and good noise tolerance than the other three prediction models. Furthermore, the prediction accuracy of the SVM model followed by instance selection is similar to the one without instance selection (i.e., the SVM baseline). In other words, the difference in performance between the SVM and the SVM baseline is the least of the three models in comparison with their corresponding baselines.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Background

Bankruptcy prediction is a very important for financial institutions in order to make the best possible lending decisions. Incorrect decisions are very likely to cause financial crises and distress [25]. As a result, many bankruptcy prediction models have been proposed using data mining techniques in the literature [18].

Pre-processing of data is an important step for good quality data mining in data mining or knowledge discovery in database (KDD) processes. For example, if too many instances are considered, it can result in large memory requirements and slow execution speed, and can cause over-sensitivity to noise [32]. The aim of data-processing is to filter out any unrepresentative features or noisy data from a given dataset, which are likely to degrade the mining performance.

One problem with using the original data points is that there may not be any located at the precise points that would make for the most accurate and concise concept description [32]. Outlier detection is designed to find those observations in a random sample that lie an abnormal distance away from other values in a population. Outliers are basically unusual observations (or bad data points) that are far removed from the mass of data [1,3]. As noted above outlier detection is an important KDD task [17], and filtering out the detected outliers is very useful for obtaining good mining results. For this purpose, classifiers trained by the selected instances as a subset of original instances can provide relatively good performance. In data mining, the aim of instance selection is similar to that of outlier detection (or record reduction) [20].

In general, outlier detection methods can be classified into (statistical) parametric or non-parametric methods. In the parametric methods, it is either assumed that there is a known underlying distribution of the observations or they are determined based on statistical estimates of unknown distribution parameters [3,12]. However, these methods are unsuitable for datasets which are high-dimensional and for cases without prior knowledge of the underlying data distribution [5].

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

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.09.017

The non-parametric methods, also called distance-based methods, are usually based on local distance measures, i.e., the distance between a point and its nearest neighbor. This distance is regarded as a measure for identifying the outliers [2,16,23]. In particular, in these distance-based outlier detection methods it is assumed that outliers are the top n data points whose distance to their kth nearest neighbor is greatest. Another approach is based on the distances to neighboring data points obtained using a clustering algorithm [11].

1.2. Motivation

The focus in most past studies related to bankruptcy prediction has been on developing new algorithms for effective prediction [18,26,9]. There have been very little done considering further examination of the data mining pre-processing stage. For example, Tsai [27] compared a number of different feature selection (or dimensionality reduction) methods using five related datasets. Li et al. [19] proposed a random subspace binary logit (RSBL) model for prediction of corporate failure in China. Their results showed the RSBL model to be a significant improvement in terms of predictive ability over classical statistical models, such as multivariate discriminant analysis, logit models, and probit models. For other related studies of feature selection, please refer to Pacheco et al. [21], Unler and Murat [28], and Bermejo et al. [6].

However, none of existing studies offer further analysis of the effect of removing different outlier volumes on bankruptcy prediction for different datasets. Thus, the following three main problems are addressed in this paper:

- Instance selection has proved to be important in the KDD process, but it has not been fully explored in the domain problem of bankruptcy prediction. Bankruptcy prediction is a special case of instance selection, because the collected data sample is usually small, especially for bankruptcy. This means that it is important to examine the performance of bankruptcy prediction models with and without instance selection, respectively.
- Which classification techniques can perform better when they are tested over the removed outliers? In general, prediction performance is based on using some method to split a given dataset into both training and testing datasets, e.g., n-fold cross validation. However, the removed outliers are not directly used as testing data to further evaluate the performance of the prediction models which offers some advantages. First, the test results can be regarded as a model of a real work problem, which is likely to be very challenging. Second, the test results can show the stability and/or robustness of the models since much more difficult data to be predicted are used.
- Intuitively, when the number of outliers removed from a data set becomes large, the accuracy of the prediction models which use the remaining dataset for training and testing will necessarily increase. This is because the 'bad' data (i.e., the outliers), which are difficult to recognize, are filtered out. However, it is unknown how many of the outliers removed could allow the prediction models to correctly distinguish all of the remaining cases. There are a number of different (public) datasets for bankruptcy prediction simulation mentioned in the literature. It is necessary to analyze this issue in relation to these datasets in order to understand their characteristics. Moreover, the analysis results can provide some guidelines about how many of the remaining data should be used to train the prediction models. The can then be compared with the prediction models without instance selection (i.e., the first question) and how many outliers can be used as a suitable validation set (i.e., the second question).

Regarding these two questions, the aim of this work is to use a simple distance-based clustering method (c.f. Section 3.2) for the detection and removal of outliers from four related bankruptcy prediction datasets. Specifically, the prediction performances of the four classification techniques obtained by removing different volumes of outliers will be examined. These classification techniques include artificial neural networks, decision trees, logistic regression, and support vector machines.

Consequently, the major contribution of this paper is twofold. First of all, this study is the first attempt to consider instance selection in the bankruptcy prediction domain problem. Secondly, the outliers removed can be used as another type of test dataset to examine the stability of the prediction models.

The rest of this paper is organized as follows. Section 2 briefly overviews the distribution, distance, and density based approaches for outlier detection. Section 3 describes the experimental setup including the datasets chosen for experiments, the process of detecting and removing outliers, the classification techniques used for comparisons, and the evaluation methods. Section 4 presents the experimental results corresponding to the two research questions. Finally, some conclusions are provided in Section 4.

2. Outlier detection

2.1. Distribution-based outlier detection

Distribution-based approaches, which originate from the field of statistics, generally rely on the assumption of an underlying known distribution of the data. Statistical models are developed from the given dataset. A statistical test is then applied to determine whether an object belongs to this model or not. Objects, which have a low probability of belonging to this statistical model, are regarded as outliers. However, the distribution-based approaches cannot be applied in a high-dimensional data set since they are univariate in nature. In addition, such approaches are difficult to apply in practical applications because of the lack of prior knowledge of data distribution [3].

2.2. Distance-based outlier detection

Distance-based outlier detection is a typical top-n outlier detection approach, in which the top-n kth-Nearest Neighbor (kNN) distance is examined to determine the outliers. In other words, the distance from an object to its kth nearest neighbor indicates the outlier-ness of the object. If the neighboring points are relatively close, then the object is considered to be normal; if the neighboring points are far away, then the object is considered to be unusual [16,4]. One advantage of this approach is that based on a distance metric it can be applied to any feature space.

For a given distance measure in a feature space, there are many different definitions of distance-based outliers. Bay and Schwabacher [4] offer the following three popular definitions of distance-based outliers:

- Examples of outliers occur when at least a fraction p of the objects lies at a greater distance d.
- Outliers are the top n examples whose distance to the kth nearest neighbor is greatest.
- Outliers are the top n examples whose average distance to the k nearest neighbors is greatest.

2.3. Density-based outlier detection

Density-based approaches are based on computing the density of regions in the data where the outliers are defined as the objects

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات