Expert Systems With Applications 73 (2017) 1–10

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A comparative study on base classifiers in ensemble methods for credit scoring

Joaquín Abellán*, Javier G. Castellano

Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

ARTICLE INFO

Article history:
Received 1 August 2016
Revised 12 December 2016
Accepted 13 December 2016
Available online 14 December 2016

Keywords:
Credit scoring
Ensembles of classifiers
Base classifier
Decision trees
Imprecise Dirichlet model
Uncertainty measures

ABSTRACT

In the last years, the application of artificial intelligence methods on credit risk assessment has meant an improvement over classic methods. Small improvements in the systems about credit scoring and bankruptcy prediction can suppose great profits. Then, any improvement represents a high interest to banks and financial institutions. Recent works show that ensembles of classifiers achieve the better results for this kind of tasks. In this paper, it is extended a previous work about the selection of the best base classifier used in ensembles on credit data sets. It is shown that a very simple base classifier, based on imprecise probabilities and uncertainty measures, attains a better trade-off among some aspects of interest for this type of studies such as accuracy and area under ROC curve (AUC). The AUC measure can be considered as a more appropriate measure in this grounds, where the different type of errors have different costs or consequences. The results shown here present to this simple classifier as an interesting choice to be used as base classifier in ensembles for credit scoring and bankruptcy prediction, proving that not only the individual performance of a classifier is the key point to be selected for an ensemble scheme.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The *sub-prime mortgage crisis of 2007* caused a ripple effect throughout the economy and it was the trigger (Longstaff, 2010) for the *global financial crisis of 2008* (also called *great credit crisis*), which is considered by many economists the worst financial crisis since the Great Depression of the 1930s (Almunia, Bénétrix, Eichengreen, O'Rourke, & Rua, 2010; Temin, 2010). Therefore, the analysis of credit risk has become more essential than ever before.

Furthermore, since the Basel second accord from 2004, known as Basel II and released by the Basel Committee on Banking Supervision, the supervised financial institutions are required to use internal ratings to measure credit risk. The need to control the credit risk has led to the banks and financial institutions to enhance the methods for this purpose. Prediction of credit risk can be performed through procedures of credit scoring. According to Hand and Henley (1997) : "Credit scoring is the term used to describe formal statistical methods used for classifying applicants for credit into 'good' and 'bad' risk classes".

As a result, the credit scoring systems are of great interest to banks and financial institutions, not only because they must measure credit risk, but because any small improvement would produce great profits (Hand & Henley, 1997), including cost reduction of credit analysis, delivery of faster decisions, guaranteed credit collection, and risk mitigation. For this task, a broad amount of methodologies have been developed (García, Marqués, & Sánchez, 2015), beginning with statistical techniques (Hand & Henley, 1997) and using mainly Artificial Intelligent methods nowadays (Lessmann, Baesens, Seow, & Thomas, 2015; Louzada, Ara, & Fernandes, 2016).

The classical statistical models assume a previous knowledge that it is not necessary when more modern artificial intelligence tools are applied. This last tools extract information directly from data without any previous conditions. In the last years, many studies have appeared showing that these techniques in the artificial intelligence, principally from the data mining area, present some improvements in the results obtained when they are compared with the ones obtained from classical statistical approaches.

Among artificial intelligence techniques, the most popular methods and the ones that show the best results are ensemble of classifiers (Ala'raj & Abbod, 2016; Hung & Chen, 2009; Marqués, García, & Sánchez, 2012; Nanni & Lumini, 2009; Sadatrasoul, Gholamian, Siami, & Hajimohammadi, 2013; Wang, Ma, Huang, & Xu, 2012; Xiao, Xiao, & Wang, 2016); Support Vector Machines (SVM) (Chen, Ma, & Ma, 2009; Harris, 2015; Hens & Tiwari, 2012; Huang, Chen, & Wang, 2007; Schebesch & Stecking, 2005; Tomczak & Zieba, 2015); and Artificial Neural Networks

* Corresponding author.
E-mail addresses: jabellan@decsai.ugr.es (J. Abellán), fjgc@decsai.ugr.es (J.G. Castellano).

http://dx.doi.org/10.1016/j.eswa.2016.12.020
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

(ANN) (Lee & Chen, 2005; West, 2000; Zhao et al., 2015). However the most outstanding are the ensembles (Marqués et al., 2012; Xiao et al., 2016). Another Artificial Intelligence methodologies that have been also used for this kind of research are decision trees (DT) (Bijak & Thomas, 2012; Makowski, 1985; Yap, Ong, & Husain, 2011); Bayesian networks (BN) (Zhu, Beling, & Overstreet, 2002; Wu, 2011); k-nearest neighbors (KNN) (Henley & Hand, 1996); and many others (Lessmann et al., 2015; Louzada et al., 2016).

Decision trees (DTs) (Quinlan, 1993) also known as classification trees or hierarchical classifiers are a fast type of classifiers with a simple structure which is easy to interpret. One important characteristic of this classifier is that few variations of the data, used to learn, produces important differences in the model, this is know as *instability* or *diversity* (Tsymbal, Pechenizkiy, & Cunningham, 2005). It is important to highlight that some schemes to create ensembles of classifier do not need to be based on very complex and accurate individual classifiers, such as ANNs or SVMs. For example, Bagging scheme (Breiman, 1996) is a well known procedure for creating ensembles of classifiers that performs best when the base classifier are not only accurate, but also unstable. Hence, DTs encourage diversity for the combination of classifiers (Breiman, 1996) and provide an excellent model for the Bagging ensemble scheme, being used advantageously for scoring problems (Abellán & Mantas, 2014; Marqués et al., 2012; Nanni & Lumini, 2009; Xiao et al., 2016). In this paper, it is shown that a single classifier is, normally, a good choice to be used in a ensemble scheme when it is accurate and presents high degree of diversity when the data vary.

On the other hand, until a few years ago, the classical theory of probability (PT) has been the fundamental tool to construct a method of classification. Many theories to represent the information have arisen as a generalization of the PT, such that: theory of evidence, measures of possibility, intervals of probability, capacities of 2-order, etc. Each one represent a model of imprecise probabilities, see Walley (1996).

The recent model of Credal Decision Tree (CDT) of Abellán and Moral (2003), uses imprecise probabilities and uncertainty measures (Klir, 2005) to build a decision tree. The CDT model represents an extension of the classical ID3 model of Quinlan (Quinlan, 1986), replacing precise probabilities and entropy with imprecise probabilities and maximum of entropy. This last measure is a well accepted measure of total uncertainty for some special type of imprecise probabilities (Abellán, Klir, & Moral, 2006).

The CDT model could be interpreted as a parametric extension of classical decision tree methods, but the use of the maximum entropy measure represents an important difference with the classical methods. The use of that measure implies a lower level of overfitting on the data used to learn. This characteristic makes the CDT model different than the classical ones.

In the paper of Marqués et al. (2012), a thorough study was performed on the use of different ensemble methods (AdaBoost, Bagging, Random Subspace, DECORATE and Rotation Forest) with the following base classifiers: 1-nearest neighbor (1-NN), naïve Bayes classifier (NBC), logistic regression (LogR), multilayer perceptron (MLP), radial basis function (RBF), Support Vector Machine (SVM) and C4.5 decision tree. Even though we do believe it is an excellent work, we think that there is some room for improvement and that is the aim of this proposal.

In this work we have made a comparison of the different ensemble procedures studied in Marqués et al. (2012). Here, we only have taken into account the base classifiers with the better results, and we have added the CDT procedure to the set of base classifiers. In the recent work of Abellán and Mantas (2014) it has been shown that the CDT model has a good performance when it is applied on credit scoring problems. This result has motivated us to analyze its behaviour when it is applied in a study as the one of Marqués et al. (2012).

In order to improve the contrast between the different procedures, the area under the ROC curve (AUC), that was not considered in Marqués et al. (2012), is now taken into account because it is more appropriate, that even the direct accuracy, for imbalanced data sets with different misclassification costs.[1] Although the AUC is less used than the type-I and type-II error rates in scoring problems (García et al., 2015), in the general machine learning literature it is acknowledged as one of the best measures for comparing classifiers in two-class problems (see Beck & Shultz, 1986; Fawcett, 2003).

Via an experimental study, it is shown that the CDT model presents a general better performance than the other methods analyzed here, when it is applied in different ensemble schemes on credit scoring. It obtains the best results in accuracy and AUC, showing that to be a good choice to use in a ensemble scheme, the individual accuracy is not the most important characteristic. We will also see that the C4.5 presents a good performance too in the aspects measured here. But the SVM method, that appears as the second best in the study of Marqués et al. (2012), suffers a worsening when it is compared here only with the best ones. Now, the LogR method appears as a very good alternative to be used in ensembles, with similar results to the ones obtained by the C4.5 method.

The rest of the paper is organized as follows. Section 2 begins with the necessary previous knowledge about the credal decision tree procedure: its differences with respect to the classic procedures, and its algorithm; following with a description of the ensemble schemes used. Section 3 describes and contains the experiments carried out. Section 4 comments the results obtained from the experiments. Finally, Section 5 is devoted to the conclusions.

2. Previous knowledge

2.1. Classic DTs vs DTs based on imprecise probabilities

Decision trees, or classification trees, are simple structures that can be used as classifiers. In situations where elements are described by one or more *attribute variables* (also called *predictive attributes* or *features*) and by a single *class variable*, which is the variable under study, classification trees can be used to predict the class value of an element by considering its attribute values. In such a structure, each non-leaf node represents an attribute variable, the edges or branches between that node and its child nodes represent the values of that attribute variable, and each leaf node normally specifies an exact value of the class variable.

The process for inferring a decision tree is mainly determined by the followings aspects:

(1) The *split criterion*, i.e. the method used to select the attribute to insert in a node and branching.
(2) The criterion to stop the branching.
(3) The method for assigning a class label or a probability distribution at the leaf nodes.

An optional final step in the procedure to build DTs, which is used to reduce the overfitting of the model to the training set, is:

(4) The post-pruning process used to simplify the tree structure.

In classic procedures for building DTs, where a measure of information based on PT is used, the criterion to stop the branching (above point (2)) is when the measure is not improved or when a threshold of gain in the measure of information is attained. With respect to the point (3), the value of the class variable inserted in

[1] The cost for a false positive in credit scoring is usually much more expensive than for a false negative (Caouette, Altman, Narayanan, & Nimmo, 2008).

Download English Version:

https://daneshyari.com/en/article/4943470

Download Persian Version:

https://daneshyari.com/article/4943470

Daneshyari.com