Expert Systems with Applications 39 (2012) 13267–13276

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

On the use of data filtering techniques for credit risk prediction with instance-based models

V. García [a], A.I. Marqués [b], J.S. Sánchez [a,*]

[a] Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
[b] Department of Business Administration and Marketing, Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain

ARTICLE INFO

Keywords:
Finance
Credit risk
Instance selection
Outlier
Filtering
Editing
Nearest neighbour rule

ABSTRACT

Many techniques have been proposed for credit risk prediction, from statistical models to artificial intelligence methods. However, very few research efforts have been devoted to deal with the presence of noise and outliers in the training set, which may strongly affect the performance of the prediction model. Accordingly, the aim of the present paper is to systematically investigate whether the application of filtering algorithms leads to an increase in accuracy of instance-based classifiers in the context of credit risk assessment. The experimental results with 20 different algorithms and 8 credit databases show that the filtered sets perform significantly better than the non-preprocessed training sets when using the nearest neighbour decision rule. The experiments also allow to identify which techniques are most robust and accurate when confronted with noisy credit data.

1. Introduction

Modern-day finance is a broad field of business activity that comprises hard decision-making problems related to risk management. Financial risk refers to the uncertainties associated with any form of financing, including credit risk, business risk, investment risk, market risk, and operational risk (Horcher, 2005). The focus of this paper is on credit risk, which denotes the probability that a loan to a borrower will not be repaid. In this context, the credit risk prediction models intend to estimate whether a new credit applicant will default (bad applicant) on a loan or not (good applicant). The development of reliable prediction models is of major importance because the use of inadequate credit risk assessment tools constitutes one of the main reasons of enterprise bankruptcy.

The development of financial prediction methods is a complex process, which involves data collection and preprocessing, model design, validation and implementation. With regard to model design, numerous strategies have been proposed, including statistical methods, computational intelligence techniques and operations research methodologies (Abdou & Pointon, 2011; Ince & Aktan, 2009; Khandani, Kim, & Lo, 2010; Khashman, 2010; Wozabal & Hochreiter, 2012). In all cases, however, the quality of the data represents a critical point to further obtain accurate predictions. This mainly depends on the adequacy of the data in terms of the number of examples, the relevance of the attributes (independent variables) used in the analysis and the presence of outliers in the data set, among other issues. As a result, data preprocessing becomes a crucial step in real-world classification and prediction problems, as is the case of credit risk assessment.

Whilst some characteristics of data have largely been studied in the credit management literature, others have received relatively little attention so far. For instance, a lot of research effort has been devoted to the problem of attribute relevance by using new and existing feature selection algorithms (Chen & Li, 2010; Liu & Schumann, 2005; Piramuthu, 1999; Shukai, Chaudhari, & Dash, 2010; Wang, Hedar, Wang, & Ma, 2012). In contrast, despite its apparent influence on the performance of the prediction models, very few studies have addressed the problem of credit data with noise and outliers. Kotsiantis, Kanellopoulos, and Tampakas (2006) presented a survey of data preprocessing techniques for financial prediction, including discretization, feature selection and instance selection. Tsai and Chou (2011) used a genetic algorithm to perform feature selection and data filtering for bankruptcy prediction. Tsai and Cheng (2012) explored the performance of artificial neural networks, decision trees, logistic regression and support vector machines after removing different amounts of outliers from credit data sets.

Outliers have traditionally been defined as observations that appear to be inconsistent with the rest of the data. Nowadays, this term is being employed to cover a broad range of circumstances, including noisy and atypical data, new unidentified objects that do not belong to any of the predefined classes, and also mislabelled instances. Since the presence of outliers adversely affect the performance of any classification or prediction model, it is necessary to filter the data in order for eliminating the instances that disturb

* Corresponding author. Tel.: +34 964 728350; fax: +34 964 728435.
E-mail address: sanchez@uji.es (J.S. Sánchez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.075

the generalization process. Identifying outliers becomes especially important with instance-based learning methods, as for example is the case of the k nearest neighbours decision rule.

Although instance selection has proved to be effective in many data mining and knowledge discovery applications, it has not been fully explored in the domain of credit management. This paper faces the problem of outlier removal in the context of credit risk prediction with the k nearest neighbours model. To this end, several instance selection algorithms of different nature will be used to filter out outliers from the training set, thus trying to obtain a cleaner representation of the credit data. It has to be pointed out that the aim of this work is not to search for the most effective instance selection method, but to test whether the application of these preprocessing techniques produces some increase in the performance of credit risk prediction models.

2. Instance-based models

From a practical point of view, the credit risk prediction problem can be deemed as a binary classification problem where a new input sample (the credit applicant) must be categorized into one of the predefined classes based on a number of observed variables or attributes related to that sample. The input of the classifier consists of a variety of information that describes socio-demographic characteristics and economic conditions of the applicant, and then the classifier has to produce the output in terms of the applicant creditworthiness.

Formally, the credit risk prediction problem can be described as follows. Given a set of applicants $T = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$, where each sample x_i is characterized by m attributes, x_{i1}, $x_{i2}, \ldots x_{im}$, and y_i denotes the class (good/bad applicant), then the credit risk classification problem consists of constructing a model δ to predict the value y for a new input applicant $\mathbf{x}$, that is, $\delta(\mathbf{x}) = y$.

A large number of prediction models compute a distance between the input applicant $\mathbf{x}$ and the stored samples in T (called training set) when generalizing. These classification models are referred to as instance-based learning algorithms, where one of the most straightforward examples is the k nearest neighbours (k-NN) decision rule (Dasarathy, 1991).

The k-NN prediction model is a standard non-parametric technique used for probability density function estimation and classification. In brief, this classifier consists of assigning a new input sample to the class most frequently represented among the k closest instances in the training set, according to a suitable dissimilarity measure (e.g., Euclidean distance, Manhattan distance). A particular case is when $k = 1$, in which an input sample is decided to belong to the class indicated by its closest neighbour.

The characteristics of the k-NN classifier need the entire training set stored in computer memory, what may cause large time and memory requirements. On the other hand, this technique is extremely sensitive to the presence of noisy, atypical and erroneously labeled examples in the training set. Nevertheless, it has been found to perform better than other more sophisticated methods when dealing with arbitrarily complex problems, and it is a fairly intuitive procedure that could be easily implemented and explained to analysts.

3. Instance selection

In machine learning and data mining, the problem of instance selection is primarily related to instance deletion as irrelevant and harmful instances are removed from the original training set while retaining only critical instances. In general, the aim of the instance selection techniques is to produce a representative and relevant subset of the training set with similar or even higher generalization performance of the prediction model. An important advantage of instance selection against instance generation or abstraction refers to the fact that it chooses relevant examples without generating new artificial data, which do not make sense in many real-world applications. For instance, the incorporation of artificially generated samples into a credit data set may distort some socio-economic conditions of the problem in hand.

Let T be a training set and let $R \subseteq T$ be the reference set, that is, the subset of examples selected by some instance selection algorithm. Usually, the reference set will be much smaller than the original training set ($|R| \ll |T|$). The class label y of a new input case $\mathbf{x}$ will be then estimated by a given prediction function δ using the reference set R instead of T. The general framework for the application of an instance selection algorithm is as follows: some instance selection procedure chooses an appropriate reference set R from the training set T, then a learning algorithm is applied to build the prediction model δ and finally, this is evaluated using an independent test set S.

Traditionally, instance selection methods have been divided into two groups: editing and condensing. The goal of editing (or filtering) algorithms is to remove noisy, atypical and mislabelled instances and clean the possible overlapping between regions from different classes, thereby producing smoother decision boundaries and improving generalization. On the other hand, condensing (or thinning) algorithms aims at discarding superfluous or redundant instances that will not significantly affect the performance of the prediction model. Apart from these two general families of techniques, we can find hybrid methods that search for a small subset of the training set by simultaneously removing both outliers and superfluous instances.

According to their location in the input space, samples can be categorized into two main types: border and internal samples. Border samples are close to the decision boundaries, while internal samples are within the region of a class. Under this taxonomy, another factor that distinguishes instance selection techniques is whether they pursue to retain border points or internal points (Wilson & Martinez, 2000). The filtering algorithms seek to remove border samples because the internal examples are considered as the most representative of each class. Conversely, the condensing algorithms discard internal points under the idea that they do not affect the decision boundaries as much as the border samples and therefore, they can be removed with relatively little effect on the prediction process.

Yet another distinction among the instance selection algorithms can be made in terms of the direction of search for a subset of instances (García, Derrac, Cano, & Herrera, 2012; Wilson & Martinez, 2000). An incremental search begins with an empty subset R and adds the training instances in T that fulfill some criterion; in this case, the order of presentation of instances can result very important. The decremental search starts putting all training instances into the reference set R, and then searches for instances to be removed from R; the order of presentation is also important, but unlike the incremental process, all training examples are available for examination at any time. Another way to apply an instance selection process is in batch mode, which involves deciding if each instance in T meets the removal criterion before discarding any of them; then, all instances that satisfy such a criterion are removed from R at once. Finally, a mixed search begins with an initially preselected subset R, obtained either by random selection or by an incremental or decremental process, and then iteratively allows for additions or removals of instances that fulfill some criterion.

It has to be pointed out that only filtering techniques will be here taken into account because the focus of the present study is mainly on improving the performance of credit risk prediction models rather than on reducing the data set size. Some of the benefits of using filters are: (i) they make a prediction problem easier