Expert Systems with Applications 41 (2014) 3825–3830

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring

Joaquín Abellán *, Carlos J. Mantas

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

ARTICLE INFO

Keywords:
Bankruptcy prediction
Credit scoring
Ensembles of classifiers
Decision trees
Imprecise Dirichlet model

ABSTRACT

Previous studies about ensembles of classifiers for bankruptcy prediction and credit scoring have been presented. In these studies, different ensemble schemes for complex classifiers were applied, and the best results were obtained using the Random Subspace method. The Bagging scheme was one of the ensemble methods used in the comparison. However, it was not correctly used. It is very important to use this ensemble scheme on weak and unstable classifiers for producing diversity in the combination. In order to improve the comparison, Bagging scheme on several decision trees models is applied to bankruptcy prediction and credit scoring. Decision trees encourage diversity for the combination of classifiers. Finally, an experimental study shows that Bagging scheme on decision trees present the best results for bankruptcy prediction and credit scoring.

1. Introduction

In supervised classification tasks, the combination or ensemble of classifiers represents an interesting way of merging information that can provide a better accuracy than each individual method. The high classification accuracy performance of these combined methods makes them very suitable for real world applications, such as bankruptcy prediction and credit scoring.

In the paper of Nanni and Lumini (2009) it is presented an interesting analysis about previous papers on bankruptcy prediction and credit scoring. The importance of this type of real application is well exposed: (a) The credit scoring models permit to discriminate between good credit group and bad credit group; (b) Developing a reliable credit scoring system offers several benefits, including cost reduction of credit analysis, delivery of faster decisions, guaranteed credit collection, and risk mitigation.

In this paper (Nanni & Lumini, 2009), the authors analyzed some well established financial decision-making methods based on machine learning to solve the financial decision-making problems mentioned above. In that work, the individual methods providing better performance were based on Artificial Neural Networks (ANNs). They also presented a thorough study about several techniques to create ensembles of classifiers based on some complex classifiers, including ANNs. All of them were applied to data sets related to the problem of bankruptcy prediction and credit scoring, with the aim of outperforming previous works, like in Tsai and Wu (2008).

It is important to highlight that some schemes to create classifier ensembles do not have to be based on very complex and accurate individual classifiers. For example, Bagging scheme (Breiman, 1996) is a well known procedure for creating ensembles of classifiers that performs best when applied to weak and unstable classifiers. However, this fact was forgotten in the previous studies about bankruptcy prediction and credit scoring.

Decision trees (DTs) represent a family of simple classifiers that can be built in very little time and have a simple structure which can be interpreted easily. An important aspect of DTs, which make them very suitable for ensembles of classifiers, is their instability: Different training sets from a given problem domain will produce very different models. Hence, DTs encourage diversity for the combination of classifiers (Breiman, 1996) and provide an excellent model for the Bagging ensemble scheme.

In Abellán and Masegosa (2012), it is shown that using Bagging ensembles on a special type of decision trees, called credal decision trees (CDTs) (Abellán & Moral, 2003b), provides an interesting tool for the classification task. CDTs are based on imprecise probabilities (more specifically, on the Imprecise Dirichlet Model (IDM); see Walley, 1996) and information/uncertainty measures (in particular, on the maximum of entropy function; see Klir, 2006, Abellán, 2011). An important characteristic of the CDT procedure is that the split criterion used to build a DT has a different treatment of the imprecision than the one used for the classic split criteria.

Hence, the main purpose of this paper is to complete previous works, especially the one presented in Nanni and Lumini (2009)

* Corresponding author. Tel.: +34 958 242376.
E-mail addresses: jabellan@decsai.ugr.es (J. Abellán), cmantas@decsai.ugr.es (C.J. Mantas).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.003

about the use of Bagging ensembles on DTs. We show that the use of CDTs in a Bagging scheme outperforms previous results for data sets related to bankruptcy prediction and credit scoring. We have used the same setting employed in Nanni and Lumini (2009): Same data sets, same type of experimentation and same measure to compare results. Moreover, we have used known statistical tests to support our results.

In order to compare the performance of the mentioned procedures in a logical way, we have used the best ensemble method described in previous works (the Random Subspace ensemble procedure (Ho, 1998)) and the Bagging scheme on DTs. In addition, the trees were built with the most successful classification method based on DTs: Quinlan's C4.5 algorithm (Quinlan, 1993); and the mentioned CDT procedure (Abellán & Moral, 2003b).

This paper is organized as follows: In Section 2, we present the necessary background about DTs, CDTs, ensemble methods and previous works made on data sets concerning bankruptcy prediction and credit scoring; in Section 3, we describe and comment on the results of the experiments carried out; and finally, Section 4 presents the conclusions.

2. Background

2.1. Decision trees

The classification task is focused on elements that are described by one or more characteristics, known as *attribute variables* (also called *predictive attributes* or *features*), and by a single *class variable*, with the aim to predict the class value of a new element by considering its attribute values.

Decision trees (also known as Classification Trees or hierarchical classifiers) started to play an important role in machine learning since the publication of Quinlan's *Iterative Dichotomiser 3*, known as ID3 (Quinlan, 1986). Subsequently, Quinlan also presented the *Classifier 4.5*, known as C4.5 (Quinlan, 1993), which is an advanced version of the ID3. Since then, C4.5 has been considered as a standard model in supervised classification. They have also been widely applied as a data analysis tool to very different fields, such as astronomy, biology, medicine, etc.

Decision trees are models based on a recursive partitioning method that divides the data set using a single variable at each level. This variable is selected by means of a given criterion. Ideally, these models define sets of cases in which all the cases in a set belong to the same class.

Their knowledge representation has a simple tree structure. It can be interpreted as a compact rule set in which each node of the tree is labeled with an attribute variable that produces a different branch for each variable value (i.e., a partition of the data set). Leaf nodes are labeled with a class label.

The process for inferring a decision tree is mainly determined by the following points: (i) The criteria used to select the attribute that should be placed in a node and branched; (ii) The criteria for stopping the tree branching process; (iii) The method for assigning a class label or a probability distribution to the leaf nodes; and (iv) The posterior pruning process used to simplify the tree structure.

Many different approaches for inferring decision trees, which depend upon the aforementioned factors, have been published. Quinlan's ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993) stand out among the most popular ones.

Decision trees are built using a data set referred to as the training data set. A different set, called the test data set, is used to check the model. When we get a new sample or instance of the test data set, we can make a decision or prediction about the state of its class variable, following the path in the tree from the root to a leaf node using the sample values and the tree structure.

2.2. Credal decision trees

2.2.1. Mathematical foundations

The split criterion employed to build CDTs (Abellán & Moral, 2003b) is based on the application of uncertainty measures on convex sets of probability distributions (credal sets). Specifically, probability intervals are extracted from the data set for each case of the class variable using Walley's Imprecise Dirichlet Model (IDM) (Walley, 1996), which represents a specific kind of convex set of probability distributions (see Abellán, 2006a).

The IDM depends on a given hyperparameter s which does not depend on the sample space (Walley, 1996). The IDM estimates that the probabilities for each value of the class variable C are within an interval defined by:

$$p(c_j) \in \left[\frac{n_{c_j}}{N+s}, \frac{n_{c_j}+s}{N+s}\right], \quad j = 1, \ldots, k;$$

with n_{c_j} as the frequency of the set of values $(C = c_j)$ in the data set, N the sample size and k the number of cases in class variable C. One important thing is that intervals are wider if the sample size is smaller. So this method produces more precise intervals as N increases.

Walley (1996) does not give a definitive recommendation for the value of the parameter s, but he suggests two candidates: $s = 1$ or $s = 2$. In our case, we will use a value $s = 1$. The reason for this is the low computational cost of the inference with credal sets for $s = 1$, as it will be shown in the following paragraph.

The entropy of this set of probability intervals will be estimated as the maximum of the entropy of all probability distributions $(q(c_1), \ldots, q(c_k))$ verifying that, for any c_i, $q(c_i)$ belongs to the estimated interval for $p(c_i)$. For $s = 1$ this entropy is very simple to compute. First, we have to determine $A = \{c_j : n_{c_j} = \min_i\{n_{c_i}\}\}$. If l is the number of elements of A, then the distribution with maximum entropy is p^*, where $p^*(c_i) = \frac{n_{c_i}}{N+s}$ if $c_i \notin A$ and $p^*(c_i) = \frac{n_{c_i}+s/l}{N+s}$ if $c_i \in A$.

This upper entropy function, denoted as S^*, is a total uncertainty measure which is well known for this type of set (see Abellán & Moral, 2003a; Abellán, Klir, & Moral, 2006; Abellán & Masegosa, 2008).

As the intervals are wider with smaller sample sizes, there will be a tendency to get greater maximum entropy values with smaller sample sizes. This property will be important to differentiate the action of the CDTs from the behavior of other kinds of DTs.

For example, if we assume $k = 2$, then the information obtained from a node A with $n_{c_1} = 400$ and $n_{c_2} = 100$ will be more reliable than that provided by a node B with $n_{c_1} = 4$ and $n_{c_2} = 1$, since the first node has a larger sample size. This fact is taken into account by the maximum entropy function (the first node will have the lowest uncertainty); in contrast, using classic entropy both nodes will have the same uncertainty.

Using S to denote the classic entropy function on a probability distribution p, and K^A, K^B to denote the sets of probabilities via the IDM (with $s = 1$) associated to nodes A and B mentioned above, we have that:

$$S\left(\frac{400}{500}, \frac{100}{500}\right) = S\left(\frac{4}{5}, \frac{1}{5}\right) = 0.5004,$$

$$S^*(K^A) = 0.5025; \quad S^*(K^B) = 0.6365$$

Hence

$$S\left(\frac{400}{500}, \frac{100}{500}\right) = S\left(\frac{4}{5}, \frac{1}{5}\right) \cong S^*(K^A) < S^*(K^B),$$

This shows that the treatment of imprecision is clearly different with the new split criterion based on imprecise probabilities of