Expert Systems With Applications 97 (2018) 228–243

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Increasing diversity in random forest learning algorithm via imprecise probabilities

Joaquín Abellán*, Carlos J. Mantas, Javier G. Castellano, Serafín Moral-García

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

ARTICLE INFO

Article history:
Received 7 September 2017
Revised 29 November 2017
Accepted 15 December 2017
Available online 15 December 2017

Keywords:
Classification
Ensemble schemes
Random forest
Imprecise probabilities
Uncertainty measures

ABSTRACT

Random Forest (RF) learning algorithm is considered a classifier of reference due its excellent performance. Its success is based on the diversity of rules generated from decision trees that are built via a procedure that randomizes instances and features. To find additional procedures for increasing the diversity of the trees is an interesting task. It has been considered a new split criterion, based on imprecise probabilities and general uncertainty measures, that has a clear dependence of a parameter and has shown to be more successful than the classic ones. Using that criterion in RF scheme, join with a random procedure to select the value of that parameter, the diversity of the trees in the forest and the performance are increased. This fact gives rise to a new classification algorithm, called Random Credal Random Forest (RCRF). The new method represents several improvements with respect to the classic RF: the use of a more successful split criterion which is more robust to noise than the classic ones; and an increasing of the randomness which facilitates the diversity of the rules obtained. In an experimental study, it is shown that this new algorithm is a clear enhancement of RF, especially when it applied on data sets with class noise, where the standard RF has a notable deterioration. The problem of overfitting that appears when RF classifies data sets with class noise is solved with RCRF. This new algorithm can be considered as a powerful alternative to be used on data with or without class noise.

1. Introduction

The classification task (Hand, 1997), in the data mining area, starts from a set of data about observations or cases described via *attributes* or *features*; where each observation has an assigned value (label) of a variable under study, also called *class variable*. The final aim of this task is to extract knowledge from data to predict the value of the label of the class variable when a new observation appears. In order to build a classifier from a data set, different approaches can be used, such as classical statistical methods (Hand, 1981), decision trees (Quinlan, 1993), artificial neural networks or Bayesian networks (Pearl, 1988).

Decision trees (DTs) also known as classification trees are a type of classifiers with a simple structure where the knowledge representation is relatively simple to interpret and it can be seen as a set of decision rules in a tree format. DTs began to increase their importance with the publication of the ID3 algorithm proposed by Quinlan (1986). Afterwards Quinlan proposed the C4.5 (Quinlan, 1993) algorithm, which is an improvement of the previous ID3 and obtains better results. One important characteristic of the standard procedures to build DTs is that few variations of the data, used to learn, produces important differences in the models. This is known as *instability* or *diversity* (Tsymbal, Pechenizkiy, & Cunningham, 2005) of decision tree classifiers, where the constructed rules may be significantly different from the original ones if the input training sample is slightly changed. That is, the rules generated from two similar samples may be very different.

The fusion of information obtained via ensembles or combination of several classifiers can improve the final process of a classification task, this can be represented via an improvement in terms of accuracy and robustness. Some of the more popular schemes are bagging (Breiman, 1996), boosting (Freund & Schapire, 1996) or Random Forest (Breiman, 2001). The inherent instability of decision trees (Breiman, 1996) makes these classifiers very suitable to be employed in ensembles. In a ensemble scheme, there is little gain combining similar classifiers, so the improvement of the ensemble relies on the diversity of the base classifiers, provided that this diversity does not diminish the accuracy of the ensemble members. A revision of ensemble methods and diversity can be found in Dietterich (2000a), Brown, Wyatt, Harris, and Yao (2005), and Ren, Zhang, and Suganthan (2016).

* Corresponding author.
E-mail addresses: jabellan@decsai.ugr.es (J. Abellán), cmantas@decsai.ugr.es (C.J. Mantas), fjgc@decsai.ugr.es (J.G. Castellano), seramoral@decsai.ugr.es (S. Moral-García).

https://doi.org/10.1016/j.eswa.2017.12.029
0957-4174/

Random Forest (RF) is a fine supervised classification method based on the combination of the Breiman's "bagging" and random selection of features (Breiman, 2001) in order to construct a collection of decision trees with controlled variance. Advanced classification models based on RF have been recently published (Menze, Kelm, Splitthoff, Koethe, & Hamprecht, 2011; Zhang & Suganthan, 2014; 2015; 2017). In the original algorithm of RF, the decision trees are built without pruning. In this way, a tree tends to be more different from the rest than the pruned version of the tree. Besides, RF algorithm has two stochastic elements: (a) Bagging employed for the selection of the instances used as input for each tree; and (b) the random set of features considered as candidates for splitting each node. These stochastic aspects increase the diversity of the trees and significantly improve the overall predictive accuracy of RF when the outputs of these trees are combined. It could be interesting to find other concepts for increasing the trees diversity in RF, without giving up the accuracy of the ensemble members. These new concepts can be found in the new theories of imprecise probabilities.

The good results obtained by the RF classifier in several areas have motivated that RF is one of the most used models in the literature of applications in the data mining area. Some very recent references about its use, combined with other models as Neural Networks (NNs), are the following ones: combinations between NNs and RF in Bai (2017), Azqhandi, Ghaedi, Yousefi, and Jamshidi (2017), and Wang et al. (2015); ensembles of NNs, RF and other models in Krauss, Do, and Huck (2017); and different applications in big data about crash risk analysis, visual classification and other ones in Gauba et al. (2017), Jiang, Abdel-Aty, Hu, and Lee (2016), and Li et al. (2016).

The classical theory of probability has been the principal tool to construct learning procedures in the data mining area. But, few years ago, generalizations of this theory have arisen, such as Klir (2005): theory of evidence, measures of possibility, intervals of probability, capacities of 2-order, etc. Each one represents a model based on imprecise probabilities (see Walley, 1996).

The Credal Decision Tree model[1] (CDT) of Abellán and Moral (2003), uses imprecise probabilities and general uncertainty measures (Klir, 2005) to build a decision tree. The CDT model represents an extension of the classical ID3 model of 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). In the last years, it has been shown that the CDT model presents good experimental results in standard classification tasks (see Abellán & Masegosa, 2009; Abellán & Moral, 2005). The treatment of the imprecision is different when imprecise probabilities are used. This fact has been experimentally shown in Abellán and Masegosa (2012), Mantas and Abellán (2014a), Abellán and Mantas (2014), and Mantas and Abellán (2014b), where the models are applied on data set with label noise, i.e. data sets where the class variable has some incorrect labels, due principally to deficiencies in the data learning and/or the process for capture of data.[2]

The performance of CDTs depends of a hyperparameter s used in its split criterion (Abellán, 2006). The adjustment of this hyperparameter is necessary in terms of the noise level of the data set to be classified (see Mantas, Abellán, & Castellano, 2016). Different values of s produce different CDTs when they are constructed to classify the same data set. In this way, diversity of CDTs without giving up accuracy can be obtained by changing the value of this parameter s when a data set is classified. Besides, as it can be read in Mantas et al. (2016), the controlled modification of the value for s do not diminish the accuracy of the decision tree drastically.

The diversity of trees in the forest created by the RF algorithm is achieved by using trees without pruning, bagging and random selection of features. If we use the split criterion of the CDT procedure in the base tree of the RF algorithm, a new element for increasing the diversity of the trees in the forest can be inserted. For each new DT in RF, a random selection for the value of s can be carried out. Thus, an increase of diversity in the trees of the forest with acceptable accuracy is obtained and this fact is important for improving the predictive accuracy of RF.

The method of the RF algorithm where the forest is built with DTs using the split criterion of the CDT and the value of the parameter s is randomly selected, will be named as *Random Credal Random Forest* (RCRF). It has been designed and implemented in this paper. Finally, an exhaustive experimental comparison has been carried out, in order to compare RCRF and other ensemble methods as the original RF algorithm and other, successful under class noise, bagging schemes. This experimental study is presented in this work in order to show that RCRF algorithm obtains better classification results than the original RF algorithm and the rest of ensemble methods. In particular, RCRF algorithm correctly classifies data sets with or without noise. This is an important improvement of the standard RF algorithm because this algorithm suffers the overfitting problem when noisy data sets are classified.

The rest of the paper is organized as follows. Section 2 presents the necessary previous knowledge about the new split criterion used and the Random Forest algorithm. Section 3 describes the RCRF algorithm and its base classifier. Section 4 justifies the definition of the new ensemble method RCRF. Section 5 describes the experimentation carried out. Section 6 comments the results of the experimentation. Finally, Section 7 is devoted to the conclusions.

2. Previous knowledge

2.1. Credal decision tree procedure

The known recursive process to build a decision tree is normally based on the followings points: (i) the use of a split criteria to select the feature to be insert in a node and branching; (ii) a criteria to stop the tree from branching; and (iii) a method for assigning a class label (or a probability distribution) at the leaf nodes. Alternatively, can be also used (iv) a post-pruning process used to simplify the tree structure.

Many different approaches for inferring decision trees, which depend upon the aforementioned points, have been published. Quinlan's ID3 and C4.5 (Quinlan, 1993) stand out among all of these. The split criteria used by these algorithms are *Info-Gain* (IG) for ID3 and *Info-Gain Ratio* (IGR) for C4.5. Both procedures have been extensively used in the literature of the area of data mining.

The use of different split criteria normally implies different graphical structures of the trees. Hence, it can be considered as the most important part of the algorithm to build a DT. The split criterion employed to build Credal Decision Trees (CDTs) Abellán and Moral (2003), is different to the classic criteria and it is based on imprecise probabilities and the application of uncertainty measures on credal sets.

2.1.1. Split criterion

The classical criteria use normally, as base measure of information, the Shannon's entropy measure; and the one that we use here, based on imprecise probabilities, uses the maximum entropy measure. The maximum entropy measure verifies an important set of properties on theories based on imprecise probabilities that are generalizations of the probability theory (see Klir, 2005). Here, we

[1] The term *credal* comes from the use of a special type of imprecise probabilities: closed and convex set of probability distributions

[2] A complete and recent revision of machine learning methods to manipulate label noise can be found in Frenay and Verleysen (2014).