Expert Systems with Applications 37 (2010) 3326–3336

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Multiple classifier application to credit risk assessment

Bhekisipho Twala *

Modelling and Digital Science, Council for Science and Industrial Research (CSIR), P.O. Box 395, Pretoria 0001, South Africa

ARTICLE INFO

Keywords:
Machine learning
Supervised learning
Statistical pattern recognition
Ensemble
Credit risk prediction
Noise

ABSTRACT

Credit risk prediction models seek to predict quality factors such as whether an individual will default (bad applicant) on a loan or not (good applicant). This can be treated as a kind of machine learning (ML) problem. Recently, the use of ML algorithms has proven to be of great practical value in solving a variety of risk problems including credit risk prediction. One of the most active areas of recent research in ML has been the use of ensemble (combining) classifiers. Research indicates that ensemble individual classifiers lead to a significant improvement in classification performance by having them vote for the most popular class. This paper explores the predicted behaviour of five classifiers for different types of noise in terms of credit risk prediction accuracy, and how such accuracy could be improved by using classifier ensembles. Benchmarking results on four credit datasets and comparison with the performance of each individual classifier on predictive accuracy at various attribute noise levels are presented. The experimental evaluation shows that the ensemble of classifiers technique has the potential to improve prediction accuracy.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

With the growth in financial services, there have been mounting losses from delinquent loans. For example, Manufacturer's Hanover's $3.5 Billion commercial property portfolio was burdened with $385 Million in non-performing loans (Rosenberg & Gleit, 1994). Thus, credit risk prediction is a critical part of a financial institution's loan approval decision processes.

According to BIS (2004), credit risk is most simply defined as the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms. Over the last decade, a number of the world's largest banks have developed sophisticated systems in an attempt to model the credit risk arising from important aspects of their business lines. Such models are intended to aid banks in quantifying, aggregating and managing risk across geographical and product lines. The outputs of these models also play increasingly major roles in banks' risk management and performance measurement processes, including performance-based compensation, customer profitability analysis, risk-based pricing and, to a lesser (but growing) extend, active portfolio management and capital structure decisions. Thus, applied finance researchers and practitioners remain concerned with prediction accuracy when building credit modelling systems.

Most techniques for predicting attributes of a credit risk system or credit data require past data from which models will be constructed and validated. One of the major problems for applying ML algorithms in credit risk prediction is the unavailability, scarcity and incompleteness (Schafer, 1997) of credit data, i.e., data for training the model. Most of the financial institutions do not share their data with other organizations so that a useful database with a great amount of data cannot be formed. In addition, surveys for collecting credit data are usually small but difficult and expensive to conduct.

Another important and common issue faced by researchers who use financial or credit datasets is the occurrence of noise in the data. Even if part of a well thought out measurement programme, credit datasets can be noisy for a number of reasons. These include inaccurate or non-reporting of information (without a direct benefit, a project manager or developer might see data collection as an overhead they can ill afford, for example), or, where data from a number of different types of customers or from a number of banks are combined, certain fields may be blank because they are not collectable for all customers. Often data is collected either with no specific purpose in mind (i.e., it is collected because it might be useful in future) or the analysis being carried out has a different goal than that for which the data was originally collected. In research datasets, e.g., experiments on human subjects to assess the effectiveness of a new credit risk technique, say, dropout or failure to follow instructions may lead to noise in data. The relevance of this issue is strictly proportional to the dimensionality of the collected data.

Economics and finance researchers have become increasingly aware of the problems and biases which can be caused by noisy data. Moreover, many credit datasets tend to be small with many different attributes – credit risk datasets grow slowly, for example

* Tel.: +27 128414711; fax: +27 128413550.
E-mail address: btwala1@csir.co.za

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.10.018

– and the numbers of available human subjects limit the size of many experimental datasets. Thus, we can ill afford to reduce our sample size further by eliminating all instances with noise. Because of the expense and difficulty of performing extensive controlled experiments on credit, case studies are often resorted to.

Handling noisy data is an important issue for classifier learning since the occurrence of noise in either the training or testing (classification) sets affect the prediction accuracy of learned classifiers. This can pose serious problems for researchers. For example, the introduction of bias and can result in misleading conclusions drawn from a research study. Also, inappropriate handling of noise in datasets can limit generalizability of the research findings (Byrne, 2000). For example, if you embark on formal statistical analysis, you may miss the important feature of your data. Also, formal statistical analysis assumes some characteristics about your data such as the number of instances, the number of attributes, the number of classes and so on. If these assumptions are wrong, the results of statistical analysis may be quite misleading and invalid. The seriousness of this problem depends in part on the proportion of noise and the type of attribute which contains noise. Therefore, it is important to consider how much noise and on which attributes is the noise when assessing the impact of attribute noise in data.

Various ML and statistical pattern recognition (SPR) techniques have been used in finance to predict credit risk. Reviews of the use of ML in report that ML in finance is a mature technique based on widely-available tools using well understood algorithms. A central concern of these applications is the need to increase the scoring accuracy of the credit decision. An improvement in accuracy or even a fraction of a percent translates into a significant future savings. In recent years, there has been an explosion of papers in the ML and statistics communities discussing how to combine models or model predictions. Many works in both the ML and statistical pattern recognition communities have shown that combining (ensemble) individual classifiers (Kittler, Hatef, Duin, & Matas, 1998) is an effective technique for improving classification accuracy.

An ensemble is generated by training multiple learners for the same task and then combining their predictions. There are different ways in which ensembles can be generated, and the resulting output combined to classify new instances. The popular approaches to creating ensembles include changing the instances used for training through techniques such as bagging (Bauer & Kohavi, 1999; Breiman, 1996), boosting (Schapire, 1990; Drucker, Cortes, Jackel, Lecun, & Vapkin, 1994; Freund & Schapire, 1996), stacking (Wolpert, 1992), changing the features used in training (Ho, 1995), introducing randomness in the classifier itself (Dietterich, 2000).

Bagging constructs a set of classifiers by sub-sampling the training examples to generate different hypotheses. After the different hypotheses are generated, they are combined by a voting mechanism. Boosting also uses the voting system to combine the classifiers. But, instead of sub-sampling the training examples, it generates the hypotheses sequentially. In each repetition, a new classifier is generated which focus in those instances that were handled incorrectly by the previous classifier. This is achieved by giving a weight to each instance in the training examples and adjusting these weights according to its importance after every iteration. Both, bagging and boosting use classifiers generated by the same base-learning algorithm and obtained from the same data. Finally, stacking can combine classifiers obtained from different learning algorithms using a high level classifier – the meta-classifier – to combine the lower level models. This is based on the fact that different classifiers are obtained from the same data and different learning algorithms use different biases to search the hypothesis space. This approach expects that the meta-classifier will be able to learn how to decide between the predictions provided by the base classifiers, in order to get accuracies better than any of them, much in the same way as a committee of experts. For purposes of this paper we follow the bagging approach.

Robustness has a twofold meaning in terms of dealing with noise using supervised classifiers. The toleration of noise in training data is one, and the toleration of noise data in test data is the other. Data presented to a given classifier, during either training or testing phase, may be noise in one or more ways. For example, attribute values and/or class labels could be noisy. For purposes of this paper we are assuming that the class labels are not noisy, i.e., only attribute values are considered as containing noise. Although the problem of noisy data has been treated adequately in various real-world datasets, there are rather few published works or empirical studies concerning the task of assessing learning and classification accuracy of supervised ML algorithms given noisy data (Aha, 1992). In addition, to the best of our knowledge, no study has been carried out on the effect of ensemble classifiers on credit risk predictive accuracy. In this paper, we first study the robustness of five classifiers on the predictive accuracy given noisy data. Then, we propose 20 ensemble methods from a combination of five classifiers. Each ensemble has two classifiers as elements. The proposed method utilizes probability patterns of classification results.

There are various reasons why the five classifiers were utilized to investigate the problem considered in this paper. Despite being one of the well known algorithms from the ML and SPR communities, they are a reasonable mix of non-parametric and parametric and they work for almost all classification problems. In addition, they can achieve good performance on many tasks.

The following two sections briefly give details of the five classifiers used in this paper. Section 4 reviews some related work to the problem of credit risk prediction in the economics and finance areas. Section 5 empirically explores the robustness and accuracy of five classifiers to four credit datasets with artificially simulated attribute noise. This section also presents empirical results from the application of the proposed ensemble procedure. We close with conclusions and directions for future research.

2. Classifiers

The most important feature of a problem domain, as far as the application of ML and SPR algorithms are concerned, is the form that the data takes and the quality of the data available. Our main focus will be on the latter. The problem of handling noise has been the focus of much attention in the ML and SPR communities. Specific ML and SPR techniques that are known to be robust enough to cope with noisy data, and to discover laws in it that may not always hold but are useful for the problem at hand, are now going to be described. These classifiers have also been used as credit scoring models (Hand & Henley, 1997). Three supervised learning (artificial neural network, decision tress and naïve Bayes classifier) and two statistical (k-nearest neighbour and logistic discrimination) techniques are examined in the presence of increasing level of artificial noise. First, the supervised learning techniques are described and a brief description of SPR techniques is briefly introduced.

2.1. Supervised learning techniques

2.1.1. Artificial neural networks

Artificial neural networks (ANNs), usually non-parametric approaches, are represented by connections between a very large number of simple computing processors or elements (neurons), have been used for a variety of classification and regression problems. These include pattern and speech recognition (Ripley,

Download English Version:

https://daneshyari.com/en/article/388701

Download Persian Version:

https://daneshyari.com/article/388701

Daneshyari.com