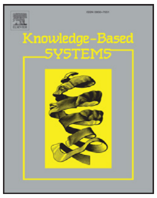
Knowledge-Based Systems 134 (2017) 72–84

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



A Random Forest approach using imprecise probabilities

Joaquín Abellán[*], Carlos J. Mantas, Javier G. Castellano

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

ARTICLE INFO

Article history:
Received 9 April 2017
Revised 20 June 2017
Accepted 17 July 2017
Available online 20 July 2017

Keywords:
Classification
Class noise
Random Forest
Imprecise probabilities
Uncertainty measures

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.

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 to interpret. DTs began to increase their importance with the publication of the ID3 algorithm proposed by Quinlan [5]. Afterwards Quinlan proposed the C4.5 [3] algorithm, which is an improvement of the previous ID3 and obtains better results. One important characteristic of the DT is that few variations of the data, used to learn, produce important differences in the model. This is known as the *instability* or *diversity* [6] of a decision tree classifier, 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 most popular schemes are bagging [7], boosting [8] and Random Forest [9]. The inherent instability of decision trees [7] makes these classifiers very suitable to be employed in ensembles. In an ensemble scheme, there is few 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.

Random Forest (RF) is a fine supervised classification method based on the combination of the Breiman's "bagging" and random selection of features [9] in order to construct a collection of decision trees with controlled variance. The "No Free Lunch" theorem applied to supervised classification [10] states that there is not a general-purpose supervised learning algorithm that always improves the rest of the methodologies, i. e. there is always a problem where a classifier is outperformed by another. Nevertheless, a thorough study comparing an extensive number of supervised classification algorithms against a large number of benchmark data sets was performed in [11], and the authors asseverate that "The classifiers most likely to be the best are the Random Forest versions".

[*] 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).

http://dx.doi.org/10.1016/j.knosys.2017.07.019
0950-7051/© 2017 Elsevier B.V. All rights reserved.

Class noise, also known as *label noise* or *classification noise*, is named to those situations which appears when data sets have incorrect class labels. This situation is principally motivated by deficiencies in the data learning and/or the process for capture of data. One of the most important procedures to have success in a classification task in situations of noisy domains, is the use or application of ensembles of classifiers. In the literature about classification on noisy domains, bagging scheme stands out as the most successful scheme. However, it has been observed that RF algorithm presents the problem of overfitting when it classifies noisy data sets. In [12,13] we can find experimental studies where RF has a bad performance under class noise when it is compared with bagging schemes of diverse decision tree models. In this paper, we present a modification of RF algorithm where the problem of overfitting on noisy data sets is solved. A complete and recent revision of machine learning methods to manipulate label noise can be found in [14].

The classical theory of probability has been the principal tool to construct learning procedures in the data mining area. Few years ago, generalizations of this theory have arisen, such as [15]: theory of evidence, measures of possibility, intervals of probability, capacities of 2-order, etc. Each one represents a model of imprecise probabilities (see [16]).

The Credal Decision Tree model (CDT) of Abellán and Moral [17], uses imprecise probabilities and general uncertainty measures [15] to build a decision tree. The CDT model represents an extension of the classical ID3 model of Quinlan [5], 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 [18]. In the last years, it has been shown that the CDT model presents good experimental results in standard classification tasks (see [19,20]). The use of a bagging ensemble with the CDT as base classifier was exposed in [21], and its application on data sets with class noise in [12,13].

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 (high instability). This process is good for reducing variance when the output of the trees is aggregated in the RF. However, when the trees are unpruned, they have the risk of excessive fitting on the data used (overfitting), mainly on noisy data sets. Hence, it could be interesting to build trees in RF with an intermediate size between unpruned trees and pruned trees. In previous studies about credal trees [21,22], it is shown that these trees have an intermediate size in comparison with the size of the standard classification trees with or without pruning. In this way, an improvement of the original RF classifier can be obtained with the use of imprecise probabilities in its algorithm. A trade-off between correlation and overfitting of the trees in the forest can be achieved.

First, in this work, it is analyzed the principal characteristics of the use of imprecise probabilities in a split criterion to build decision trees. It is shown that: (i) the inclusion of imprecise probabilities can stop the branching in a DT before than precise probabilities do, avoiding excessive overfitting; (ii) the different treatment of the imprecision makes the trees, built with imprecise probabilities, more robust on data sets with label noise; (iii) the use of the maximum entropy function on a credal set for calculating the split criterion encourages the diversity of the trees in a forest.

The decision tree model which uses the above mentioned new split criterion and the same randomization process of features than the one used in the base classifier of the RF procedure, that we have called *Credal Random Tree* (CRT), has been compared with the DT used in the original method of the RF. An experimentation has been carried out to compare them with other single similar DT classifiers: the C4.5 procedure and the CDT. We have analyzed the results considering accuracy and the average number of nodes built with each model, with the aim to discover the principal differences among the models when they are applied on data set with label noise. We find that the level of overfitting and bad performance are related when the level of noise increases.

In previous studies [12,13], the CDT model presented the best single performance under label noise, and the best general performance for a classifier when it is used in a bagging scheme on noise data sets. However, now the Credal Random Tree, when it is used in the RF scheme, can even improve the results of a bagging CDT on data sets with different levels of label noise. This is our new final proposal as classifier in this work, that we have named as *Credal Random Forest* (CRT). It will be seen that the CRF corrects the problems of the RF in situations with label noise. The improvement of the new model with respect to the classic RF is clear also when no label noise is added.

An experimental comparison on a large set of data sets has been also carried out in order to compare the new proposal (CRF), the original RF and bagging schemes with decision tree models as base classifier: C4.5 and CDT. A set of tests and measures have been used in order to study the different performance of the models. From this experimentation we can see that the CRF model improves the rest of models when data sets with or without noise are classified.

It must be remarked, as a consequence of this work, that the use of the following elements: (i) a randomization of features and data, and (ii) the application of imprecise probabilities and general uncertainty measures; is a winner combination as it will be seen with the new Credal Random Forest. Those elements represents an excellent trade-off between accuracy and instability for a base classifier, with the aim to be used in an ensemble scheme. The new model presented here, improves the performance of the RF on all situations of possible label noise.

The rest of the paper is organized as follows. In Section 2 it is presented all the necessary previous knowledge about the decision tree models used and the Random Forest algorithm. Sections 3 describes the CRF algorithm and some of its principal differences with respect to the standard RF. Section 4 describes the experimentation carried out. Section 5 comments the results of the experimentation. Finally, Section 6 is devoted to conclusions.

2. Previous knowledge

2.1. Decision trees

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 in the classical theory of probability 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 that measure is attained. With respect to the (3) point, the value of the class variable inserted in a leaf node is the one with more frequency in the partition of the data associated with that leaf node; also can be inserted the distribution of probabilities associated with that partition set. Then the principal difference among all the procedures to build DTs is the point (1), i.e. the split criterion used to select the attribute variable to insert in a node.

Considering classic split criteria and split criteria based on imprecise probabilities, a basic point to differentiate them is how they obtain probabilities from data.

In the following subsection, the classic *Info-Gain* of Quinlan [5] based on precise probabilities will be compared with the *Imprecise Info-Gain* of Abellán and Moral [17], based on imprecise probabilities. It will be seen that the classical criteria use normally, as base measure of information, the Shannon's measure; and the ones based on imprecise probabilities use the maximum entropy measure. This measure is based on the *principle of maximum uncertainty* [15], widely used in classic information theory, where it is known as *maximum entropy principle* [23]. The maximum entropy measure verifies an important set of properties on theories based on imprecise probabilities that are generalizations of the probability theory (see [15]).

2.2. Info-Gain vs. Imprecise Info-Gain

Following the above notation, let X be a general feature with values belong to $\{x_1, \ldots, x_t\}$. The Info-Gain (IG) criterion was introduced by Quinlan as the basis for his ID3 model [5], and it is explained as follows:

- The entropy of C for the data set $\mathcal{D}$ is the Shannon's entropy [24] and it is defined as:

  $$H^{\mathcal{D}}(C) = \sum_i p(c_i) log_2(1/p(c_i)) \qquad (1)$$

  where $p(c_i)$ represents the probability of the class i in $\mathcal{D}$.
- The average entropy generated by the attribute X is:

  $$H^{\mathcal{D}}(C|X) = \sum_i P^{\mathcal{D}}(X = x_i) H^{\mathcal{D}_i}(C|X = x_i) \qquad (2)$$

  where $P^{\mathcal{D}}(X = x_i)$ represents the probability that $X = x_i$ in $\mathcal{D}$. $\mathcal{D}_i$ is the subset of $\mathcal{D}$ ($\mathcal{D}_i \subset \mathcal{D}$) where $(X = x_i)$.

Finally we can define the *Info-Gain* as follows:

$$IG(C,X)^{\mathcal{D}} = H^{\mathcal{D}}(C) - H^{\mathcal{D}}(C|X) \qquad (3)$$

The Imprecise Info-Gain (IIG) [17] is based on imprecise probabilities and the utilization of uncertainty measures on credal sets.[1] It was introduced to build the called *Credal Decision Tree* model (CDT). Probability intervals are obtained from the data set using the Walley's Imprecise Dirichlet Model (IDM) [16] (a special type of credal sets [25]), which arise from an optimization process of the Dirichlet's distribution.

With the above notation, $p(c_j), j = 1, .., k$ defined for each value c_j of the variable C, is obtained via the IDM on the following way:

$$p(c_j) \in \left[\frac{n_{c_j}}{N+s}, \frac{n_{c_j}+s}{N+s}\right], \quad j = 1, .., k; \qquad (4)$$

with n_{c_j} as the frequency of the set of values $(C = c_j)$ in the data set, N the sample size and s a given hyperparameter. The value of parameter s regulates the convergence speed of the upper and lower probability when the sample size increases. Higher values of s produce an additional cautious inference. Walley [16] does not give a decisive recommendation for the value of the parameter s, but he proposes two candidates: $s = 1$ or $s = 2$, nevertheless he recommend the value $s = 1$. It is easy to check that the size of the intervals increases when the value of s increases.

Table 1
Variable selected for branching (X_{sel}) by each split criterion.

IG	$X_{sel}: \ Arg - \min_X \left\{H^{\mathcal{D}}(C|X)\right\}$
IIG	$X_{sel}: \ Arg - \min_X \left\{H^*(K^{\mathcal{D}}(C|X))\right\}$

This representation gives rise to a specific kind of credal set on the variable C, $K^{\mathcal{D}}(C)$ [25]. The set is defined as

$$K^{\mathcal{D}}(C) = \left\{p \,|\, p(c_j) \in \left[\frac{n_{c_j}}{N+s}, \frac{n_{c_j}+s}{N+s}\right], \quad j = 1, .., k\right\}. \qquad (5)$$

On this type of sets (really credal sets, [25]), uncertainty measures can be applied. The procedure to build CDTs uses the maximum of entropy function on the above defined credal set.[2] This function, denoted as H^*, is defined as follows:

$$H^*(K^{\mathcal{D}}(C)) = max\left\{H^{\mathcal{D}}(p) \,|\, p \in K^{\mathcal{D}}(C)\right\} \qquad (6)$$

The procedure to obtain H^* for the special case of the IDM reaches the lowest computational cost for $s \le 1$ (see [25] for more details). Moreover, to compute H^* is very simple for $s = 1$ (this value for s will be used in the experimentation section). Firstly, the procedure consists in determining the set

$$A = \{c_j \mid n_{c_j} = \min_i\{n_{c_i}\}\} \qquad (7)$$

then the distribution with maximum entropy is

$$p^*(c_i) = \begin{cases} \frac{n_{c_i}}{N+s} & \text{if } c_i \notin A \\ \frac{n_{c_i}+s/|A|}{N+s} & \text{if } c_i \in A \end{cases}, \quad i = 1, .., k$$

The scheme to induce CDTs is like the one used by the classical ID3 algorithm [5], replacing its *Info-Gain* Split criterion with the *Imprecise Info-Gain* (IIG) split criterion which can be defined by the following way:

$$IIG^{\mathcal{D}}(C,X) = H^*(K^{\mathcal{D}}(C)) - \sum_i P^{\mathcal{D}}(X = x_i) H^*(K^{\mathcal{D}_i}(C|X = x_i)), \qquad (8)$$

where $P^{\mathcal{D}}(X = x_i)$ represents the probability that $X = x_i$ in $\mathcal{D}$, $K^{\mathcal{D}}(C)$ and $K^{\mathcal{D}_i}(C|X = x_i)$ are the credal sets obtained via the IDM for the C and $(C|X = x_i)$ variables respectively, for $\mathcal{D}$ and $\mathcal{D}_i$ partitions of the data set (see [17]). $\mathcal{D}_i$ is the subset of $\mathcal{D}$ ($\mathcal{D}_i \subset \mathcal{D}$) where $(X = x_i)$.

It should be taken into account that for a variable X and a data set $\mathcal{D}$, $IIG^{\mathcal{D}}(C,X)$ can be negative. This situation does not occur with the Info-Gain criterion. This important characteristic allows that the IIG criterion discards variables that worsen the information on the class variable. This is an important characteristic of the model that can be considered as an additional criterion to stop the branching of the tree.

For IG and IIG the first part of the criterion is a constant value for each attribute variable. Hence, both criteria select the variable with lowest value of uncertainty (greatest gain of information) expressed by the second parts in Eqs. (3) and (8), respectively. This situation can be seen as a scheme in Table 1.

[1] Closed and convex sets of probability distributions.

[2] This measure verifies a set of important properties on credal sets [18]. In the last years other measures on imprecise probability theories have arisen [26–28], but unfortunately they do not verify all the needed properties [29,30]

2.3. Random Forest algorithm

In the previous sections, the principal steps to build a DT and the split criterion used by classical DTs have been exposed. Now, the RF algorithm will be explained. This algorithm builds a forest of DTs. If a new instance must be classified, the features of this instance are presented to each DT in the forest. Each DT returns a classification value, a vote for that class. Finally, the classification value given by the RF is the one associated with the most voted state of the class variable, over all the DTs in the forest.

Each DT is built with the following characteristics:

1. If N is the number of instances in a data set, then RF selects a random sample with replacement of N instances from the original data. This sample will be the training set for building the tree.
2. If M is the number of features in a data set then a number $m < < M$ is specified. This value of m is held constant during the forest building.
3. At each node of the tree,
   3.1. m features are selected at random out of the M original features.
   3.2. The split criterion is calculated on these m features. The feature with the best value is used to split the node.
4. There is no pruning after building each decision tree.

In the original work where RF was presented ([9]), the author proposed two approaches of RF:

- Random Forest Using Random Input Selection (Random Forest-RI): It is the most common. In this approach m variables are selected at random out of the available attributes and the best split on these m is used. The number of attributes used in random selection by the author were 1 and the first integer less than $log_2(M) + 1$.
- Random Forest Using Linear Combination of Inputs (Random Forest-RC): Before the selection of the best variable to split, more attributes are created by taking random linear combinations of L variables (using $L = 3$).

The chosen approach is the RF with random input selection, which is the most common.[3] The number of features chosen for the split is the first integer less than $log_2(M) + 1$, where M is the number of input variables (features). We do not find any meaning in the use of only 1 variable, due to the use of only one random attribute is just random splitting which decreases the decision tree accuracy [31].

The original split criterion used by RF was the Gini Index, also based on classical probabilities, which was used by the CART[4] algorithm [32]. In this work the Information Gain criterion is used due to the fact that *Weka* software [33] has been used for the experimentation and this software utilizes the Info-Gain criterion in the RF implementation. Nonetheless, the Gini Index and the Info-Gain measure disagree only in 2% of all cases [34], which explains why empirical works (see [34,35]) concluded that there is not significant variation in accuracy, i.e. it is not feasible to determine which one of the two split criterion performs better.

3. Credal Random Forest algorithm and properties

3.1. Algorithm

The *Credal Random Forest* (CRF) algorithm is similar to the original RF algorithm. The main change is that CRF utilizes the Imprecise Info-Gain measure to split instead of using Info-Gain or Gini Index. This implies also an important difference expressed in step 3.3 of Fig. 1, where it is described the characteristics of the credal trees designed by CRF. This distinction provides new properties to the CRF algorithm that will be analyzed in the next subsections.

3.2. Size of the credal trees

The use of the Imprecise Info-Gain in the CRF procedure provides a trade-off between correlation of trees in the forest and overfitting of the constructed models. The decision trees designed by the classic RF algorithm are unpruned in order to achieve a low correlation among them. However, this fact implies that these trees have the risk of overfitting. Unpruned trees build with the IIG criterion have an intermediate size between pruned and unpruned decision trees. Hence, the trees built with the IIG criterion have a lower risk of overfitting than the unpruned decision trees built with the IG criterion. But, also the first ones have higher risk of correlation than the second ones.

To understand how the trees built with the IIG criterion can be smaller than the ones built with the IG criterion, we show the following example where the use of the IG implies a branching and the IIG implies a leaf node. This is motivated by an important property of the IIG criterion: it can be negative; whereas IG always gives us a positive value.

Example 1. Let C be a class variable with two possible states $\{c_1, c_2\}$. We consider that in a node J, for a DT, we have the following frequencies $\{c_1: 9, c_2: 4\}$. In this node, we also consider that we have only 2 attribute variables X_1, X_2, with possible values $X_1 \in \{x_1^1, x_2^1\}$, and $X_2 \in \{x_1^2, x_2^2, x_3^2\}$. The frequencies of each combination of states in the node J are the following ones:

$$\begin{array}{ll} X_1 = x_1^1 \rightarrow & (5 \text{ of class } c_1, 3 \text{ of class } c_2) \\ X_1 = x_2^1 \rightarrow & (4 \text{ of class } c_1, 1 \text{ of class } c_2) \\ X_2 = x_1^2 \rightarrow & (2 \text{ of class } c_1, 2 \text{ of class } c_2) \\ X_2 = x_2^2 \rightarrow & (5 \text{ of class } c_1, 2 \text{ of class } c_2) \\ X_2 = x_3^2 \rightarrow & (2 \text{ of class } c_1, 0 \text{ of class } c_2) \end{array}$$

Considering the IG criterion, we always have an improvement in the gain of information. The values obtained with this criterion are the following ones:

$$IG(C, X_1) = 0.8905 - \frac{8}{13}0.9544 - \frac{5}{13}0.7219 = 0.0255$$

$$IG(C, X_2) = 0.8905 - \frac{4}{13}1.0 - \frac{7}{13}0.8631 - \frac{2}{13}0.0 = 0.1180$$

Then the feature X_2 is inserted in the node J, because it produces the greater gain of information by the IG criterion.

But with the IIG criterion (and $s = 1$) we have the following values:

$$IIG(C, X_1) = 0.9403 - \frac{8}{13}0.9911 - \frac{5}{13}0.9183 = -0.0227$$

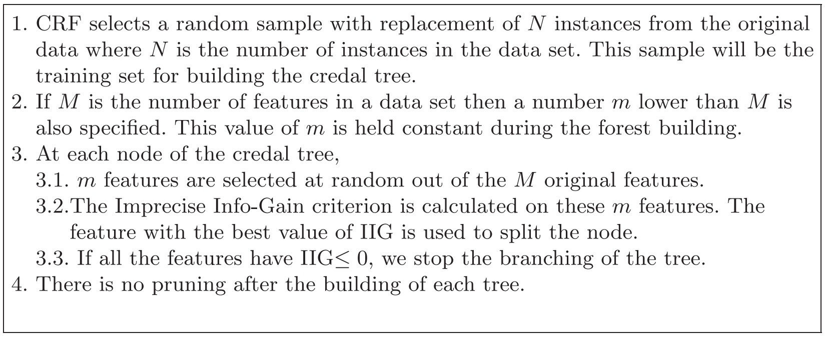
$$IIG(C, X_2) = 0.9403 - \frac{4}{13}1.0 - \frac{7}{13}0.9544 - \frac{2}{13}0.9183 = -0.0226$$

Now, the values of Imprecise Information Gain are negative. Therefore, with this criterion, there is no branching in the node J, and a leaf node is produced.

With this example we can observe that credal decision trees normally have a lower size than the ones built with classic split criteria. In the experiments of this work, it is shown that the base classifiers used for CRF algorithm (Credal Random Trees (CRTs)) have a size smaller than the base classifiers used for the RF algorithm (standard Random Tree (RT)).

[3] In fact, it is the only version described by Leo Breiman in his web about Random Forest at https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[4] Classification and regression tree.



```
1. CRF selects a random sample with replacement of $N$ instances from the original
   data where $N$ is the number of instances in the data set. This sample will be the
   training set for building the credal tree.
2. If $M$ is the number of features in a data set then a number $m$ lower than $M$ is
   also specified. This value of $m$ is held constant during the forest building.
3. At each node of the credal tree,
   3.1. $m$ features are selected at random out of the $M$ original features.
   3.2.The Imprecise Info-Gain criterion is calculated on these $m$ features. The
        feature with the best value of IIG is used to split the node.
   3.3. If all the features have IIG$\leq 0$, we stop the branching of the tree.
4. There is no pruning after the building of each tree.
```

Fig. 1. Characteristics of the credal trees designed by CRF algorithm.

3.3. Robustness to noise of the CDTs

The second important characteristic of the IIG criterion is that it is more robust to noise than the IG criterion. This is shown with the following example:

Example 2. Let us suppose a noisy data set composed by 15 instances, 9 instances of class A and 6 instances of class B. We consider that there are two binary feature variables X_1 and X_2. According with the values of these variables, the instances are organized in the following way:

$$\begin{array}{ll} X_1 = 0 \rightarrow & (3 \text{ of class } A, 6 \text{ of class } B) \\ X_1 = 1 \rightarrow & (6 \text{ of class } A, 0 \text{ of class } B) \\ X_2 = 0 \rightarrow & (1 \text{ of class } A, 5 \text{ of class } B) \\ X_2 = 1 \rightarrow & (8 \text{ of class } A, 1 \text{ of class } B) \end{array}$$

If this data set is associated with a node in a tree, then a classic DT chooses the variable X_1 for splitting the node because

$$H^{\mathcal{D}}(C|X_1) = 0.5510 < H^{\mathcal{D}}(C|X_2) = 0.5619$$

where $\mathcal{D}$ is the noisy data set composed by the 15 instances.

We can suppose that the data set is noisy because it has an outlier point when $X_2 = 1$ and class is B. In this way, the clean distribution is composed by 10 instances of class A and 5 instances of class B, that are organized as follows:

$$\begin{array}{ll} X_1 = 0 \rightarrow & (4 \text{ of class } A, 5 \text{ of class } B) \\ X_1 = 1 \rightarrow & (6 \text{ of class } A, 0 \text{ of class } B) \\ X_2 = 0 \rightarrow & (1 \text{ of class } A, 5 \text{ of class } B) \\ X_2 = 1 \rightarrow & (9 \text{ of class } A, 0 \text{ of class } B) \end{array}$$

When this data set is the one associated with a node of a tree, then a classic DT chooses the variable X_2 for splitting the node because

$$H^{\mathcal{D}}(C|X_2) = 0.2600 < H^{\mathcal{D}}(C|X_1) = 0.5946$$

where $\mathcal{D}$ is the clean data set composed by the 15 instances.

We can observe that a classic DT built with the IG criterion, generates an incorrect subtree when noisy data are processed, because it considers that the data set is reliable. However, a tree built with the IIG criterion (and $s = 1$), chooses the variable X_2 for splitting the node in both cases, when the data are noisy and when they are clean. That is,

$$H^*(K^{\mathcal{D}}(C|X_2)) = 0.7784 < H^*(K^{\mathcal{D}}(C|X_1)) = 0.8192$$

where $\mathcal{D}$ is the noisy data set, and

$$H^*(K^{\mathcal{D}}(C|X_2)) = 0.6266 < H^*(K^{\mathcal{D}}(C|X_1)) = 0.8366$$

where $\mathcal{D}$ is the clean data set.

Hence, we can see with this example the difference with respect to the robustness. In the experiments of this paper, it will be shown that the base classifiers of the CRF algorithm (Credal Random Trees) are more robust to noise than the base classifiers of the RF algorithm (classic Random Tree).

3.4. Diversity of the credal trees

The third important difference of the CRF with respect to the RF is based on the diversity of the trees in the forest used by the RF procedure. This characteristic is achieved by means of:

(1) *Bagging employed for the selection of the instances used as input for each tree*. In this way, a tree is focused on a subset of the original instances of the training data set. In this subset, it can be found properties of the data that are hidden by the global population of instances.

(2) *The random set of features considered as candidates for each node*. In this way, very strong predictors for the class variable are avoided in some nodes and thus, input variables with different information about the data can be selected in these nodes.

When the IDM is used to represent the information and the maximum entropy function is calculated on this representation, it is considered that there is a mass of unknown instances by means of the hyperparameter s. When $IIG^{\mathcal{D}}$ is calculated (Eq. (8)), the maximum entropy function H^* assigns this mass of unknown instances to the less frequent class (see Eqs. (6) and (7)). This characteristic also helps to the diversity of trees in the Random Forest. Instances with the less frequent class are enhanced. Normally, these instances are not taken into account when a tree is designed because they are hidden by the presence of more frequent instances with a different class. In this way, the distribution of unknown instances carried out by the function H^* promotes the consideration in the trees of instances and variables that are not usually taken into account.

Normally, the hyperparameter s has a low value, such as 1 or 2. This implies that, when H^* is calculated, the effect of s is negligible if the size of the data set is high. The action of s is increased when the size is low, this fact normally happens in the nodes of the lower levels of the trees. Therefore, the trees built by the RF algorithm and the trees designed by the CRF will be mainly different in the lower levels. In the CRF, the nodes of these levels encourage the instances with the less frequent class, which implies to consider instances and variables that are not usually taken into account when the trees are built. This additional diversity for the trees is obtained by the use of the IDM and the maximum entropy function in the CRF algorithm.

3.5. Considerations about properties of the CRF

Via the above paragraphs, we show that the trees built with the IIG criterion have three important properties with respect to the ones built with the classical IG criterion:

(i) With the IIG, the trees have a lower risk of overfitting.
(ii) With the IIG, a model more robust to noise is built.
(iii) With the IIG, the diversity in the lower levels of the trees is encouraged.

Using the IIG criterion in the classical RF algorithm, we obtain some important advantages with respect to the classical split criteria. These facts will be experimentally shown in the following section.

Finally, the new procedure of the CRF can be considered as a bagging scheme using a new model to build decision trees as base classifier (Credal Random Tree). This model of DT uses the CDT procedure with a randomization process of the features. That process is the same used by the base classifier of the RF, known as *Random Tree*. Hence, the new CRF has the advantage of the use of decision trees with high instability produced by the randomization process of features; and the ones of the credal trees, above explained. In the following section about experimentation, the difference of performance of these two single models on data sets with and without label noise will be exposed.

4. Experimentation

4.1. Data sets

In this section the experiments carried out are described. 50 well-known data sets in the field of machine learning have been selected, obtained from the *UCI repository of machine learning* [36]. The chosen data sets are very different in terms of their sample size, number and type of attribute variables, number of states of the class variable, etc. Table 2 gives a brief description of the characteristics of the data sets used.

4.2. Experimental setup

Two studies have been performed. In the first study, the models used as base classifiers for the ensemble algorithms are compared when they classify data with added noise. These models are: C4.5 and CDT [17] that are the base classifiers for the bagging ensembles in the principal experimentation (second study); and Random Tree (RT) and Credal Random Tree (CRT) algorithms that are the base classifiers in the RF and CRF algorithms, respectively. All the models are used without a post-pruning process in order to be coherent with the characteristics of the original RF algorithm. The average results about accuracy and tree size of the base classifiers are shown. The methods compared in this first experimentation have been noted as follows:

- C4.5
- Credal Decision Tree (CDT)
- Random Tree (RT)
- Credal Random Tree (CRT)

In the second study, CRF algorithm is compared with the original RF algorithm and with bagging schemes of the other tree based models: C4.5 [37] and CDT [21]. All the trees of the previous ensemble methods are used without pruning process in order to keep the same experimental conditions for all the algorithms to be compared. Hence, the algorithms considered in the second study are the following ones:

- Bagging C4.5 (BA-C4.5)
- Bagging CDT (BA-CDT)
- Random Forest (RF)
- Credal Random Forest (CRF)

Table 2
Data set description. Column ‘N’ is the number of instances in the data sets, column ‘Feat’ is the number of features or attribute variables, column ‘Num’ is the number of numerical variables, column ‘Nom’ is the number of nominal variables, column ‘k’ is the number of cases or states of the class variable (always a nominal variable) and column ‘Range’ is the range of states of the nominal variables of each data set.

Data set	N	Feat	Num	Nom	k	Range
anneal	898	38	6	32	6	2–10
arrhythmia	452	279	206	73	16	2
audiology	226	69	0	69	24	2–6
autos	205	25	15	10	7	2–22
balance-scale	625	4	4	0	3	–
breast-cancer	286	9	0	9	2	2-13
wisconsin-breast-cancer	699	9	9	0	2	–
car	1728	6	0	6	4	3–4
cmc	1473	9	2	7	3	2–4
horse-colic	368	22	7	15	2	2–6
credit-rating	690	15	6	9	2	2–14
german-credit	1000	20	7	13	2	2–11
dermatology	366	34	1	33	6	2–4
pima-diabetes	768	8	8	0	2	–
ecoli	366	7	7	0	7	–
glass	214	9	9	0	7	–
haberman	306	3	2	1	2	12
cleveland-14-heart-disease	303	13	6	7	5	2–14
hungarian-14-heart-disease	294	13	6	7	5	2–14
heart-statlog	270	13	13	0	2	–
hepatitis	155	19	4	15	2	2
hypothyroid	3772	30	7	23	4	2–4
ionosphere	351	35	35	0	2	–
iris	150	4	4	0	3	–
kr-vs-kp	3196	36	0	36	2	2–3
letter	20000	16	16	0	26	–
liver-disorders	345	6	6	0	2	–
lymphography	146	18	3	15	4	2–8
mfeat-pixel	2000	240	0	240	10	4–6
nursery	12960	8	0	8	4	2–4
optdigits	5620	64	64	0	10	–
page-blocks	5473	10	10	0	5	–
pendigits	10992	16	16	0	10	–
primary-tumor	339	17	0	17	21	2–3
segment	2310	19	16	0	7	–
sick	3772	29	7	22	2	2
solar-flare2	1066	12	0	6	3	2–8
sonar	208	60	60	0	2	-
soybean	683	35	0	35	19	2–7
spambase	4601	57	57	0	2	–
spectrometer	531	101	100	1	48	4
splice	3190	60	0	60	3	4–6
sponge	76	44	0	44	3	2–9
tae	151	5	3	2	3	2
vehicle	946	18	18	0	4	–
vote	435	16	0	16	2	2
vowel	990	11	10	1	11	2
waveform	5000	40	40	0	3	–
wine	178	13	13	0	3	–
zoo	101	16	1	16	7	2

In the two studies, the algorithms are compared by using the original data sets obtained from the UCI repository and these data sets where a percentage of random label noise equal to 5%, 10% and 20% are added only in the training data set.

The *Weka* software [33] has been used for the experimentation. The methods CRF and Credal Random Tree were implemented using data structures of *Weka*. Several methods were added to the implementation of the algorithms RF and Random Tree provided by *Weka* software in order to design the CRF and Credal Random Tree with the same experimental conditions.

The implementation of RF algorithm provided by *Weka* was used with its default configuration where the number of randomly chosen attributes at each node is equal to the first integer less than log_2 (number of features)+1. The only difference with the default configuration is that the number of trees used for that method was equal to 100 decision trees. The same number was used for CRF and the bagging algorithms. Although the number of trees can strongly affect the ensemble performance, this is a reasonable number of trees for the low-medium size of the data sets used in this study, and moreover it was the number of trees used in related research, such as [8].

The parameter of the IDM was set to $s = 1$ for the Imprecise Info-Gain used in CDT, CRF and Credal Random Tree. This value was used in the original method of Abellán and Moral [19] where the CDT was presented, and also in [21] where the CDT is used in a bagging scheme for classification noise tasks. The reasons to use this value were principally that it was the value recommended by Walley [16]; and the procedure to obtain the maximum entropy value reaches its lowest computational cost for this value (see [25]).

Using *Weka's* filters, random noise has been added to the class variable with the following percentages: 0%, 5%, 10% and 20%, only in the training data set. The procedure to introduce noise was the following: a given percentage of instances of the training data set was randomly selected and, then, their current class values were randomly changed to other possible values. The instances belonging to the test data set were left unmodified.

A 10-fold cross validation procedure was repeated 10 times for each data set. It is a very known and used validation procedure.

4.3. Evaluation criteria

In order to evaluate the methods of the first experiment, the measures of average accuracy and size of the base classifiers have been used. In the second study, the average accuracy has been also used to compare the ensemble methods. The particular accuracy of each method on each data set will be shown in Appendix section.

Besides, following the recommendation of Demšar [38], a series of tests were used in order to compare the ensemble methods using the *Keel* software [39]. The following tests to compare multiple classifiers on multiple data sets were utilized:

Friedman test [40,41]: a non-parametric test that ranks the algorithms separately for each data set, the best performing algorithm being assigned the rank of 1, the second best, rank 2, etc. The null hypothesis is that all the algorithms are equivalent. If the null-hypothesis is rejected, we can compare all the algorithms to each other using the **Nemenyi test** [42]. All the tests were carried out with a level of significance $\alpha = 0.05$.

To extend the comparison of the ensemble algorithms when the methods are applied on data sets with label noise, it has been used a recent measure to quantify the degree of robustness of a classifier under noise. The measure is the *Equalized Loss of Accuracy* (ELA) of Sáez et al. [43], and it can be defined as follows:

- The Equalized Loss of Accuracy (*ELA*) measure is a new behavior-against-noise measure that allows us to characterize the behavior of a method with noisy data considering performance and robustness. *ELA* measure is expressed as follows:

$$ELA_{x\%} = \frac{100 - A_{x\%}}{A_{0\%}} \qquad (9)$$

where $A_{0\%}$ is the accuracy of the classifier when it is applied on a data set without added noise and $A_{x\%}$ is the accuracy of the classifier with it is applied on a data set with level of added noise of x%.

The *ELA* measure quantifies the performance without noise considering which classifier is more suitable to work with noisy data sets. This characteristic makes it particularly useful when comparing two different classifiers over the same data set. The classifier with the lowest value for $ELA_{x\%}$ will be the most robust classifier.

Table 3
Average accuracy results of the different base classifiers when they are applied on data sets with added noise (in bold it is marked the best one and in italic the second best).

Algorithm	noise 0%	noise 5%	noise 10%	noise 20%
C4.5	**82.13**	*80.19*	*77.74*	*72.01*
CDT	*81.93*	**80.76**	**79.16**	**74.68**
RT	78.99	75.66	72.16	65.51
CRT	80.29	78.27	75.72	70.15

Table 4
Average results about the tree size of the different base classifiers when they are applied on data sets with added noise (in bold it is marked the best one and in italic the second best).

Algorithm	noise 0%	noise 5%	noise 10%	noise 20%
C4.5	*216.98*	*291.90*	*376.37*	*536.09*
CDT	**160.35**	**184.30**	**217.54**	**314.12**
RT	564.72	959.78	1238.42	1649.45
CRT	351.98	431.63	528.83	755.05

Table 5
Average accuracy results of the ensemble methods when they are applied on data sets with added noise (in bold it is marked the best one and in italic the second best).

Algorithm	noise 0%	noise 5%	noise 10%	noise 20%
BA-C4.5	85.28	84.78	84	81.5
BA-CDT	84.9	84.57	84.06	*82.15*
RF	*86.24*	*85.32*	*84.17*	80.99
CRF	**86.59**	**85.96**	**85.11**	**82.68**

Table 6
Friedman's ranks about the accuracy of the ensemble methods when they are applied on data sets with different percentages of added noise (in bold it is marked the best one and in italic the second best).

Algorithm	noise 0%	noise 5%	noise 10%	noise 20%
BA-C4.5	2.72	2.8	2.94	2.87
BA-CDT	3.23	2.99	*2.59*	*2.30*
RF	*2.26*	*2.56*	2.78	2.97
CRF	**1.76**	**1.65**	**1.69**	**1.86**

4.4. Results

Tables 3 and 4 show the results obtained by the first study, that is, the average results of accuracy and tree size for the base classifiers of the ensemble methods (C4.5, CDT, Random Tree and Credal Random Tree) for each added noise level. The best algorithm for each noise level is emphasized using bold fonts, the second best is marked with italic fonts. The previous information (average accuracy and tree size obtained by each base classifier) is illustrated in Figs. 2 and 3.

In Appendix A, it can be found the tables which show the accuracy of the ensemble methods obtained in the second study, when they classify data sets with different added noise levels. Table 5 presents a summary for the ensemble methods with the average accuracy results where the best algorithm for each added noise level is emphasized using bold fonts, the second best is marked with italic fonts. This information is graphically presented in Fig. 4.

From the statistical tests carried out, Table 6 show Friedman's ranks about the accuracy of the methods when they are applied on

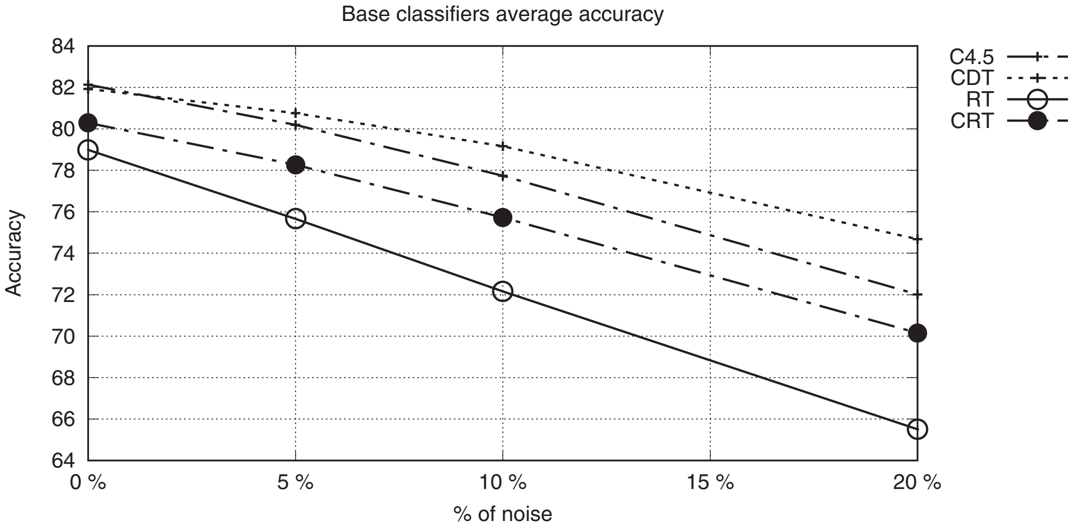


Fig. 2. Average accuracy for the base classifiers when they are applied on data sets with added noise.

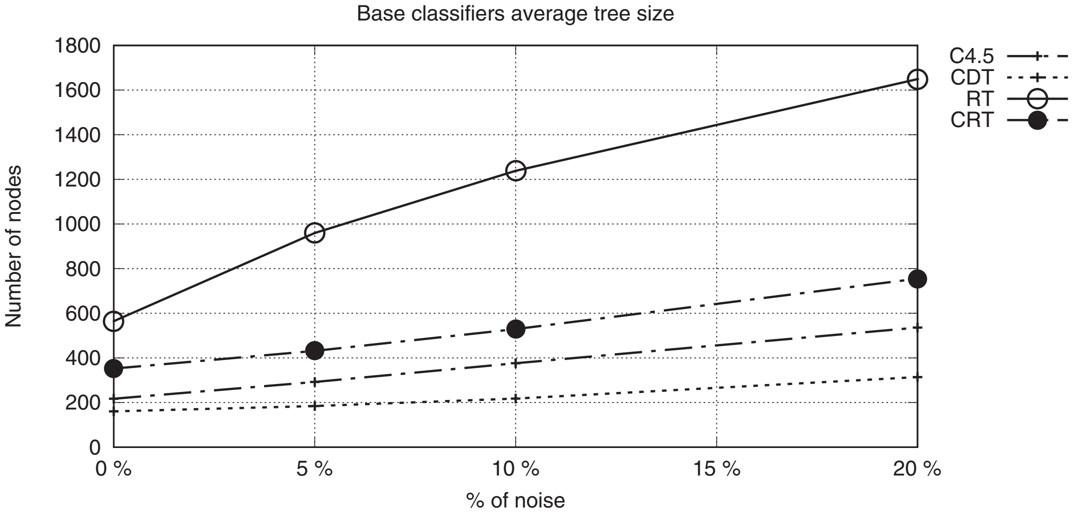


Fig. 3. Average tree size for the base classifiers when they are applied on data sets with added noise.

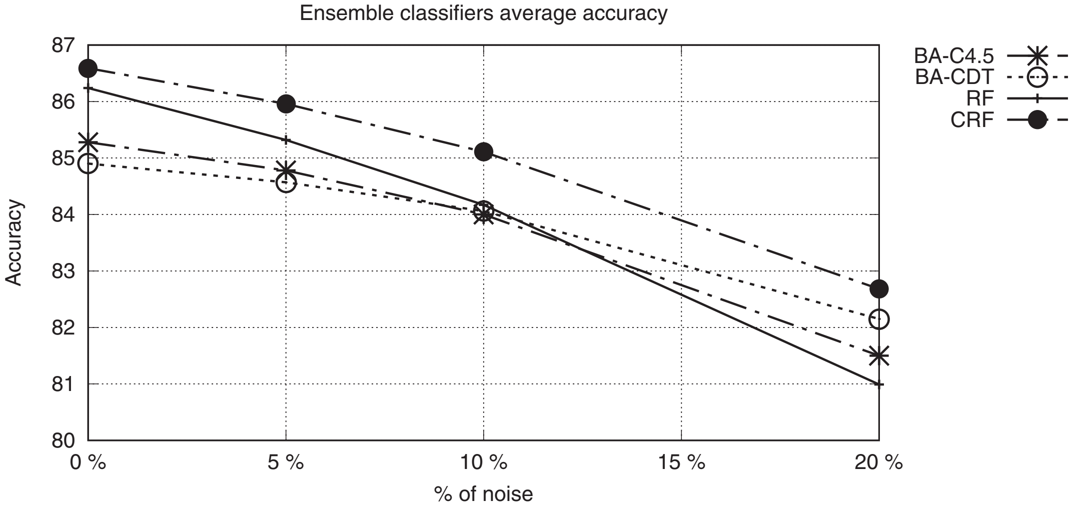


Fig. 4. Average accuracy results of the ensemble methods when they classify data sets with added noise.

Table 7
p-values of the Nemenyi test about the accuracy on data sets without added noise.

i	Algorithms	*p*
6	BA-CDT vs. **CRF**	0
5	BA-CDT vs. **RF**	0.000172
4	BA-C4.5 vs. **CRF**	0.000316
3	BA-C4.5 vs. BA-CDT	0.048243
2	RF vs. CRF	0.068713
1	BA-C4.5 vs. RF	0.074819

Table 8
p-values of the Nemenyi test about the accuracy on data sets with 5% of added noise.

i	Algorithms	*p*
6	BA-CDT vs. **CRF**	0
5	BA-C4.5 vs. **CRF**	0.000008
4	RF vs. **CRF**	0.000424
3	BA-CDT vs. RF	0.095836
2	BA-C4.5 vs. RF	0.352622
1	BA-C4.5 vs. BA-CDT	0.461812

Table 9
p-values of the Nemenyi test about the accuracy on data sets with 10% of added noise.

i	Algorithms	*p*
6	BA-C4.5 vs. **CRF**	0.000001
5	RF vs. **CRF**	0.000024
4	BA-CDT vs. **CRF**	0.000491
3	BA-C4.5 vs. BA-CDT	0.175244
2	BA-CDT vs. RF	0.461812
1	BA-C4.5 vs. RF	0.53547

Table 10
p-values of the Nemenyi test about the accuracy on data sets with 20% of added noise.

i	Algorithms	*p*
6	RF vs. **CRF**	0.000017
5	BA-C4.5 vs. **CRF**	0.000092
4	BA-CDT vs. RF	0.009462
3	BA-C4.5 vs. BA-CDT	0.027272
2	BA-CDT vs. CRF	0.08836
1	BA-C4.5 vs. RF	0.698535

data sets with different levels of added noise. The best algorithm for each noise level is emphasized using bold fonts, the second one is marked with italic fonts.Tables 7–10 show the p-values of the Nemenyi test on the pairs of comparisons when they are applied on data sets with different percentage of added noise. In all the cases, Nemenyi test rejects the hypotheses that the algorithms are equivalent if the corresponding p-value is ≤ 0.008333. When there is a significative difference, the best algorithm is distinguished with bold fonts.

In Table 11 we can see the average results of the *ELA* measure for each ensemble method that is used to classify data sets with added noise.

Table 11
Average results of the *ELA* measure for each ensemble method and noise level (in bold it is marked the best one and in italic the second best).

Algorithm	noise 5%	noise 10%	noise 20%
BA-C4.5	0.1785	0.1876	0.2169
BA-CDT	0.1817	0.1878	*0.2102*
RF	*0.1702*	*0.1836*	0.2204
CRF	**0.1621**	**0.1720**	**0.2000**

5. Comments on the results

From the results of the first study about the single methods, we can express the following comments:

- The Credal Random Tree improves the results of the Random Tree classifier for each noise level. The first one builds trees with a lower number of nodes. This fact, generally, implies a lower degree of overfitting. The difference of tree size increases with the level of noise, and this involves an greater difference of accuracy in favor of the Credal Random Tree.
- Credal Random Tree does not improve the results of the C4.5 and CDT classifiers. These two models build trees with less size and with better accuracy. Here, the randomization process of the features, that implies that not all the features can be taken into account in each node, is the reason of that difference against the Credal Random Tree.
- The Random Tree is the worse single classifier for all the levels of noise. The rest of procedures built trees with a size that depends on the level of noise, but this increasing follows similar proportionality for the three methods. However, the Random Tree increases the size of the trees in a greater level of proportionality. This can be the reason of a strong overfitting and bad accuracy when the noise increases: all the methods decrease their accuracy a percentage of 7–10% when they pass from 0% to 20% of class noise, whereas the Random Tree decreases more than 13%.

The second study is the principal one in this work. It is about the comparison of the new CRF with the RF procedure adding successful ensemble methods as reference. It must be remarked that the BA-CDT method differs from the CRF only in the process of randomization used on the features to build the trees; and, observing the results, we can see that it is a crucial point of the CRF procedure to obtain good performance. That process increases the instability of the trees used, which is an important characteristic for a single procedure to be used in a bagging scheme.

In a general comparative, it is observed that always the CRF improves the RF for each level of added noise. The difference is notably greater when the noise increases. The RF is some similar, but worse, to the CRF when noise is not added, in this particular case BA-CDT is the worst one. The situation with the greatest level of noise is the contrary one, now the RF is the worst one and BA-CDT is some similar to the CRF. The BA-C4.5 has an intermediate situation for all the levels of noise, but always the differences of this method with respect to the CRF are significative in favor of the CRF.

The following comments can be indicated in order to detail the results with respect to the measures used, focusing more on the differences between RF and the new CRF:

- **Accuracy**: Here, the results of accuracy percentage and the tests carried out are taken into account. Always, the CRF obtains the best Friedman's rank in all the comparatives carried out, with and without noise. The CRF and RF algorithms are the best models to classify data without noise, but when the level of noise increases, the RF suffers a clear deterioration motivated by the performing of the Random Tree, analyzed in the first study. CRF is always significatively better (Nemenyi's test) with respect to 2 or 3 of the rest of the models for each level of noise. Only there is no significant differences when no noise is added, with respect to the RF; and when the greatest level of noise is added, with respect to the BA-CDT. RF is significantly better than the BA-CDT method when no noise is added, but it is significantly worse than the same method when the greatest level of noise is added. The BA-C4.5 method is the worst one with 10% of level of noise and the second worst in the rest of situations.

- **ELA measure**: According to this measure, the most robust classifier to noise is the CRF algorithm, for each level of noise. The second one is the RF algorithm for levels of noise equal to 5% and 10%. However, for the greatest level of noise, the RF is the worst one. In that case the BA-CDT is the second best. The BA-C4.5 method is here the worst one with the lowest level of noise and the second worst in the rest of situations.

From the previous comments, we can conclude that CRF is the best classifier when data sets with or without noise must be classified. RF algorithm is a good classifier for data sets without noise, but it suffers the overfitting problem when data sets with noise must be classified. BA-CDT has an excellent performance to classify data sets with high level of noise but, in that case, it is not better than the CRF method. BA-C4.5 is a good method to classify data sets without noise but, in that case, it is not better than the CRF and RF method.

6. Conclusion

The Random Forest procedure is based on a randomization process of data and features. Its principal drawback is that it has a bad performance when it is applied on data sets with class noise. That is motivated by the base classifier used, built via precise probabilities and classic information measures, because it suffers of a strong overfitting when the level of class noise increases.

In this paper, the base classifier used in the Random Forest procedure is modified by replacing precise probabilities with imprecise probabilities; and classic information measures with new general information measures. The differences between the new base classifier and the original one used by the RF classifier have been shown. In an experimental study, It is also exposed the different performance of those two base models in situations where there are data sets with class noise.

The combination of the randomization of features and data, join with the application of imprecise probabilities and general uncertainty measures implies that the new base classifier obtains a good trade-off between accuracy and instability. These are important characteristics to be used in an ensemble scheme for classification.

This new base classifier is used in a Random Forest scheme to present here a new classifier called Credal Random Forest. In a second experimental study, it is shown that this new method of classification represents an improvement of the Random Forest when both are applied on data sets without class noise. The improvement is very important when they are applied on data sets with class noise, where the new method is significantly better than the old method. It is exposed that the new method is even better than other successful ensemble schemes of classification. It has been shown, via a measure of robustness under noise, that the new method is also very robust on the presence of class noise. Hence, the new method of the Credal Random Forest can be considered as an important alternative to be used in grounds where the standard Random Forest is applied.

Acknowledgments

This work has been supported by the Spanish "Ministerio de Economía y Competitividad" and by "Fondo Europeo de Desarrollo Regional" (FEDER) under Project TEC2015-69496-R.

Appendix A. Tables about accuracy results

Tables 12–15 are presented in this Appendix. They show the accuracy results obtained by the ensemble methods when they classify data sets with different added noise levels.

Table 12
Accuracy results of the ensemble methods when they are used on data sets without added noise.

Dataset	BA-C4.5	BA-CDT	RF	CRF
anneal	98.9	98.89	99.68	99.71
arrhythmia	75.35	74.49	69.12	69.93
audiology	81.83	80.41	80.36	81.28
autos	85.45	80.27	84.29	85.32
balance-scale	81.56	82.41	80.3	81.94
breast-cancer	70.43	70.35	70.02	73.53
wisconsin-breast-cancer	96.45	96.14	96.58	96.55
car	94.33	93.55	94.7	94.44
cmc	52.19	53.21	50.69	52.09
horse-colic	85.51	84.91	85.59	85.18
credit-rating	85.68	86.07	86.14	86.87
german-credit	73.01	74.64	76.08	76.38
dermatology	97.13	94.18	96.91	97.87
pima-diabetes	76.14	75.8	76.01	75.86
ecoli	84.88	83.82	84.67	85.27
Glass	74.49	75.51	79.71	78.87
haberman	70.17	73.76	65.44	72.56
cleveland-14-heart-diseas	80.23	78.68	81.56	81.26
hungarian-14-heart-diseas	78.92	81.18	80.25	80.54
heart-statlog	80.96	81.41	82.26	82
hepatitis	81.76	80.99	83.58	83.37
hypothyroid	99.62	99.59	99.51	99.7
ionosphere	92.57	91.23	93.48	93.65
iris	94.47	95.07	94.53	94.6
kr-vs-kp	99.46	99.4	99.27	99.34
letter	94.03	92.44	96.6	96.54
liver-disorders	73.42	72.21	72.03	72.6
lymphography	79.96	76.24	83.42	82.34
mfeat-pixel	83.86	87.2	96.37	96.65
nursery	98.68	96.66	99.17	96.86
optdigits	95.84	95.55	98.3	98.38
page-blocks	97.36	97.32	97.46	97.6
pendigits	98.32	98.45	99.21	99.21
primary-tumor	44.22	43.93	43.45	44.72
segment	97.75	97.45	98.16	98.19
sick	98.97	98.97	98.43	98.59
solar-flare2	99.49	99.53	99.43	99.53
sonar	80.07	80.78	84.63	84.55
soybean	92.28	90.47	93.31	94.95
spambase	94.73	94.65	95.68	95.57
spectrometer	56.61	54.48	57.42	57.91
splice	94.7	94.4	95.88	96.31
sponge	93.91	92.63	95	95
tae	60.88	60.88	68.25	67.37
vehicle	75.22	74.78	75.18	74.96
vote	96.78	96.34	96.43	96.55
vowel	94.04	92.17	98.16	98.22
waveform	83.4	83.51	85.2	85.15
wine	95.34	95.84	97.74	97.51
zoo	92.8	92.4	96.33	96.25
Average	85.28	84.9	*86.24*	**86.59**
Standard desv.	13.17	13.02	13.55	13.10

Table 13
Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 5%.

Dataset	BA-C4.5	BA-CDT	RF	CRF
anneal	98.83	98.78	98.2	99.11
arrhythmia	75.05	74.27	68.56	69.16
audiology	81.32	80.36	78.82	80.52
autos	83.54	78.56	80.04	82.2
balance-scale	81.71	82.25	79.21	82.05
breast-cancer	69.03	70.63	68.26	72.56
wisconsin-breast-cancer	95.94	96.04	95.81	96.52
car	92.88	93.28	94.12	93.99
cmc	51.17	52.5	49.55	50.94
horse-colic	85.21	84.15	85.12	84.88
credit-rating	84.87	85.77	85.23	86.39
german-credit	72.81	73.96	74.96	75.91
dermatology	96.53	94.31	96.53	97.37
pima-diabetes	75.88	74.88	74.88	75.15
ecoli	84.05	83.93	84.25	84.64
Glass	74.67	75.03	77.97	77.99
haberman	70.07	72.23	64.63	70.97
cleveland-14-heart-diseas	79.9	79.67	80.84	80.73
hungarian-14-heart-diseas	79.23	79.9	80.37	80.28
heart-statlog	79.89	79.85	80.89	80.85
hepatitis	81.12	81.76	82.88	83.25
hypothyroid	99.53	99.55	99.35	99.62
ionosphere	92.25	91.54	92.97	93.14
iris	94.33	94.6	91.87	92.93
kr-vs-kp	99.08	99.16	98.3	98.87
letter	93.85	92.58	95.41	96.23
liver-disorders	72.22	71.65	71.24	71.3
lymphography	79.63	77.02	83.77	82.52
mfeat-pixel	83.71	86.63	96.11	96.4
nursery	97.87	97.01	98.59	97.34
optdigits	95.65	95.73	98.31	98.42
page-blocks	97.31	97.33	97.13	97.45
pendigits	98.46	98.44	99.15	99.17
primary-tumor	43.01	43.28	43.15	44.54
segment	97.59	97.38	97.09	97.42
sick	98.68	98.66	98.4	98.5
solar-flare2	99.16	99.5	98.61	99.53
sonar	79.43	80.35	83.18	83.92
soybean	91.95	90.5	92.17	94.61
spambase	94.12	94.17	94.58	94.54
spectrometer	55.86	52.94	56.84	57.24
splice	94.02	94.04	94.89	95.53
sponge	92.75	92.57	94.61	94.32
tae	58.75	59	64.61	63.27
vehicle	74.51	74.1	74.92	74.97
vote	96.04	95.79	95.56	96.04
vowel	93.35	91.78	95.41	96.15
waveform	83.18	83.14	85.03	84.94
wine	95.89	95.27	97.85	97.24
zoo	93.01	92.5	95.67	96.18
Average	84.78	84.57	*85.32*	**85.96**
Standard desv.	13.46	13.37	13.74	13.39

Table 14
Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 10%.

Dataset	BA-C4.5	BA-CDT	RF	CRF
anneal	98.05	98.5	96.44	98.34
arrhythmia	74.29	73.88	67.74	68.83
audiology	80.84	79.28	75.72	78.83
autos	80.44	75.79	77.21	79.06
balance-scale	81.09	81.97	78.03	81.26
breast-cancer	67.17	69.87	66.77	70.89
wisconsin-breast-cancer	95.49	95.75	94.61	96.01
car	90.92	92.34	93.3	93.44
cmc	50.12	51.82	48.51	50.17
horse-colic	84.55	83.71	83.61	83.31
credit-rating	83.3	84.77	84.01	85.26
german-credit	72.67	73.43	74.79	75.05
dermatology	95.46	93.82	96.25	96.88
pima-diabetes	75.59	74.48	74.24	74.16
ecoli	84.82	84.7	83.87	84.76
Glass	73.33	74.37	76.82	76.27
haberman	69.05	70.44	62.66	69.72
cleveland-14-heart-diseas	80.3	79.73	80.76	80.6
hungarian-14-heart-diseas	78.96	79.46	79.56	79.77
heart-statlog	79.7	79.26	79.37	79.78
hepatitis	80.63	81.53	82.71	82.14
hypothyroid	99.3	99.48	99.24	99.48
ionosphere	91.8	90.58	92.31	92.42
iris	93.8	94.2	90.07	92.33
kr-vs-kp	98.02	98.72	96.57	98.09
letter	93.56	92.56	94.04	95.87
liver-disorders	70.47	69.43	69.38	69.87
lymphography	79.58	77.02	83.09	82.62
mfeat-pixel	83.09	86.71	95.82	96.33
nursery	96.27	97.11	97.55	97.55
optdigits	95.7	95.81	98.26	98.34
page-blocks	97.11	97.2	96.49	97.15
pendigits	98.43	98.43	99.08	99.08
primary-tumor	41.62	43.06	42.15	43.36
segment	96.75	97.08	95.92	96.34
sick	98.08	98.47	98.17	98.28
solar-flare2	98.58	99.47	97.56	99.46
sonar	77.45	79.47	81.61	82.08
soybean	91.22	90.25	90.41	94.03
spambase	93.23	93.32	93.13	93.33
spectrometer	55.42	51.85	56.39	56.41
splice	93.11	93.54	93.98	94.73
sponge	91.39	92.68	92.98	93.52
tae	56.17	57.15	61.69	60.28
vehicle	73.88	73.54	74.48	74.49
vote	95.22	95.35	94.11	95.28
vowel	92.73	90.74	92.18	93.28
waveform	83.16	83.16	84.94	84.9
wine	94.44	94.5	96.86	96.19
zoo	93.66	93.37	92.97	95.86
Average	84	84.06	*84.17*	**85.11**
Standard desv.	13.56	13.54	13.77	13.60

Table 15
Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 20%.

Dataset	BA-C4.5	BA-CDT	RF	CRF
anneal	95.34	97.42	91.16	95.51
arrhythmia	73.87	72.84	66.75	66.73
audiology	76.25	75.57	71.28	75.23
autos	73.34	69.8	70.63	73.4
balance-scale	79.26	80.97	75.28	80.38
breast-cancer	63.4	66.2	62.02	66.79
wisconsin-breast-cancer	93.41	94	90.83	93.48
car	85.43	89.72	90.48	91.53
cmc	48.38	50.14	46.58	48.7
horse-colic	81.46	80.73	80.7	81.3
credit-rating	79.41	82.67	80	82.77
german-credit	69.91	71.38	71.8	72.71
dermatology	92.73	93.52	94.86	95.76
pima-diabetes	74.62	72.6	71.85	72.68
ecoli	82.56	82.91	80.74	81.48
Glass	70.61	72.67	72.72	71.94
haberman	66.55	66.33	59.43	64.05
cleveland-14-heart-diseas	79.02	79.15	79.48	79.91
hungarian-14-heart-diseas	78.14	78.36	77.81	78.75
heart-statlog	76.93	76.81	76.93	77
hepatitis	79.35	79.95	79.69	79.56
hypothyroid	98.34	99.37	98.65	99.14
ionosphere	87.84	86.7	88.39	88.15
iris	90.07	90.93	82.8	88.87
kr-vs-kp	92.68	95.63	90.37	93.27
letter	92.57	92.32	90.57	94.6
liver-disorders	67.08	66.45	65.84	65.7
lymphography	75.99	76	78.08	80.58
mfeat-pixel	82.19	86.6	95.32	95.85
nursery	90.42	96.5	93.74	97.2
optdigits	95.73	96.07	98.01	98.08
page-blocks	96.33	96.79	94.68	95.97
pendigits	98.08	98.19	98.75	98.83
primary-tumor	40.2	41.03	40.53	41.8
segment	94.29	95.83	93.48	93.92
sick	96.14	97.87	96.82	97.3
solar-flare2	96.45	99.23	94.76	98.99
sonar	74.77	76.27	78.54	79
soybean	88.07	87.7	84.83	92.33
spambase	90.39	89.95	89.33	89.63
spectrometer	54.15	49.97	55.86	55.03
splice	90.87	91.5	91.52	92.22
sponge	87.89	90.57	89.45	90.82
tae	53.13	54.8	54.87	54.82
vehicle	72.59	72.41	72.52	72.87
vote.	92.59	93.93	90.55	93.26
vowel	88.88	84.42	84.23	85.79
waveform	82.7	82.8	84.46	84.41
wine	91.35	90.68	93.61	92.6
zoo	93.5	93.27	87.83	93.1
Average	81.5	*82.15*	80.99	**82.68**
Standard desv.	13.51	13.93	13.82	13.98

References

[1] D.J. Hand, Construction and Assessment of Classification Rules, John Wiley and Sons, New York, 1997.

[2] D. Hand, Discrimination and Classification, John Wiley, 1981.

[3] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[4] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, CA, 1988.

[5] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106, doi:10.1023/A:1022643204877.

[6] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection, Inf. Fusion 6 (1) (2005) 83–98, doi:10.1016/j.inffus.2004.04.003.

[7] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140, doi:10.1023/A:1018054314350.

[8] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: L. Saitta (Ed.), Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996), Morgan Kaufmann, 1996, pp. 148–156.

[9] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, doi:10.1023/A:1010933404324.

[10] D.H. Wolpert, The Supervised Learning No-Free-Lunch Theorems, Springer London, London, pp. 25–42. doi:10.1007/978-1-4471-0123-9-3.

[11] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1) (2014) 3133–3181.

[12] J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Inf. Fusion 14 (4) (2013) 423–430.

[13] J. Abellán, An application of non-parametric predictive inference on multi-class classification high-level-noise problems, Expert Syst. Appl. 40 (11) (2013) 4585–4592, doi:10.1016/j.eswa.2013.01.066.

[14] B. Frenay, M. Verleysen, Classification in the presence of label noise: a survey, Neural Netw. Learn. Syst. IEEE Trans. 25 (5) (2014) 845–869, doi:10.1109/TNNLS.2013.2292894.

[15] G.J. Klir, Uncertainty and Information: Foundations of Generalized Information Theory, John Wiley And Sons, Inc., 2005, doi:10.1002/0471755575.

[16] P. Walley, Inferences from multinomial data; learning about a bag of marbles (with discussion)., J. R. Stat. Soc. Ser. B (Methodological) 58 (1) (1996) 3–57, doi:10.2307/2346164.

[17] J. Abellán, S. Moral, Building classification trees using the total uncertainty criterion, Int. J. Intell. Syst. 18 (12) (2003) 1215–1225, doi:10.1002/int.10143.

[18] J. Abellán, G. Klir, S. Moral, Disaggregated total uncertainty measure for credal sets, Int. J. Gen. Syst. 35 (1) (2006) 29–44, doi:10.1080/03081070500473490.

[19] J. Abellán, S. Moral, Upper entropy of credal sets. applications to credal classification imprecise Probabilities and Their Application, Int. J. Approximate Reasoning 39 (2–3) (2005) 235–255, doi:10.1016/j.ijar.2004.10.001. Imprecise Probabilities and Their Applications

[20] J. Abellán, A. Masegosa, A filter-wrapper method to select variables for the naive bayes classifier based on credal decision trees, Int. J. Uncertainty Fuzziness Knowl. Based Syst. 17 (06) (2009) 833–854, doi:10.1142/S0218488509006297.

[21] J. Abellán, A.R. Masegosa, Bagging schemes on the presence of class noise in classification, Expert Syst. Appl. 39 (8) (2012) 6827–6837, doi:10.1016/j.eswa.2012.01.013.

[22] J. Abellán, A. Masegosa, An experimental study about simple decision trees for bagging ensemble on datasets with classification noise, in: C. Sossai, G. Chemello (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Lecture Notes in Computer Science, 5590, Springer Berlin Heidelberg, 2009, pp. 446–456, doi:10.1007/978-3-642-02906-6_39.

[23] E. Jaynes, On the rationale of maximum-entropy methods, Proc. IEEE 70 (9) (1982) 939–952, doi:10.1109/PROC.1982.12425.

[24] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.

[25] J. Abellán, Uncertainty measures on probability intervals from the imprecise dirichlet model, Int. J. Gen. Syst. 35 (5) (2006) 509–528, doi:10.1080/03081070600687643.

[26] A. Shahpari, S.A. Seyedin, Using mutual aggregate uncertainty measures in a threat assessment problem constructed by dempster–shafer network, IEEE Trans. Syst. Man Cybern. 45 (6) (2015) 877–886, doi:10.1109/TSMC.2014.2378213.

[27] Y. Deng, Deng entropy, Chaos Solitons Fractals 91 (2016) 549–553. doi: 110.1016/j.chaos.2016.07.014.

[28] Y. Yang, D. Han, A new distance-based total uncertainty measure in the theory of belief functions, J. Knowl. Based Syst. 94 (C) (2016) 114–123, doi:10.1016/j.knosys.2015.11.014.

[29] J. Abellán, É. Bossé, Drawbacks of uncertainty measures based on the pignistic transformation, IEEE Trans. Syst. Man Cybern. (99) (2016) 1–7, doi:10.1109/TSMC.2016.2597267.

[30] J. Abellán, Analyzing properties of deng entropy in the theory of evidence, Chaos Solitons Fractals 95 (2017) 195–199, doi:10.1016/j.chaos.2016.12.024.

[31] W. Buntine, T. Niblett, A further comparison of splitting rules for decision-tree induction, Mach. Learn. 8 (1) (1992) 75–85, doi:10.1023/A:1022686419106.

[32] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Statistics/Probability Series, Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[33] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Series in Data Management Systems, 2nd, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[34] L.E. Raileanu, K. Stoffel, Theoretical comparison between the gini index and information gain criteria, Ann. Math. Artif. Intell. 41 (1) (2004) 77–93, doi:10.1023/B:AMAI.0000018580.96245.c6.

[35] V.Y. Kulkarni, M. Petare, P.K. Sinha, Analyzing random forest classifier with different split measures, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), Springer India, New Delhi, 2012, pp. 691–699, doi:10.1007/978-81-322-1602-5_74.

[36] M. Lichman, UCI machine learning repository, 2013.

[37] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Mach. Learn. 40 (2) (2000) 139–157, doi:10.1023/A:1007607513941.

[38] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[39] J. Alcalá-Fdez, L. Sánchez, S. García, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernández, F. Herrera, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2009) 307–318, doi:10.1007/s00500-008-0323-y.

[40] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (1937) 675–701.

[41] M. Friedman, A comparison of alternative tests of significance for the problem of *m* rankings, Ann. Math. Stat. 11 (1) (1940) 86–92, doi:10.1214/aoms/1177731944.

[42] P. Nemenyi, Distribution-free multiple comparisons, Princeton University, 1963 Doctoral dissertation. New Jersey, USA.

[43] J.A. Sáez, J. Luengo, F. Herrera, Evaluating the classifier behavior with noisy data considering performance and robustness: the equalized loss of accuracy measure, Neurocomputing 176 (2014) 26–35. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems, selected papers from the {HAIS} 2013 conference doi: 10.1016/j.neucom.2014.11.086.