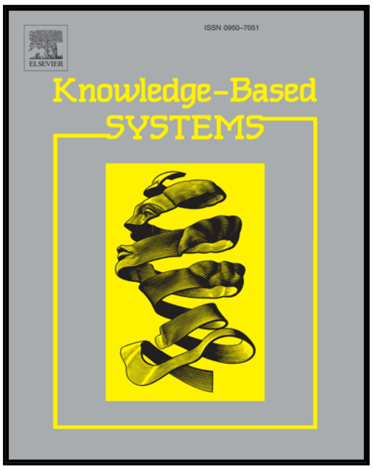

A Random Forest approach using imprecise probabilities

Joaquín Abellán, Carlos J. Mantas, Javier G. Castellano

PII: S0950-7051(17)30341-6
DOI: 10.1016/j.knosys.2017.07.019
Reference: KNOSYS 3983

To appear in: *Knowledge-Based Systems*

Received date: 9 April 2017
Revised date: 20 June 2017
Accepted date: 17 July 2017

Please cite this article as: Joaquín Abellán, Carlos J. Mantas, Javier G. Castellano, A Random Forest approach using imprecise probabilities, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.07.019

A Random Forest approach using imprecise probabilities

Joaquín Abellán, Carlos J. Mantas and Javier G. Castellano

Department of Computer Science and
Artificial Intelligence
University of Granada, Granada, Spain
{jabellan,cmantas,fjgc}@decsai.ugr.es

Abstract. The Random Forest classifier has been considered as an important reference in the data mining area. The building procedure of its base classifier (a decision tree) is principally based on a randomization process of data and features; and on a split criterion, which uses classic precise probabilities, to quantify the gain of information. One drawback found on this classifier is that it has a bad performance when it is applied on data sets with class noise. Very recently, it is proved that a new criterion which uses imprecise probabilities and general uncertainty measures, can improve the performance of the classic split criteria. In this work, the base classifier of the Random Forest is modified using that new criterion, producing also a new single decision tree model. This model join with the randomization process of features is the base classifier of a new procedure similar to the Random Forest, called Credal Random Forest. The principal differences between those two models are presented. In an experimental study, it is shown that the new method represents an improvement of the Random Forest when both are applied on data sets without class noise. But this improvement is notably greater when they are applied on data sets with class noise.

Keywords: Classification; class noise; Random Forest; imprecise probabilities; uncertainty measures

1 Introduction

The task of supervised classification [1] 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 [2], decision trees [3], artificial neural networks or Bayesian networks [4].

Decision trees (DTs) also known as classification trees are a type of classifiers with a simple structure where the knowledge representation is relatively simple