# Increasing diversity in Random Forest learning algorithm via imprecise probabilities

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

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

**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.

**Keywords**: Classification, ensemble schemes, Random Forest, imprecise probabilities; uncertainty measures

## 1 Introduction

The classification task (D. J. 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 (D. 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, & Yao, 2005; Ren, Zhang, & Suganthan, 2016).

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, & Jamshidi, 2017; Wang et al., 2015); ensembles of NNs, RF and other models in (Krauss, Do, & 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, & Lee, 2016; 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 & 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 & Moral, 2005), (Abellán & Masegosa, 2009)). The treatment of the imprecision is different when imprecise probabilities are used. This fact has been experimentally shown in (Abellán & Masegosa, 2012; Mantas & Abellán, 2014a; Abellán & Mantas, 2014; Mantas & 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

[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 & Verleysen, 2014).

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 & 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 will introduce the split criterion used by the CDT algorithm in a comparative way with the classic ID3. The new criterion can be considered as a parametric extension of the one of the ID3.

Let $C$ be the class variable with states $\{c_1, \cdots, c_k\}$; and $X$ be a general feature whose values belong to $\{x_1, \ldots, x_t\}$. Let $\mathcal{D}$ be a data set. The Info-Gain (IG) criterion was introduced by Quinlan as the basis for his ID3 model (Quinlan, 1986), and it is explained as follows:

- The entropy of C for the data set $\mathcal{D}$ is the Shannon's entropy (Shannon, 1948) and it is defined as:

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

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

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

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

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

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

The feature that represents the greatest gain in information is selected for branching.

The Imprecise Info-Gain (IIG) (Abellán & Moral, 2003) is based on imprecise probabilities and the application of uncertainty measures on credal sets. It was introduced to build the called *Credal Decision Tree* model (CDT). Probability intervals are obtained from the data set using Walley's Imprecise Dirichlet

Model (IDM) (Walley, 1996) (a special type of credal sets (Abellán, 2006)). The mathematical basis applied is described below.

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:

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

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 parameter. The value of the 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, 1996) 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.

This representation gives rise to a specific kind of credal set on the variable $C$, $K^{\mathcal{D}}(C)$ (Abellán, 2006). The set is defined as

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

In the Example 1 we can see a practical case where a credal set associated with the IDM is shown.

*Example 1.* Let $C$ be a class variable with three possible states $\{c_1, c_2, c_3\}$. We consider a data set, $\mathcal{D}$, where we have the following frequencies $\{c_1 : 1,\ c_2 : 2, c_3 : 4\}$. Then the associated credal set from the IDM, for $s = 1$, is the following set of probability distributions:

$$K^{\mathcal{D}}(C) = \left\{ p \,|\, p(c_1) \in \left[ \frac{1}{8}, \frac{2}{8} \right]; p_2 \in \left[ \frac{2}{8}, \frac{3}{8} \right]; p_2 \in \left[ \frac{4}{8}, \frac{5}{8} \right] \right\}.$$

Hence,

$$K^{\mathcal{D}}(C) = CH \left\{ (\frac{1}{8}, \frac{2}{8}, \frac{5}{8}); (\frac{1}{8}, \frac{3}{8}, \frac{4}{8}); (\frac{2}{8}, \frac{2}{8}, \frac{4}{8}) \right\},$$

where with $CH$ we express the *convex hull* of those probability distributions.

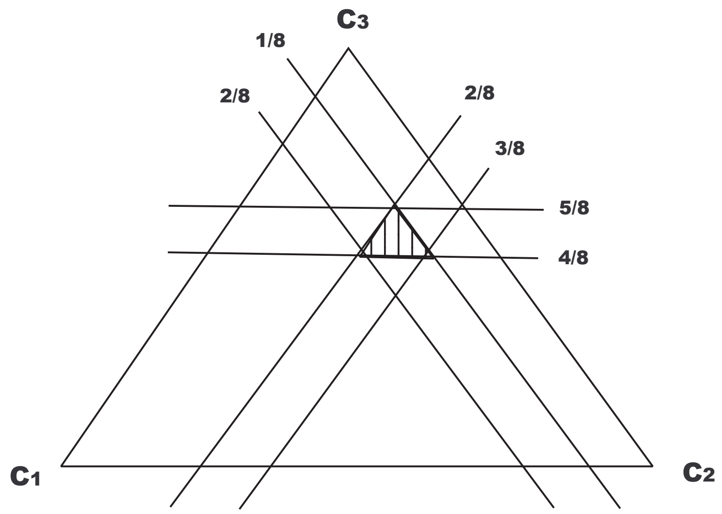
This credal set can be seen in Figure 1, where we use a simplex 2-dimensional representation of a credal set on the 3-dimensional space $I\!R^3$.

On this type of sets of probability distributions (convex and closed sets, i.e. credal sets (Abellán, 2006)), uncertainty measures can be applied. The procedure to build CDTs uses the maximum of entropy function on the above defined credal set. This function, denoted as $H^*$, is defined as follows:

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

The procedure to obtain $H^*$ for the special case of the IDM reaches its lowest computational cost for $s \leq 1$ (see (Abellán, 2006) for more details). For

**Fig. 1.** Simplex representation of the credal set from Example 1. One point in the triangle of height one represents a probability distribution $(p(c_1), p(c_2), p(c_3)) \in \mathbb{R}^3$ such that $p(c_i)$ is the distance of this point to the opposite side of the vertex $c_i$.




that value, that procedure is simple and treats to share a mass of $s$ among on all the cases of the class variable with minimum frequency, starting from the lower possible values of probability taken from the intervals of the IDM. In the Example 1, the value of $s = 1$ will be assigned to the case $c_1$

$$p(c_1) = \frac{1}{8} \rightarrow \frac{2}{8}$$
$$p(c_2) = \frac{2}{8} \rightarrow \frac{2}{8}$$
$$p(c_3) = \frac{4}{8} \rightarrow \frac{4}{8}$$

Hence, the value of maximum entropy is attained on the probability distribution $(\frac{2}{8}, \frac{2}{8}, \frac{4}{8})$. If $s$ has a value upper 1 then the procedure will be repeated using portions of mass $\leq 1$ (see (Mantas et al., 2016)). For example, for $s = 2.5$, the procedure can be called 3 times.[3]. (for the values $s = 1$, $s = 1$ and $s = 0.5$)

The scheme to induce CDTs is like the one used by the classical ID3 algorithm (Quinlan, 1986), replacing its *Info-Gain* Split criterion with the *Imprecise Info-Gain* (IIG) split criterion which can be defined by the following way:

$$IIG^{\mathcal{D}}(C, X) = H^*(K^{\mathcal{D}}(C)) - H^*(K^{\mathcal{D}}(C|X)), \tag{7}$$

[3] The more efficient general algorithm of (Abellán & Moral, 2003) can be applied too.

where $H^*(K^{\mathcal{D}}(C|X))$ is calculated via a similar way than $H^{\mathcal{D}}(C|X)$ in the IG criterion.[4] Here, the feature with the greatest gain of information is selected for branching, as with the IG criterion. The criterion has a clear dependence of the parameter $s$, then it can be noted as $IIG_s^{\mathcal{D}}(C,X)$.

It must be taken into account that for a variable X and a data set $\mathcal{D}$, $IIG_s^{\mathcal{D}}(C,X)$ 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 property of the CDT model, which uses the IIG criterion, that can be considered as an additional criterion to stop the branching of the tree. In the Example 2 we can see a case where both criteria give us different type of situation for branching.

*Example 2.* 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:

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

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

$$IG(C, X_1) = 0.6172 - \frac{8}{13}0.6615 - \frac{5}{13}0.5004 = 0.0177$$

$$IG(C, X_2) = 0.6172 - \frac{4}{13}0.6931 - \frac{7}{13}0.5983 - \frac{2}{13}0 = 0.0818$$

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 we have the following values, for $s = 1$:

$$IIG_{s=1}(C, X_1) = 0.6518 - \frac{8}{13}0.6850 - \frac{5}{13}0.6368 = -0.0002$$

$$IIG_{s=1}(C, X_2) = 0.6518 - \frac{4}{13}0.6931 - \frac{7}{13}0.6615 - \frac{2}{13}0.6368 = -0.0157$$

Now, with this criterion, there is no branching in the node $J$ and a leaf node is produced.

[4] For a more extended explanation see (Mantas & Abellán, 2014b).

### 2.1.2 Algorithm

The procedure for building credal trees is very close to the one used in the well-known Quinlan's ID3 algorithm (Quinlan, 1986), replacing its *Info-Gain* split criterion with the *Imprecise Info-Gain* . It can be described as follows.

Each node $No$ in a decision tree, produces a partition of the data set (for the root node, $\mathcal{D}$ is considered to be the entire data set). Furthermore, each node $No$ has an associated list $\mathcal{L}$ of feature labels (that are not in the path from the root node to $No$). The procedure for building CDTs is explained in the algorithm in Figure 2. Here, the procedure starts with $No$ as the root node, and for the first call to the algorithm (the one associated with the root node), the data set $\mathcal{D}' = \mathcal{D}$.

Procedure `CDT(`$No$`,`$\mathcal{L}$`,`$\mathcal{D}$`,`$s$`)`

1. If $\mathcal{L} = \emptyset$, then `Exit`.
2. Let $\mathcal{D}'$ be the partition associated with node $No$
3. Compute the value
$$\alpha = \max_{X_j \in \mathcal{L}} \left\{ IIG_s^{\mathcal{D}'}(C, X_j) \right\}$$
4. If $\alpha \leq 0$ then `Exit`
5. Else
   6. Let $X_l$ be the variable for which the maximum $\alpha$ is attained
7. Remove $X_l$ from $\mathcal{L}$
8. Assign $X_l$ to node $No$
9. For each possible value $x_l$ of $X_l$
   10. Add a node $No_l$
   11. Make $No_l$ a child of $No$
   12. Call `CDT(`$No_l$`,`$\mathcal{L}$`,`$\mathcal{D}'$`,`$s$`)`

**Fig. 2.** Procedure to build a CDT.

In this algorithm, when an *Exit* situation is attained, i.e. when there are no more features to introduce in a node, or when the uncertainty is not reduced (steps 1 and 4 of the algorithm, respectively), a leaf node is produced.

In a leaf node, the most probable state or value of the class variable for the partition associated with that leaf node is inserted.[5]

## 2.2 Random Forest

The base classifier of the RF algorithm is denominated *Random Tree* (RT). The RF algorithm builds a forest of RTs. If $M$ is the number of features in a data set

[5] To avoid obtaining unclassified instances, if we do not have one single most probable class value, we can select the one obtained in its parent node, and so recursively (see (Abellán & Masegosa, 2010)).

then a number $m << M$ is specified.[6] This value of $m$ is held constant during the forest building and will be used to select features in each node randomly. For each RT, 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 DT. This type of decision tree, RT, is built with the following characteristics:

1. At each node of the random tree,
   1.1. $m$ features are selected at random out of the $M$ original features.
   1.2. The split criterion is calculated on these $m$ features. The feature with the best value is used to split the node.
2. There is no pruning after building each random tree.

If a new instance must be classified in the RF algorithm, the features of this instance are presented to each RT in the forest. Each RT returns a classification value, a vote for that class. Finally, the classification value given by RF is the one associated with the most voted state of the class variable, over all the DTs in the forest.

The original split criterion used by RF was the Gini Index, also based on classical probabilities, which was used by the CART[7] algorithm (Breiman, Friedman, Olshen, & Stone, 1984). In this work the Information Gain criterion is used due to the fact that *Weka* software (Witten & Frank, 2005) 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 (Raileanu & Stoffel, 2004), which explains why empirical works (see (Raileanu & Stoffel, 2004; Kulkarni, Petare, & Sinha, 2012)) 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 The Random Credal Random Forest classifier

The RCRF procedure is similar to the RF approach presented in the previous paragraph. The main difference is that RCRF uses a new base classifier, called *Random Credal Random Tree* (RCRT), instead of RT algorithm. RCRT utilizes the Imprecise Info-Gain measure to split each node instead of using Info-Gain or Gini Index. Besides, a random value for the parameter $s$ is selected for each new built RCRT.

The procedure for building a RCRT is similar to the one of the CDT, adding random procedures to select features and value of $s$ parameter. With the same notation than the one used for the algorithm of the CDT, the algorithm of RCRT is explained in Figure 3.

The selected feature to split each node in step 9 and the ramification of RCRT depends on the IIG criterion. This criterion is calculated in terms of the

[6] Normally the value used for $m$ is the integer part of $\log_2$ (number of features)+1

[7] ClAssification and Regression Tree.

Procedure `RCRT(`$No$`,`$\mathcal{L}$`,`$\mathcal{D}$`,`$s$`)`

1. If $\mathcal{L} = \emptyset$, then `Exit`.
2. Calculate the $m$ value.
3. Let $\mathcal{L}'$ a subset of $m$ features randomly selected from $\mathcal{L}$
4. Let $\mathcal{D}'$ be the partition of $\mathcal{D}$ associated with node $No$
5. Compute the value

$$\alpha = \max_{X_j \in \mathcal{L}'} \left\{ IIG_s^{\mathcal{D}'}(C, X_j) \right\}$$

6. If $\alpha \leq 0$ then `Exit`
7. Else
   8. Let $X_l$ be the variable for which the maximum $\alpha$ is attained
9. Remove $X_l$ from $\mathcal{L}$
10. Assign $X_l$ to node $No$
11. For each possible value $x_l$ of $X_l$
    12. Add a node $No_l$
    13. Make $No_l$ a child of $No$
    14. Call `RCRT(`$No_l$`,`$\mathcal{L}$`,`$\mathcal{D}'$`,`$s$`)`

**Fig. 3.** Procedure to build a RCRT.

parameter $s$. If the value for $s$ is randomly selected, different RCRTs are obtained by the features inside the nodes and by the size of the trees. This property is important in order to provide diversity for the ensemble members in the RCRF algorithm.

Now, the procedure to obtain the forest of DTs from RCRF can be exposed in the Figure 4.

Procedure `RCRF(`$\mathcal{D}$`,`$\mathcal{L}$`)`

1. Fix $nT$ the number of trees to be used
2. For $i = 1$ to $nT$ do
   3. Select randomly the value of $s$ from the set $\{1, 1.5, 2, 2.5, 3, 3.5\}$ and name it $s'$
   4. Let $\mathcal{D}'$ be a partition of size $|\mathcal{D}|$ obtained from $\mathcal{D}$ with replacement
   5. Call RCRT($No$,$\mathcal{L}$,$\mathcal{D}'$,$s'$) to build DT$_i$

**Fig. 4.** Procedure to obtain a forest of DTs via RCRF.

As with RF, when a new instance must be classified in the RCRF algorithm, the features of this instance are presented to each DT (RCRT) of the forest $\{DT_i\}_{i=1}^{nT}$. Each DT$_i$ returns a classification value and the final classification of RCRF is the one associated with the most voted state of the class variable.

## 4 Justification of the new classifier

The base classifier of RCRF, RCRT, represents a modification of the CDT algorithm adding randomness in the set of features taken into account to be inserted in a node; and on the value of the $s$ parameter. In the RCRF procedure, if the value for $s$ is randomly changed for each RCRT, then the diversity of the trees is increased (as we will see in the next subsection) and this produces a richer variety of rules from the decision trees. In other words, the information obtained from data is increased. Moreover, as it has been shown in previous works (Abellán, 2013; Abellán & Masegosa, 2012; Mantas & Abellán, 2014b), that the new split criterion $IIG$, used in the base classifier of RCRF, gives us better results than the classic ones, specially in noise domains. Hence, the new model has an increased randomness join with a more success base classifier procedure. Hence, the principal differences of the new model with respect to the RF procedure can be summarizes as follows:

· The randomness in the forest of RCRF is increased with respect to the one of RF.
· RCRF uses as base classifier a DT with a more successful split criterion than the one used by the DT in RF.
· The above mentioned split criterion, based on imprecise probabilities, produces more robust to noise models.

All these characteristics imply an important improvement of the RF ensemble method. The experimental results of the RCRF algorithm will show this.

### 4.1 Diversity of the trees

An important property for the classical RF algorithm is to have diversity in the elements of the forest. In RF, this diversity is obtained with the Bagging of instances used as input for each tree, the random selection of variables for each node and the absence of post-pruning process when the tree building is finished. In the RCRF algorithm, this diversity is increased via the random procedure to select the $s$ parameter. The RCRT base classifier is constructed selecting the value of the parameter $s$ via a random procedure for each tree. This characteristic provides a new element to add diversity in the trees of the forest.

In the following proposition we will see that increasing the value of $s$ we have greater probability intervals that imply a greater value of the measure of information of the IIG criterion (the maximum entropy measure). This is the reason that explain why the choice of a feature can change when the parameter $s$ is increased:

**Proposition 1.** *Let $\mathcal{D}$ a data set of size $N$. Suppose that $C$ is the class variable and its possible values are $\{c_1, \ldots, c_k\}$. Given a specific $s$, consider the following convex set of probability distributions (credal set):*

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

*where* $n_{c_j}$ *is the number of instances in* $\mathcal{D}$ *whose class value is* $c_j$, $\forall j = 1, \ldots, k$.

*Then* $s_1 > s_2 \Rightarrow H^*(K_{s_1}^{\mathcal{D}}(C)) \geq H^*(K_{s_2}^{\mathcal{D}}(C))$.

*Proof.*

Let $1 \leq i \leq k$ and $s_1 > s_2$. It is easy to prove that $\frac{n_{c_i}}{N+s_1} < \frac{n_{c_i}}{N+s_2}$ :

$$\frac{n_{c_i}}{N+s_1} < \frac{n_{c_i}}{N+s_2} \Leftrightarrow \frac{1}{N+s_1} < \frac{1}{N+s_2} \Leftrightarrow N+s_2 < N+s_1 \Leftrightarrow s_2 < s_1.$$

In the same way, it is possible to prove that $\frac{n_{c_i+s_1}}{N+s_1} > \frac{n_{c_i}+s_1}{N+s_2}$, because

$$\frac{n_{c_i+s_1}}{N+s_1} > \frac{n_{c_i}+s_1}{N+s_2} \Leftrightarrow (N+s_2)(n_{c_i}+s_1) > (N+s_1)(n_{c_i}+s_2) \Leftrightarrow$$

$$\Leftrightarrow Ns_1 + s_2 n_{c_i} > Ns_2 + s_1 n_{c_i} \Leftrightarrow N(s_1 - s_2) > n_{c_i}(s_1 - s_2).$$

Since $s_1 > s_2$ the above inequality is fulfilled if and only if $N > n_{c_i}$ and we know that it is right. In consequence, we have that

$$\left[\frac{n_{c_i}}{N+s_1}, \frac{n_{c_i}+s_1}{N+s_1}\right] \subset \left[\frac{n_{c_i}}{N+s_1}, \frac{n_{c_i}+s_1}{N+s_1}\right] \quad \forall i = 1, \ldots, k$$

and this fact implies that

$$K_{s_2}^{\mathcal{D}}(C) \subset K_{s_1}^{\mathcal{D}}(C).$$

Hence,

$$H^*(K_{s_2}^{\mathcal{D}}(C)) \leq H^*(K_{s_1}^{\mathcal{D}}(C)).$$

□

Using the above proposition, we can prove mathematically that a change in the value of the parameter can change the feature selected, via the following proposition.

**Proposition 2.** *A change in the value of the s parameter can produce different feature selected by the IIG split criterion*

*Proof.* Suppose that we have two features $X_1$ and $X_2$ whose possible values are, respectively, $\{x_1^1, \dots, x_{t_1}^1\}$ and $\{x_1^2, \dots, x_{t_2}^2\}$ and that, for a given parameter $s_2$,

$$IIG_{s_2}^{\mathcal{D}}(C, X_1) = H^*(K_{s_2}^{\mathcal{D}}(C)) - H^*(K_{s_2}^{\mathcal{D}}(C|X_1)) > H^*(K_{s_2}^{\mathcal{D}}(C)) - H^*(K_{s_2}^{\mathcal{D}}(C|X_2))$$

$$= IIG_{s_2}^{\mathcal{D}}(C, X_2)$$

If $s_1$ is another parameter of the IDM and $s_1 > s_2$, according with Proposition 1,

$$H^*(K_{s_1}^{\mathcal{D}}(C)) > H^*(K_{s_2}^{\mathcal{D}}(C)),$$

but also

$$H^*(K_{s_1}^{\mathcal{D}}(C|X_j)) > H^*(K_{s_2}^{\mathcal{D}}(C|X_j)), j = 1, 2.$$

The IIG measure for each one of these features with this new parameter is

$$IIG_{s_1}^{\mathcal{D}}(C, X_i) =$$

$$= H^*(K_{s_1}^{\mathcal{D}}(C)) - H^*(K_{s_1}^{\mathcal{D}}(C|X_i)), i = 1, 2$$

Hence:

$$IIG_{s_1}^{\mathcal{D}}(C, X_i) - IIG_{s_2}^{\mathcal{D}}(C, X_i) =$$

$$= H^*(K_{s_1}^{\mathcal{D}}(C)) - H^*(K_{s_2}^{\mathcal{D}}(C)) - (H^*(K_{s_1}^{\mathcal{D}}(C|X_i)) - H^*(K_{s_2}^{\mathcal{D}}(C|X_i))), i = 1, 2$$

The difference $H^*(K_{s_1}^{\mathcal{D}}(C)) - H^*(K_{s_2}^{\mathcal{D}}(C))$ is the same for $X_1$ and $X_2$, unlike $H^*(K_{s_1}^{\mathcal{D}}(C|X_i)) - H^*(K_{s_2}^{\mathcal{D}}(C|X_i))$, which depends on the partitions generated by $X_i$, for $i = 1, 2$.

For this reason, it is possible that $IIG_{s_2}^{\mathcal{D}}(C, X_1) > IIG_{s_2}^{\mathcal{D}}(C, X_2)$ although $IIG_{s_1}^{\mathcal{D}}(C, X_1) < IIG_{s_1}^{\mathcal{D}}(C, X_2)$.

Then, a change in the parameter $s$ can change the choice of the split feature in the tree. □

The fact proven in Proposition 2 is shown with the Example 3. In this example, a toy data set of binary classification is used. It will be seen that three distinct features for splitting the data set are selected in a node by using three different values for $s$ (0, 1 and 2).

*Example 3.* 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 $\{c_1 : 5,\ c_2 : 10\}$. In this node, we also consider that we have only 3 attribute variables $X_1, X_2, X_3$, with possible values $X_1 \in \{x_1^1, x_2^1\}$, $X_2 \in \{x_1^2, x_2^2\}$ and $X_3 \in \{x_1^3, x_2^3\}$. The frequencies of each combination of states in the node $J$ are the following ones:

$$
\begin{array}{l}
X_1 = x_1^1 \rightarrow (4 \text{ of class } c_1,\ 10 \text{ of class } c_2)\\
X_1 = x_2^1 \rightarrow\ (1 \text{ of class } c_1,\ 0 \text{ of class } c_2)\\
X_2 = x_1^2 \rightarrow\ (4 \text{ of class } c_1,\ 4 \text{ of class } c_2)\\
X_2 = x_2^2 \rightarrow\ (1 \text{ of class } c_1,\ 6 \text{ of class } c_2)\\
X_3 = x_1^3 \rightarrow\ (3 \text{ of class } c_1,\ 9 \text{ of class } c_2)\\
X_3 = x_2^3 \rightarrow\ (2 \text{ of class } c_1,\ 1 \text{ of class } c_2)
\end{array}
$$

Considering the IIG criterion of (7), the following values of information gain are obtained for each variable $X_1, X_2, X_3$, and for values of $s = 0, 1$ and 2 (for $s = 0$ the IIG criterion is equivalent to the IG criterion):

$$IIG_{s=0}(C, X_1) = 0.9183 - (0.8056 + 0.0000) = 0.9183 - 0.8056 = 0.1127$$
$$IIG_{s=0}(C, X_2) = 0.9183 - (0.5333 + 0.2761) = 0.9183 - 0.8094 = 0.1089$$
$$IIG_{s=0}(C, X_3) = 0.9183 - (0.6490 + 0.1837) = 0.9183 - 0.8327 = 0.0856$$

$$IIG_{s=1}(C, X_1) = 0.9544 - (0.8570 + 0.0667) = 0.9544 - 0.9237 = 0.0307$$
$$IIG_{s=1}(C, X_2) = 0.9544 - (0.5333 + 0.3786) = 0.9544 - 0.9119 = 0.0425$$
$$IIG_{s=1}(C, X_3) = 0.9544 - (0.7124 + 0.2000) = 0.9544 - 0.9124 = 0.0420$$

$$IIG_{s=2}(C, X_1) = 0.9774 - (0.8908 + 0.0667) = 0.9774 - 0.9575 = 0.0199$$
$$IIG_{s=2}(C, X_2) = 0.9774 - (0.5333 + 0.4286) = 0.9774 - 0.9619 = 0.0155$$
$$IIG_{s=2}(C, X_3) = 0.9774 - (0.7522 + 0.2000) = 0.9774 - 0.9522 = 0.0252$$

From the previous values, it can be seen that the selected feature in the node $J$ is different if we consider different values for the $s$ parameter. In all cases, the feature with greater gain of information is selected:

- The feature $X_1$ is inserted in the node $J$ when $s = 0$ is used.
- The feature $X_2$ is inserted in the node $J$ when $s = 1$ is used.
- The feature $X_3$ is inserted in the node $J$ when $s = 2$ is used.

Deepening in the results of Example 3, with $s = 1$ the chosen variable for split is $X_2$, whereas with $s = 2$ the attribute for split is $X_3$. Here we have the following situations:

- $H^*(K^{\mathcal{D}}_{s=1}(C|X_2)) = \frac{8}{15}H^*(K^{\mathcal{D}}_{s=1}(C|X_2 = x^2_1)) + \frac{7}{15}H^*(K^{\mathcal{D}}_{s=1}(C|X_2 = x^2_2)) = \frac{8}{15} \times 1 + \frac{7}{15} \times 0.8113 = 0.5333 + 0.3786 = 0.9119$.

- $H^*(K^{\mathcal{D}}_{s=2}(C|X_2)) = \frac{8}{15}H^*(K^{\mathcal{D}}_{s=2}(C|X_2 = x^2_1)) + \frac{7}{15}H^*(K^{\mathcal{D}}_{s=2}(C|X_2 = x^2_2)) = \frac{8}{15} \times 1 + \frac{7}{15} \times 0.9182958 = 0.5333 + 0.4285 = 0.9619$.

- $H^*(K^{\mathcal{D}}_{s=1}(C|X_3)) = \frac{12}{15}H^*(K^{\mathcal{D}}_{s=1}(C|X_3 = x^3_1)) + \frac{3}{15}H^*(K^{\mathcal{D}}_{s=1}(C|X_3 = x^3_2)) = \frac{12}{15} \times 0.8904916 + \frac{3}{15} \times 1 = 0.7124 + 0.2 = 0.9124$.

- $H^*(K^{\mathcal{D}}_{s=2}(C|X_3)) = \frac{12}{15}H^*(K^{\mathcal{D}}_{s=2}(C|X_3 = x^3_1)) + \frac{3}{15}H^*(K^{\mathcal{D}}_{s=2}(C|X_3 = x^3_2)) = \frac{12}{15} \times 0.940286 + \frac{3}{15} \times 1 = 0.7522 + 0.2 = 0.9522$.

It can be observed that the maximum of entropy of the partitions generated by $X_2 = x^2_1$ and $X_3 = x^3_2$ do not change when the $s$ value passes from 1 to 2 because in both cases, the maximum of entropy reaches its maximum value when $s = 1$. However, the change of the maximum of entropy of the partitions generated by $X_2 = x^2_2$ and $X_3 = x^3_1$ are notable, being more significative in the first partition. According with Eq. (5), the credal sets are smaller as long as $N$ is bigger and thus, it is easy to check that, in general, the smaller is the sample size the more notable is the difference in the maximum of entropy when the $s$ value increases. The size of the partition generated by $X_3 = x^3_1$ is bigger than the size of the partition generated by $X_2 = x^2_1$. This is the reason why when the $s$ value is incremented from 1 to 2 the increment of $H^*(K^{\mathcal{D}}_s(C|X = X_i))$ is higher for $i = 2$ than for $i = 3$ and, therefore, $IIG_{s=2}(C, X)$ has a higher value for $X = X_3$ than for $X = X_2$, although $IIG_{s=1}(C, X)$ has a higher value for $X = X_2$.

From the above example, propositions and reasonings, it can be seen that if the values of $s$ are modified when RCRTs are built, then different trees can be obtained according the features selected in each node. It can be concluded that the random selection for the value $s$ increases the diversity of the elements of a forest when the RCRF algorithm is used.

## 5 Experiments

To compare the results of the new method, presented here, with the ones of the original RF and other ensemble classifiers with excellent performing under label noise, we have carried out a series of experiments on 100 well-known data sets in the field of machine learning, obtained from the *UCI repository of machine learning* (Lichman, 2013). 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 1 gives a brief description of the characteristics of the data sets used.

Two experimental studies have been carried out. In the first one, the RCRT base classifier of RCRF is checked. In this experiment, the value of the parameter $s$ is fixed instead of having a random value. This RCRT procedure with a fix

**Table 1.** 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.

| Dataset | N | Feat | Num | Nom | k | Range | Dataset | N | Feat | Num | Nom | k | Range |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acute-inflamm-nephritis | 120 | 6 | 1 | 5 | 2 | 2 | mol-splice-junction | 3190 | 60 | 0 | 60 | 3 | 4-5 |
| acute-inflamm-urinary | 120 | 6 | 1 | 5 | 2 | 2 | monks1 | 556 | 6 | 0 | 6 | 2 | 2-4 |
| anneal | 898 | 38 | 6 | 32 | 6 | 2-10 | monks2 | 601 | 6 | 0 | 6 | 2 | 2-4 |
| appendicitis | 106 | 7 | 7 | 0 | 2 | - | monks3 | 554 | 6 | 0 | 6 | 2 | 2-4 |
| arrhythmia | 452 | 279 | 206 | 73 | 16 | 2 | mushroom | 8124 | 22 | 0 | 22 | 2 | 1-10 |
| audiology | 226 | 69 | 0 | 69 | 24 | 2-6 | nursery | 12960 | 8 | 0 | 8 | 4 | 2-4 |
| autos | 205 | 25 | 15 | 10 | 7 | 2-22 | optdigits | 5620 | 64 | 64 | 0 | 10 | - |
| balance-scale | 625 | 4 | 4 | 0 | 3 | - | page-blocks | 5473 | 10 | 10 | 0 | 5 | - |
| bank-marketing | 4521 | 16 | 7 | 9 | 2 | 2-12 | parkinsons | 195 | 22 | 22 | 0 | 2 | - |
| banknote-authe | 1372 | 4 | 4 | 0 | 2 | - | pendigits | 10992 | 16 | 16 | 0 | 10 | - |
| blogger | 100 | 5 | 0 | 4 | 2 | 2-5 | phoneme | 5404 | 5 | 5 | 0 | 2 | - |
| breast-cancer | 286 | 9 | 0 | 9 | 2 | 2-13 | pima-diabetes | 768 | 8 | 8 | 0 | 2 | - |
| bridges-version1 | 107 | 11 | 3 | 8 | 6 | 2-54 | postoperative-patient | 90 | 8 | 8 | 0 | 3 | 2-4 |
| bridges-version2 | 107 | 11 | 0 | 11 | 6 | 2-54 | primary-tumor | 339 | 17 | 0 | 17 | 21 | 2-3 |
| bupa | 345 | 6 | 6 | 9 | 2 | - | qsar-biodegradation | 1055 | 41 | 41 | 0 | 2 | - |
| car | 1728 | 6 | 0 | 6 | 4 | 3-4 | qualitative-bankruptcy | 250 | 6 | 0 | 6 | 2 | 3 |
| cleveland-heart-dis | 303 | 13 | 6 | 7 | 5 | 2-14 | robot-failure-lp1 | 88 | 90 | 90 | 0 | 4 | - |
| cmc | 1473 | 9 | 2 | 7 | 3 | 2-4 | robot-failure-lp2 | 47 | 90 | 90 | 0 | 5 | - |
| credit-rating | 690 | 15 | 6 | 9 | 2 | 2-14 | robot-failure-lp3 | 47 | 90 | 90 | 0 | 4 | - |
| crx | 690 | 15 | 6 | 9 | 2 | 2-14 | robot-failure-lp4 | 117 | 90 | 90 | 0 | 3 | - |
| cylinder-bands | 540 | 39 | 18 | 21 | 2 | 2-429 | robot-failure-lp5 | 164 | 90 | 90 | 0 | 5 | - |
| dermatology | 366 | 34 | 1 | 33 | 6 | 2-4 | saheart | 462 | 9 | 8 | 1 | 2 | 2 |
| dresses-sales | 500 | 12 | 1 | 11 | 2 | 5-25 | seeds | 210 | 7 | 7 | 0 | 3 | - |
| ecoli | 366 | 7 | 7 | 0 | 7 | - | segment | 2310 | 19 | 16 | 0 | 7 | - |
| fertility-diagnosis | 100 | 9 | 9 | 0 | 2 | - | seismic-bumps | 2584 | 18 | 14 | 4 | 2 | 2-3 |
| flags | 194 | 29 | 2 | 27 | 8 | 4-194 | sick | 3772 | 29 | 7 | 22 | 2 | 2 |
| german-credit | 1000 | 20 | 7 | 13 | 2 | 2-11 | solar-flare2 | 1066 | 12 | 0 | 6 | 3 | 2-8 |
| glass | 214 | 9 | 9 | 0 | 7 | - | sonar | 208 | 60 | 60 | 0 | 2 | - |
| glioma16 | 50 | 16 | 16 | 0 | 2 | - | soybean | 683 | 35 | 0 | 35 | 19 | 2-7 |
| haberman | 306 | 3 | 2 | 1 | 2 | 12 | spambase | 4601 | 57 | 57 | 0 | 2 | - |
| hayes-roth | 160 | 4 | 4 | 0 | 4 | - | spect | 267 | 22 | 0 | 22 | 2 | 2 |
| heart-statlog | 270 | 13 | 13 | 0 | 2 | - | spectf | 349 | 44 | 44 | 0 | 2 | - |
| hepatitis | 155 | 19 | 4 | 15 | 2 | 2 | spectrometer | 531 | 101 | 100 | 1 | 48 | 4 |
| horse-colic | 368 | 22 | 7 | 15 | 2 | 2-6 | splice | 3190 | 60 | 0 | 60 | 3 | 4-6 |
| hungarian-heart-dis | 294 | 13 | 6 | 7 | 5 | 2-14 | sponge | 76 | 44 | 0 | 44 | 3 | 2-9 |
| hypothyroid | 3772 | 30 | 7 | 23 | 4 | 2-4 | synthetic-control | 600 | 61 | 61 | 0 | 6 | - |
| ionosphere | 351 | 35 | 35 | 0 | 2 | - | tae | 151 | 5 | 3 | 2 | 3 | 2 |
| iris | 150 | 4 | 4 | 0 | 3 | - | teaching-assistant-eval | 151 | 5 | 3 | 2 | 3 | 2 |
| kr-vs-kp | 3196 | 36 | 0 | 36 | 2 | 2-3 | thoracic-surgery | 470 | 16 | 3 | 13 | 2 | 2-7 |
| labor | 57 | 16 | 8 | 8 | 2 | 2-3 | tic-tac-toe | 958 | 9 | 0 | 9 | 2 | 3 |
| leaf | 340 | 15 | 15 | 0 | 30 | - | trains | 10 | 32 | 0 | 32 | 2 | 1-8 |
| letter | 20000 | 16 | 16 | 0 | 26 | - | turkiye-student | 5820 | 32 | 32 | 0 | 13 | - |
| leukemia-haslinger | 100 | 50 | 50 | 0 | 2 | - | user-knowledge | 403 | 5 | 5 | 0 | 5 | - |
| liver-disorders | 345 | 6 | 6 | 0 | 2 | - | vehicle | 946 | 18 | 18 | 0 | 4 | - |
| lsvt-voice-rehab | 126 | 310 | 310 | 0 | 2 | - | vote | 435 | 16 | 0 | 16 | 2 | 2 |
| lymphography | 146 | 18 | 3 | 15 | 4 | 2-8 | vowel | 990 | 11 | 10 | 1 | 11 | 2 |
| mfeat-morphological | 2000 | 6 | 6 | 0 | 10 | - | waveform | 5000 | 40 | 40 | 0 | 3 | - |
| mfeat-pixel | 2000 | 240 | 0 | 240 | 10 | 4-6 | wine | 178 | 13 | 13 | 0 | 3 | - |
| mol-biology-promoters | 106 | 57 | 0 | 57 | 2 | 4 | wisconsin-breast-cancer | 699 | 9 | 9 | 0 | 2 | - |
| mol-promotor-gene | 106 | 57 | 0 | 57 | 2 | 4 | zoo | 101 | 16 | 1 | 16 | 7 | 2 |

value of $s$ will be called *Credal Random Tree* (CRT). CRT has been performed with different values of $s$. The aim of this study is established a good range to be used in the RCRT procedure, that is, we look for a a good interval to use for the random process to select the value of $s$ in the base classifier RCRT. The average results about accuracy for CRT with different values of $s$ will be described later on.

In the second study, RCRF algorithm is compared with the original RF algorithm and bagging schemes of other two based models: C4.5 (Dietterich, 2000b) and CDT (Abellán & Masegosa, 2012). The motive of the use of bagging schemes is that this scheme has shown the best performance in noise domains in the literature. We also use as reference the algorithm that we call *Credal Random Forest* (CRF) that is the RCRF ensemble method using the base classifier CRT with $s = 1$. This value for $s$ is the standard one used in the previous papers about credal trees (Abellán & Moral, 2005; Mantas & Abellán, 2014b), motivated principally by computational reasons (Mantas & Abellán, 2014b) and by its origin (Walley, 1996). All the trees of the previous ensemble methods are used without a pruning process in order to keep the same experimental conditions for all the algorithms to be compared. Resuming, 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)
- Random Credal Random Forest (RCRF)

In the two studies, the algorithms are compared using the original data sets obtained from the UCI repository, adding different percentages of random label noise only in the training set.

The *Weka* software (Witten & Frank, 2005) has been used for the experimentation. The methods RCRF, CRF, RCRT and CRT were implemented using data structures of *Weka*. We added the necessary methods to the implementation of the algorithms RF and RT provided by *Weka* software to design RCRF, CRF, RCRT and CRT 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 RCRF, 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 researches, such as (Freund & Schapire, 1996).

Using *Weka's* filters, the following percentages of random noise to the class variable: $0\%, 5\%, 10\%, 20\%$ and $30\%$, have been only added 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. To compare the results of all the classifiers, 10 times a 10-fold cross validation procedure was repeated for each data set.

### 5.1 Results

Table 2 shows the results obtained in the first study. It presents the average results of accuracy of the CRT algorithm for each added noise level with the following values for the parameter $s = 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$ and $4.0$. These values are similar to those used for comparing credal trees with different values of $s$ in (Mantas et al., 2016) and (Abellán, Mantas, & Castellano, 2018). In this table, the best algorithm for each noise level is emphasized using bold fonts, the second best is marked with italic fonts. It can be observed that CRT algorithm with the values for $s$ close to 0 do not obtain good results in any noise level. For data sets without noise or with low level noise (5%), it can be seen that the values for $s$ close to 1.0 are the bests. On the other hand, when the noise level is increased, high values for $s$ are the best choice. Based on the different values of $s$ studied in (Mantas et al., 2016; Abellán et al., 2018) and the results obtained in this analysis, it can be concluded that a range of values for $s$ equal to $\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5\}$ could be useful in the RCRF algorithm.[8] The aim is to achieve a new algorithm with acceptable results without having to adjust parameters in terms of the noise level of the data sets.

**Table 2.** Average accuracy results of the CRT algorithm with different values of $s$ when data sets with added noise are classified.

| Algorithm | noise 0% | noise 5% | noise 10% | noise 20% | noise 30% |
|---|---|---|---|---|---|
| RT ($CRT_{s=0}$) | 76.75 | 73.64 | 70.62 | 64.08 | 58.24 |
| $CRT_{s=0.5}$ | 77.47 | 74.75 | 72.13 | 65.48 | 59.23 |
| $CRT_{s=1.0}$ | *78.10* | 76.00 | 73.74 | 67.67 | 61.39 |
| $CRT_{s=1.5}$ | **78.16** | 76.51 | 74.89 | 69.45 | 63.29 |
| $CRT_{s=2.0}$ | 78.09 | **76.65** | 75.23 | 70.57 | 65.12 |
| $CRT_{s=2.5}$ | 78.03 | **76.65** | 75.40 | 71.35 | 66.21 |
| $CRT_{s=3.0}$ | 77.75 | *76.58* | *75.41* | 71.70 | 67.09 |
| $CRT_{s=3.5}$ | 77.73 | 76.54 | 75.34 | *71.81* | *67.64* |
| $CRT_{s=4.0}$ | 77.67 | 76.37 | **75.56** | **71.91** | **67.94** |

With the interval of values for randomly selecting the parameter $s$ obtained in the previous paragraph, the RCRF algorithm was used to classify the data

[8] We can observe in Table 2 that for $s = 4$ we can obtain good results for high level of label noise, but we have checked that they are very similar than the ones obtained with $s = 3.5$. This is the reason to consider that set of values for $s$. It is a shorter and more compact set than the one considered in (Mantas et al., 2016; Abellán et al., 2018).

sets. Table 3 presents the average accuracy results of the methods used in the second study: BA-C4.5, BA-CDT, RF, CRF and RCRF. In this table, the best algorithm for each added noise level is emphasized using bold fonts, the second best is marked with italic fonts. Tables that present the detailed accuracy results of the ensemble methods obtained in the second study, when they classify data sets with different levels of label noise, are described in Appendix A.

**Table 3.** Average accuracy results of the ensemble methods when they are built from data sets with added noise.

| Algorithm | noise 0% | noise 5% | noise 10% | noise 20% | noise 30% |
|---|---|---|---|---|---|
| BA-C4.5 | 82.84 | 82.15 | 81.02 | 77.82 | 72.76 |
| BA-CDT | 82.34 | 81.81 | 81.23 | 77.81 | 73.98 |
| RF | 84.01 | 82.95 | 81.86 | 77.76 | 72.60 |
| CRF | *84.89* | *84.04* | *83.10* | *79.58* | *74.64* |
| RCRF | **84.98** | **84.35** | **83.73** | **80.93** | **76.71** |

Following the recommendation of (Demšar, 2006), a series of tests have been used in order to compare the ensemble methods of the second study using the *Keel* software (Alcalá-Fdez et al., 2009). The following tests to compare multiple classifiers on multiple data sets have been utilized:

> **Friedman test** (Friedman, 1937, 1940): 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, all the algorithms can be compared to each other using the **Nemenyi test** (Nemenyi, 1963).

All the tests were carried out with a level of significance of $\alpha = 0.05$. Hence, Table 4 show Friedman's ranks about the accuracy of the methods when they are applied on data sets with different levels of added noise. The best algorithm for each noise level is emphasized using bold fonts, the second best is marked with italic fonts. Tables 11, 12, 13, 14 and 15 in the Appendix A, 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[9] if the corresponding p-value is $\leq 0.005$. When there is a significative difference, the best algorithm is distinguished with bold fonts.

For the sake of simplicity, the results of the Nemenyi's test about the pairwise comparisons can be seen graphically in Figure 5. Here the critical difference is expressed as a vertical segment and the columns express the values of the Friedman's ranks. When the high of the correspondent segment is lower than

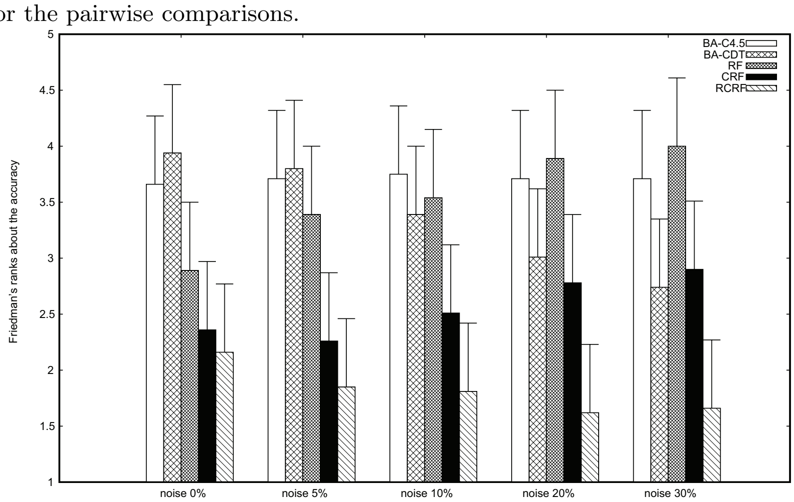
[9] In this case, the *critical difference* used on the Friedman's ranks is 0.610 (see (Demšar, 2006)).

the high of the columns, the differences are statistically significant in favor of the algorithm represented in the lower column.

**Table 4.** Friedman's ranks about the accuracy of the ensemble methods when they are applied on data sets with different percentages of added noise.

| Algorithm | noise 0% | noise 5% | noise 10% | noise 20% | noise 30% |
|---|---|---|---|---|---|
| BA-C4.5 | 3.66 | 3.71 | 3.75 | 3.71 | 3.71 |
| BA-CDT | 3.94 | 3.80 | 3.39 | 3.01 | 2.74 |
| RF | 2.89 | 3.39 | 3.54 | 3.89 | 4.00 |
| CRF | *2.36* | *2.26* | *2.51* | *2.78* | *2.90* |
| RCRF | **2.16** | **1.85** | **1.81** | **1.62** | **1.66** |

**Fig. 5.** Values of the Friedman's rank of the methods. The segment on the top expresses the size of the critical difference associated with the experiments and the Nemenyi's test for the pairwise comparisons.




To extend the comparison of the ensemble algorithms when the methods are applied on data sets with label noise, a recent measure to quantify the degree of robustness of a classifier under noise have been used. The measure is the *Equalized Loss of Accuracy* (ELA) of (Sáez, Luengo, & Herrera, 2014), 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 characterise the behavior of a method with noisy data considering performance and robustness. $ELA$ measure is expressed as follows:

$$ELA_{x\%} = \frac{100 - A_{x\%}}{A_{0\%}} \tag{8}$$

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.

In Table 5, it can be seen the average results of the $ELA$ measure for each ensemble method of the second study.

**Table 5.** 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% | noise 30% |
|---|---|---|---|---|
| BA-C4.5 | 0.2155 | 0.2291 | 0.2677 | 0.3288 |
| BA-CDT | 0.2209 | 0.2280 | 0.2695 | 0.3160 |
| RF | 0.2030 | 0.2159 | 0.2647 | 0.3262 |
| CRF | *0.1880* | *0.1991* | *0.2405* | *0.2987* |
| RCRF | **0.1842** | **0.1915** | **0.2244** | **0.2741** |

## 6 Analysis of results

From the results above presented, the following points can be exposed taking into account the general and particular statistical comparatives (Friedman and Nemenyi's test):

- In general, RCRF is the algorithm with the best results on data sets with and without label noise. The Friedman's ranks always show that the best results are obtained with this procedure. RCRF has always significantly better results than RF.
- RCRF, RF and CRF are the best algorithms when they classify data sets without noise according to the tests carried out, being RCRF significantly better accuracy than RF. But the statistical differences are not significative among RCRF and CRF or CRF and RF. In this situation the worse methods are the BA-CDT and BA-C4.5. When these two methods are compared with the rest via the Nemenyi's test, only RF has no significant differences with BA-C4.5, the rest of comparisons shown statistical significant differences. Hence, RF can be considered the third best in the comparative study.
- With low level of label noise (5%), RCRF and CRF have positive statistical significant differences with respect to the rest. Here RF has no differences with the worse methods: again BA-CDT and BA-C4.5. As occurs with no noise, RCRF and CRF have not significant differences.

- With medium level of noise (10% and 20%) we can observe important changes. With 10% RCRF wins to all the methods (via Nemenyi's test). Here, RF is the second worse procedure. With 20%, RCRF has also the best possible results: it wins statistically, via the Nemenyi's test, to all the rest of procedures. Here, RF has a bad performance being the worse method.
- With the highest level of noise (30%), RF is the worse method with important differences with the rest. Here, RCRF wins to all the rest of methods and CRF only wins to RF and BA-C4.5. Now, BA-CDT has good results being the winner in the statistical comparative with RF and BA-C4.5. BA-C4.5 is the second worse procedure.
- According to the ELA, RCRF algorithm is the most robust classifier for all noise levels. The second one is the CRF algorithm. On the other hand, RF is more robust to noise than the bagging schemes for levels of noise equal to 5%, 10% and 20%, and RF is even most robust than BA-C4.5 with the greatest level of noise (30%).

From the previous comments, the experiments show that it can be concluded that RCRF is a good classification algorithm for data sets with or without noise. The overfitting problem of the RF algorithm is avoided with the new algorithm presented in this paper.

## 7 Conclusions

In this paper the scheme of the Random Forest method has been modified using a new split criterion based on imprecise probabilities, called Imprecise Info-Gain. The performance of this new split criterion depends of a parameter $s$. The value of $s$ has been also randomly selected for each tree of the forest. In this way, the diversity of the trees in the RF algorithm is increased without diminishing the accuracy of the ensemble members. This is a good property for improving the classification accuracy of the RF ensemble. These modifications of RF represents a new method of classification with important characteristics that imply some advantages with respect to the classic RF algorithm: (i) the use of a more successful split criteria; (ii) an increasing of the randomness to obtain more diversity in the forest; and (iii) the application of imprecise probabilities that imply a more robust to noise model.

It has been shown, via an experimental study on a large set of data sets, that the new procedure improves significatively the original RF. Besides, when noisy data sets are classified, this improvement increases and it is also statistically significant. Classic bagging ensembles, very successful models on noise domains, are also compared with the new procedure in an experimental study. The new procedure achieves better results than the ones of the rest of method used here as reference. All these assertions have been reinforced via appropriate statistical tests.

Hence, a new method of supervised classification has been presented: Random Credal Random Forest. This model solves the problem of overfitting that presented the RF method when noisy data sets are classified. The new classifier represents a very powerful tool to be applied on data sets without worrying about the noise level of the data. In all the grounds where RF has a good performance, the new classifier can be a better alternative.

**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.

## References

Abellán, J. (2006). Uncertainty measures on probability intervals from the imprecise dirichlet model. *International Journal of General Systems*, *35*(5), 509-528. doi: 10.1080/03081070600687643

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

Abellán, J., Klir, G., & Moral, S. (2006). Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, *35*(1), 29-44. doi: 10.1080/03081070500473490

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *41*(8), 3825 - 3830. doi: 10.1016/j.eswa.2013.12.003

Abellán, J., Mantas, C. J., & Castellano, J. G. (2018). AdaptativeCC4.5: Credal C4.5 with a rough class noise estimator. *Expert Systems with Applications*, *92*(Supplement C), 363 - 379. doi: 10.1016/j.eswa.2017.09.057

Abellán, J., & Masegosa, A. (2009). A filter-wrapper method to select variables for the naive bayes classifier based on credal decision trees. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *17*(06), 833-854. doi: 10.1142/S0218488509006297

Abellán, J., & Masegosa, A. R. (2010). An ensemble method using credal decision trees. *European journal of operational research*, *205*(1), 218–226.

Abellán, J., & Masegosa, A. R. (2012). Bagging schemes on the presence of class noise in classification. *Expert Systems with Applications*, *39*(8), 6827 - 6837. doi: 10.1016/j.eswa.2012.01.013

Abellán, J., & Moral, S. (2003). Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, *18*(12), 1215–1225. doi: 10.1002/int.10143

Abellán, J., & Moral, S. (2005). Upper entropy of credal sets. applications to credal classification. *International Journal of Approximate Reasoning*, *39*(2–3), 235 - 255. (Imprecise Probabilities and Their Applications) doi: 10.1016/j.ijar.2004.10.001

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., ... Herrera, F. (2009). Keel: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307-318. doi: 10.1007/s00500-008-0323-y

Azqhandi, M. A., Ghaedi, M., Yousefi, F., & Jamshidi, M. (2017). Application of random forest, radial basis function neural networks and central composite design for modeling and/or optimization of the ultrasonic assisted adsorption of brilliant green on zns-np-ac. *Journal of Colloid and Interface Science*(in press), -. doi: 10.1016/j.jcis.2017.05.098

Bai, S. (2017, 4). Growing random forest on deep convolutional neural networks for scene categorization. *Expert Systems with Applications*, *71*, 279–287.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi: 10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, California, U.S.A.: Wadsworth Publishing Company.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Journal of Information Fusion*, *6*, 5–20.

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

Dietterich, T. G. (2000a). Ensemble methods in machine learning. In *Proceedings of the first international workshop on multiple classifier systems* (pp. 1–15). London, UK, UK: Springer-Verlag.

Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. doi: 10.1023/A:1007607513941

Frenay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845-869. doi: 10.1109/TNNLS.2013.2292894

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta (Ed.), *Proceedings of the thirteenth international conference on machine learning (icml 1996)* (p. 148-156). Morgan Kaufmann.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675-701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of $m$ rankings. *The Annals of Mathematical Statistics*, *11*(1), 86–92. doi: 10.1214/aoms/1177731944

Gauba, H., Kumar, P., Roy, P. P., Singh, P., Dogra, D. P., & Raman, B. (2017). Prediction of advertisement preference by fusing {EEG} response and sentiment analysis. *Neural Networks*, *92*, 77 - 88. (Advances in Cognitive Engineering Using Neural Networks) doi: 10.1016/j.neunet.2017.01.013

Hand, D. (1981). *Discrimination and classification.* John Wiley.

Hand, D. J. (1997). *Construction and assessment of classification rules.* John Wiley and Sons, New York.

Jiang, X., Abdel-Aty, M., Hu, J., & Lee, J. (2016). Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. *Neurocomputing*, *181*, 53 - 63. (Big Data Driven Intelligent Transportation Systems) doi: 10.1016/j.neucom.2015.08.097

Klir, G. J. (2005). *Uncertainty and information: Foundations of generalized information theory.* John Wiley And Sons, Inc. doi: 10.1002/0471755575

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, *259*(2), 689 - 702. doi: 10.1016/j.ejor.2016.10.031

Kulkarni, V. Y., Petare, M., & Sinha, P. K. (2012). Analyzing random forest classifier with different split measures. In *Proceedings of the second international conference on soft computing for problem solving (socpros 2012)* (pp. 691–699). New Delhi: Springer India. doi: 10.1007/978-81-322-1602-5_74

Li, T., Ni, B., Wu, X., Gao, Q., Li, Q., & Sun, D. (2016). On random hyper-class random forest for visual classification. *Neurocomputing*, *172*, 281 - 289. doi: 10.1016/j.neucom.2014.10.101

Lichman, M. (2013). *UCI machine learning repository.* Retrieved from http://archive.ics.uci.edu/ml

Mantas, C. J., & Abellán, J. (2014a). Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data. *Expert Systems with Applications*, *41*(5), 2514 - 2525. doi: 10.1016/j.eswa.2013.09.050

Mantas, C. J., & Abellán, J. (2014b). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, *41*(10), 4625 - 4637. doi: 10.1016/j.eswa.2014.01.017

Mantas, C. J., Abellán, J., & Castellano, J. G. (2016). Analysis of Credal-C4.5 for classification in noisy domains. *Expert Systems with Applications*, *61*, 314 - 326. doi: 10.1016/j.eswa.2016.05.035

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., & Hamprecht, F. A. (2011). On oblique random forests. In *Proceedings of the 2011 european conference on machine learning and knowledge discovery in databases - volume part ii* (pp. 453–469). Berlin, Heidelberg: Springer-Verlag.

Nemenyi, P. (1963). *Distribution-free multiple comparisons* (Doctoral dissertation). Princeton University, New Jersey, USA.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann, San Mateo, CA.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. doi: 10.1023/A:1022643204877

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93. doi: 10.1023/B:AMAI.0000018580.96245.c6

Ren, Y., Zhang, L., & Suganthan, P. N. (2016, Feb). Ensemble classification and regression-recent developments, applications and future directions [review article]. *IEEE Computational Intelligence Magazine*, *11*(1), 41-53. doi: 10.1109/MCI.2015.2471235

Sáez, J. A., Luengo, J., & Herrera, F. (2014). Evaluating the classifier behavior with noisy data considering performance and robustness: the equalized loss of accuracy measure. *Neurocomputing*, *176*, 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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion*, *6*(1), 83 - 98. doi: 10.1016/j.inffus.2004.04.003

Walley, P. (1996). Inferences from multinomial data; learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 3-57. doi: 10.2307/2346164

Wang, S., Yin, Y., Cao, G., Wei, B., Zheng, Y., & Yang, G. (2015). Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing*, *149*, 708 - 717. doi: 10.1016/j.neucom.2014.07.059

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (Second ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Zhang, L., & Suganthan, P. N. (2014). Random forests with ensemble of feature spaces. *Pattern Recognition*, *47*(10), 3429 - 3437. doi: https://doi.org/10.1016/j.patcog.2014.04.001

Zhang, L., & Suganthan, P. N. (2015). Oblique decision tree ensemble via multisurface proximal support vector machine. *IEEE Trans. Cybernetics*, *45*(10), 2165–2176. doi: 10.1109/TCYB.2014.2366468

Zhang, L., & Suganthan, P. N. (2017). Benchmarking ensemble classifiers with novel co-trained kernal ridge regression and random vector functional link ensembles [research frontier]. *IEEE Comp. Int. Mag.*, *12*(4), 61–72. doi: 10.1109/MCI.2017.2742867

## Appendix A Tables about accuracy results

Tables 6, 7, 8, 9 and 10 show the accuracy results obtained by the ensemble methods when they classify data sets with different added noise levels.

Tables 11, 12, 13, 14 and 15 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's procedures rejects the hypotheses which have a corresponding p-value $\leq 0.005$. When there is a significative difference, the best algorithm is distinguished with bold fonts.

**Table 6.** Accuracy results of the ensemble methods when they are used on data sets without added noise.

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| acute-inflamm-nephritis | 100 | 100 | 100 | 100 | 100 |
| acute-inflamm-urinary | 100 | 99.42 | 100 | 100 | 100 |
| anneal | 98.9 | 98.89 | 99.68 | 99.71 | 99.63 |
| appendicitis | 86.50 | 87.15 | 85.88 | 86.80 | 87.71 |
| arrhythmia | 75.35 | 74.49 | 69.12 | 69.93 | 73.14 |
| audiology | 81.83 | 80.41 | 80.36 | 81.28 | 80.92 |
| autos | 85.45 | 80.27 | 84.29 | 85.32 | 85.02 |
| balance-scale | 81.56 | 82.41 | 80.3 | 81.94 | 82.76 |
| bank-marketing | 89.88 | 89.76 | 89.68 | 89.86 | 89.76 |
| banknote-auth | 98.95 | 98.77 | 99.34 | 99.31 | 99.26 |
| blogger | 75.90 | 76.10 | 82.40 | 82.60 | 81.70 |
| breast-cancer | 70.43 | 70.35 | 70.02 | 73.53 | 73.79 |
| bridges-version1 | 64.75 | 63.15 | 54.88 | 68.61 | 70.52 |
| bridges-version2 | 60.71 | 61.15 | 54.55 | 65.95 | 66.41 |
| bupa | 71.54 | 70.46 | 72.03 | 72.60 | 73.51 |
| car | 94.33 | 93.55 | 94.7 | 94.44 | 93.3 |
| cleveland-heart-dis | 80.23 | 78.68 | 81.56 | 81.26 | 81.18 |
| cmc | 52.19 | 53.21 | 50.69 | 52.09 | 53.52 |
| credit-rating | 85.68 | 86.07 | 86.14 | 86.87 | 86.91 |
| crx | 85.71 | 86.13 | 86.14 | 86.87 | 86.91 |
| cylinder-bands | 57.91 | 74.15 | 76.28 | 81.83 | 84.26 |
| dermatology | 97.13 | 94.18 | 96.91 | 97.87 | 97.7 |
| dresses-sales | 55.78 | 60.26 | 56.38 | 59.28 | 60.24 |
| ecoli | 84.88 | 83.82 | 84.67 | 85.27 | 84.91 |
| fertility-diagnosis | 87.80 | 88.00 | 85.30 | 86.50 | 86.90 |
| flags | 58.94 | 58.53 | 61.40 | 63.67 | 63.09 |
| german-credit | 73.01 | 74.64 | 76.08 | 76.38 | 76.42 |
| glass | 74.49 | 75.51 | 79.71 | 78.87 | 77.85 |
| glioma16 | 77.60 | 81.40 | 79.80 | 79.20 | 79.60 |
| haberman | 70.17 | 73.76 | 65.44 | 72.56 | 73.02 |
| hayes-roth | 81.63 | 81.00 | 81.63 | 81.19 | 81.94 |
| heart-statlog | 80.96 | 81.41 | 82.26 | 82 | 82.19 |
| hepatitis | 81.76 | 80.99 | 83.58 | 83.37 | 83.94 |
| horse-colic | 85.51 | 84.91 | 85.59 | 85.18 | 85.19 |
| hungarian-heart-dis | 78.92 | 81.18 | 80.25 | 80.54 | 82.69 |
| hypothyroid | 99.62 | 99.59 | 99.51 | 99.7 | 99.73 |
| ionosphere | 92.57 | 91.23 | 93.48 | 93.65 | 93.74 |
| iris | 94.47 | 95.07 | 94.53 | 94.6 | 94.87 |
| kr-vs-kp | 99.46 | 99.4 | 99.27 | 99.34 | 99.26 |
| labor | 82.60 | 83.87 | 87.10 | 87.53 | 87.87 |
| leaf | 69.94 | 70.35 | 77.24 | 77.41 | 76.09 |
| letter | 94.03 | 92.44 | 96.6 | 96.54 | 96.18 |
| leukemia-haslinger | 80.00 | 78.30 | 85.70 | 85.70 | 85.60 |
| liver-disorders | 73.42 | 72.21 | 72.03 | 72.6 | 73.51 |
| lsvt-voice-rehab | 78.67 | 81.75 | 82.56 | 82.88 | 83.67 |
| lymphography | 79.96 | 76.24 | 83.42 | 82.34 | 81.99 |
| mfeat-morphological | 72.82 | 73.68 | 70.06 | 73.96 | 74.40 |
| mfeat-pixel | 83.86 | 87.2 | 96.37 | 96.65 | 96.61 |
| mol-biology-promoters | 85.02 | 83.61 | 90.81 | 92.85 | 92.86 |
| mol-promotor-gene | 85.11 | 83.58 | 91.77 | 93.00 | 92.55 |

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| mol-splice-junction | 94.51 | 93.70 | 95.85 | 96.39 | 96.43 |
| monks1 | 100 | 97.00 | 100 | 99.95 | 99.85 |
| monks2 | 65.50 | 63.74 | 66.63 | 71.04 | 72.86 |
| monks3 | 98.92 | 98.92 | 97.98 | 98.92 | 98.92 |
| mushroom | 100 | 100 | 100 | 100 | 100 |
| nursery | 98.68 | 96.66 | 99.17 | 96.86 | 96.01 |
| optdigits | 95.84 | 95.55 | 98.3 | 98.38 | 98.4 |
| page-blocks | 97.36 | 97.32 | 97.46 | 97.6 | 97.6 |
| parkinsons | 89.87 | 88.43 | 92.11 | 91.87 | 91.81 |
| pendigits | 98.32 | 98.45 | 99.21 | 99.21 | 99.15 |
| phoneme | 89.11 | 87.93 | 91.40 | 91.25 | 90.87 |
| pima-diabetes | 76.14 | 75.8 | 76.01 | 75.86 | 76.13 |
| postoperative-patient | 68.67 | 70.89 | 61.00 | 70.56 | 70.78 |
| primary-tumor | 44.22 | 43.93 | 43.45 | 44.72 | 44.9 |
| qsar-biodegradation | 86.14 | 85.25 | 87.13 | 87.08 | 87.07 |
| qualitative-bankruptcy | 98.36 | 98.28 | 99.72 | 99.72 | 99.88 |
| robot-failure-lp1 | 77.35 | 78.06 | 86.40 | 85.42 | 86.44 |
| robot-failure-lp2 | 65.15 | 56.20 | 65.95 | 66.90 | 66.85 |
| robot-failure-lp3 | 62.95 | 56.05 | 71.85 | 71.65 | 72.10 |
| robot-failure-lp4 | 86.54 | 84.57 | 91.37 | 90.92 | 91.27 |
| robot-failure-lp5 | 66.00 | 68.65 | 73.27 | 72.93 | 72.90 |
| saheart | 68.61 | 69.82 | 68.02 | 67.83 | 68.31 |
| seeds | 92.71 | 91.19 | 93.57 | 93.57 | 93.71 |
| segment | 97.75 | 97.45 | 98.16 | 98.19 | 98.03 |
| seismic-bumps | 92.82 | 93.39 | 93.07 | 93.06 | 93.06 |
| sick | 98.97 | 98.97 | 98.43 | 98.59 | 98.64 |
| solar-flare2 | 99.49 | 99.53 | 99.43 | 99.53 | 99.53 |
| sonar | 80.07 | 80.78 | 84.63 | 84.55 | 84.97 |
| soybean | 92.28 | 90.47 | 93.31 | 94.95 | 94.6 |
| spambase | 94.73 | 94.65 | 95.68 | 95.57 | 95.44 |
| spect | 82.18 | 83.38 | 81.99 | 83.24 | 83.23 |
| spectf | 89.75 | 83.54 | 91.63 | 91.63 | 91.98 |
| spectrometer | 56.61 | 54.48 | 57.42 | 57.91 | 56.1 |
| splice | 94.7 | 94.4 | 95.88 | 96.31 | 96.48 |
| sponge | 93.91 | 92.63 | 95 | 95 | 95 |
| synthetic-control | 95.67 | 94.27 | 98.22 | 98.80 | 98.95 |
| tae | 60.88 | 60.88 | 68.25 | 67.37 | 64.25 |
| teaching-assistant-eval | 59.03 | 54.73 | 68.25 | 67.37 | 64.25 |
| thoracic-surgery | 84.02 | 85.06 | 83.28 | 83.77 | 84.55 |
| tic-tac-toe | 93.05 | 90.19 | 97.10 | 96.95 | 95.04 |
| trains | 78.00 | 56.00 | 54.00 | 62.00 | 62.00 |
| turkiye-student | 36.38 | 38.63 | 36.51 | 37.87 | 39.00 |
| user-knowledge | 90.33 | 89.98 | 91.31 | 90.79 | 90.60 |
| vehicle | 75.22 | 74.78 | 75.18 | 74.96 | 75.36 |
| vote | 96.78 | 96.34 | 96.43 | 96.55 | 96.59 |
| vowel | 94.04 | 92.17 | 98.16 | 98.22 | 97.62 |
| waveform | 83.4 | 83.51 | 85.2 | 85.15 | 85.01 |
| wine | 95.34 | 95.84 | 97.74 | 97.51 | 97.51 |
| wisconsin-breast-cancer | 96.45 | 96.14 | 96.58 | 96.55 | 96.81 |
| zoo | 92.8 | 92.4 | 96.33 | 96.25 | 96.05 |

**Table 7.** Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 5%.

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| acute-inflamm-nephritis | 99.67 | 99.67 | 96.67 | 99.58 | 99.75 |
| acute-inflamm-urinary | 99.50 | 98.75 | 95.75 | 98.42 | 98.75 |
| anneal | 98.83 | 98.78 | 98.2 | 99.11 | 99.28 |
| appendicitis | 85.06 | 86.34 | 84.37 | 85.18 | 86.49 |
| arrhythmia | 75.05 | 74.27 | 68.56 | 69.16 | 71.68 |
| audiology | 81.32 | 80.36 | 78.82 | 80.52 | 80.56 |
| autos | 83.54 | 78.56 | 80.04 | 82.2 | 82.5 |
| balance-scale | 81.71 | 82.25 | 79.21 | 82.05 | 82.94 |
| bank-marketing | 89.48 | 89.53 | 89.35 | 89.76 | 89.64 |
| banknote-auth | 98.65 | 98.78 | 98.75 | 98.99 | 99.04 |
| blogger | 75.30 | 75.90 | 81.60 | 81.90 | 81.00 |
| breast-cancer | 69.03 | 70.63 | 68.26 | 72.56 | 73.09 |
| bridges-version1 | 62.28 | 57.05 | 52.32 | 62.95 | 69.55 |
| bridges-version2 | 58.92 | 56.46 | 51.83 | 62.78 | 66.46 |
| bupa | 69.93 | 68.98 | 71.24 | 71.30 | 72.08 |
| car | 92.88 | 93.28 | 94.12 | 93.99 | 93.11 |
| cleveland-heart-dis | 79.9 | 79.67 | 80.84 | 80.73 | 80.5 |
| cmc | 51.17 | 52.5 | 49.55 | 50.94 | 52.65 |
| credit-rating | 84.87 | 85.77 | 85.23 | 86.39 | 86.72 |
| crx | 85.97 | 85.78 | 85.23 | 86.39 | 86.72 |
| cylinder-bands | 57.98 | 70.94 | 74.26 | 78.85 | 81.94 |
| dermatology | 96.53 | 94.31 | 96.53 | 97.37 | 97.7 |
| dresses-sales | 55.72 | 58.80 | 55.46 | 58.32 | 58.98 |
| ecoli | 84.05 | 83.93 | 84.25 | 84.64 | 84.56 |
| fertility-diagnosis | 86.00 | 87.90 | 85.30 | 86.20 | 86.60 |
| flags | 56.49 | 58.79 | 60.79 | 62.61 | 62.80 |
| german-credit | 72.81 | 73.96 | 74.96 | 75.91 | 76.15 |
| glass | 74.67 | 75.03 | 77.97 | 77.99 | 76.81 |
| glioma16 | 79.00 | 78.80 | 81.20 | 82.60 | 81.60 |
| haberman | 70.07 | 72.23 | 64.63 | 70.97 | 73.65 |
| hayes-roth | 81.44 | 81.00 | 80.81 | 81.06 | 80.63 |
| heart-statlog | 79.89 | 79.85 | 80.89 | 80.85 | 81.26 |
| hepatitis | 81.12 | 81.76 | 82.88 | 83.25 | 83.26 |
| horse-colic | 85.21 | 84.15 | 85.12 | 84.88 | 84.83 |
| hungarian-heart-dis | 79.23 | 79.9 | 80.37 | 80.28 | 82.28 |
| hypothyroid | 99.53 | 99.55 | 99.35 | 99.62 | 99.63 |
| ionosphere | 92.25 | 91.54 | 92.97 | 93.14 | 92.77 |
| iris | 94.33 | 94.6 | 91.87 | 92.93 | 94 |
| kr-vs-kp | 99.08 | 99.16 | 98.3 | 98.87 | 98.95 |
| labor | 82.93 | 82.83 | 86.93 | 86.63 | 87.27 |
| leaf | 70.44 | 69.71 | 75.41 | 76.03 | 74.88 |
| letter | 93.85 | 92.58 | 95.41 | 96.23 | 95.99 |
| leukemia-haslinger | 79.80 | 79.70 | 84.50 | 85.30 | 85.10 |
| liver-disorders | 72.22 | 71.65 | 71.24 | 71.3 | 72.08 |
| lsvt-voice-rehab | 77.58 | 80.27 | 81.72 | 81.47 | 81.87 |
| lymphography | 79.63 | 77.02 | 83.77 | 82.52 | 81.94 |
| mfeat-morphological | 71.18 | 73.26 | 69.34 | 71.50 | 72.83 |
| mfeat-pixel | 83.71 | 86.63 | 96.11 | 96.4 | 96.49 |
| mol-biology-promoters | 81.95 | 84.01 | 89.22 | 89.94 | 89.80 |
| mol-promotor-gene | 82.04 | 83.91 | 88.36 | 90.43 | 91.05 |

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| mol-splice-junction | 94.11 | 93.44 | 95.16 | 95.60 | 96.05 |
| monks1 | 99.60 | 96.73 | 97.00 | 98.47 | 98.62 |
| monks2 | 63.49 | 63.11 | 65.68 | 69.23 | 70.35 |
| monks3 | 98.77 | 98.76 | 95.65 | 98.72 | 98.83 |
| mushroom | 99.99 | 99.98 | 99.95 | 100 | 100 |
| nursery | 97.87 | 97.01 | 98.59 | 97.34 | 96.1 |
| optdigits | 95.65 | 95.73 | 98.31 | 98.42 | 98.46 |
| page-blocks | 97.31 | 97.33 | 97.13 | 97.45 | 97.49 |
| parkinsons | 87.48 | 87.73 | 90.43 | 91.11 | 90.80 |
| pendigits | 98.46 | 98.44 | 99.15 | 99.17 | 99.12 |
| phoneme | 88.00 | 87.74 | 90.06 | 90.42 | 90.11 |
| pima-diabetes | 75.88 | 74.88 | 74.88 | 75.15 | 75.48 |
| postoperative-patient | 67.56 | 70.44 | 59.67 | 69.11 | 70.44 |
| primary-tumor | 43.01 | 43.28 | 43.15 | 44.54 | 44.37 |
| qsar-biodegradation | 85.38 | 85.00 | 86.02 | 86.10 | 86.36 |
| qualitative-bankruptcy | 98.68 | 98.36 | 97.68 | 99.36 | 99.48 |
| robot-failure-lp1 | 76.60 | 77.19 | 86.24 | 85.78 | 86.36 |
| robot-failure-lp2 | 64.65 | 55.30 | 65.35 | 66.20 | 66.05 |
| robot-failure-lp3 | 59.00 | 55.10 | 70.60 | 69.35 | 70.00 |
| robot-failure-lp4 | 85.76 | 82.75 | 90.93 | 90.68 | 91.09 |
| robot-failure-lp5 | 66.60 | 67.90 | 73.19 | 73.11 | 72.91 |
| saheart | 67.68 | 68.76 | 67.87 | 67.91 | 67.92 |
| seeds | 91.57 | 91.67 | 92.48 | 92.10 | 92.52 |
| segment | 97.59 | 97.38 | 97.09 | 97.42 | 97.71 |
| seismic-bumps | 92.68 | 93.38 | 92.66 | 92.93 | 93.01 |
| sick | 98.68 | 98.66 | 98.4 | 98.5 | 98.55 |
| solar-flare2 | 99.16 | 99.5 | 98.61 | 99.53 | 99.53 |
| sonar | 79.43 | 80.35 | 83.18 | 83.92 | 83.03 |
| soybean | 91.95 | 90.5 | 92.17 | 94.61 | 94.61 |
| spambase | 94.12 | 94.17 | 94.58 | 94.54 | 94.79 |
| spect | 81.91 | 82.64 | 81.16 | 82.82 | 83.57 |
| spectf | 87.40 | 83.39 | 89.74 | 89.57 | 89.77 |
| spectrometer | 55.86 | 52.94 | 56.84 | 57.24 | 56.26 |
| splice | 94.02 | 94.04 | 94.89 | 95.53 | 96.13 |
| sponge | 92.75 | 92.57 | 94.61 | 94.32 | 94.73 |
| synthetic-control | 95.77 | 94.33 | 98.28 | 98.50 | 98.80 |
| tae | 58.75 | 59 | 64.61 | 63.27 | 62.46 |
| teaching-assistant-eval | 57.17 | 53.60 | 64.61 | 63.27 | 62.46 |
| thoracic-surgery | 83.68 | 84.98 | 82.55 | 83.34 | 84.32 |
| tic-tac-toe | 91.06 | 88.73 | 94.91 | 95.04 | 93.67 |
| trains | 78.00 | 56.00 | 54.00 | 62.00 | 62.00 |
| turkiye-student | 35.85 | 38.48 | 36.15 | 37.19 | 38.56 |
| user-knowledge | 90.00 | 89.87 | 90.42 | 90.47 | 90.30 |
| vehicle | 74.51 | 74.1 | 74.92 | 74.97 | 74.87 |
| vote | 96.04 | 95.79 | 95.56 | 96.04 | 95.88 |
| vowel | 93.35 | 91.78 | 95.41 | 96.15 | 96.25 |
| waveform | 83.18 | 83.14 | 85.03 | 84.94 | 84.74 |
| wine | 95.89 | 95.27 | 97.85 | 97.24 | 96.95 |
| wisconsin-breast-cancer | 95.94 | 96.04 | 95.81 | 96.52 | 96.74 |
| zoo | 93.01 | 92.5 | 95.67 | 96.18 | 95.78 |

**Table 8.** Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 10%.

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| acute-inflamm-nephritis | 99.33 | 99.42 | 92.25 | 98.33 | 98.83 |
| acute-inflamm-urinary | 99.17 | 98.58 | 91.42 | 96.83 | 97.25 |
| anneal | 98.05 | 98.5 | 96.44 | 98.34 | 98.91 |
| appendicitis | 85.52 | 84.96 | 82.45 | 82.74 | 84.73 |
| arrhythmia | 74.29 | 73.88 | 67.76 | 68.83 | 70.51 |
| audiology | 80.84 | 79.28 | 75.72 | 78.83 | 78.96 |
| autos | 80.44 | 75.79 | 77.16 | 79.06 | 80.04 |
| balance-scale | 81.09 | 81.97 | 78.03 | 81.26 | 82.57 |
| bank-marketing | 89.13 | 89.33 | 88.62 | 89.40 | 89.50 |
| banknote-auth | 98.26 | 98.56 | 97.19 | 98.34 | 98.70 |
| blogger | 75.60 | 73.40 | 79.60 | 79.90 | 80.40 |
| breast-cancer | 67.17 | 69.87 | 66.77 | 70.89 | 72.88 |
| bridges-version1 | 57.16 | 50.93 | 51.68 | 59.63 | 66.98 |
| bridges-version2 | 55.62 | 51.72 | 50.35 | 59.29 | 64.56 |
| bupa | 68.25 | 69.45 | 69.38 | 69.87 | 70.22 |
| car | 90.92 | 92.34 | 93.3 | 93.44 | 92.72 |
| cleveland-heart-dis | 80.3 | 79.73 | 80.73 | 80.6 | 81.4 |
| cmc | 50.12 | 51.82 | 48.51 | 50.17 | 51.88 |
| credit-rating | 83.3 | 84.77 | 84.01 | 85.26 | 86.07 |
| crx | 85.59 | 85.58 | 84.01 | 85.28 | 86.07 |
| cylinder-bands | 57.96 | 66.28 | 71.91 | 75.61 | 79.39 |
| dermatology | 95.46 | 93.82 | 96.25 | 96.88 | 97.26 |
| dresses-sales | 55.00 | 58.18 | 54.20 | 56.58 | 57.60 |
| ecoli | 84.82 | 84.7 | 83.87 | 84.76 | 84.55 |
| fertility-diagnosis | 84.10 | 88.20 | 82.40 | 83.40 | 85.20 |
| flags | 47.14 | 58.12 | 58.73 | 62.01 | 61.80 |
| german-credit | 72.67 | 73.43 | 74.79 | 75.05 | 75.26 |
| glass | 73.33 | 74.37 | 76.82 | 76.27 | 76.17 |
| glioma16 | 77.60 | 79.40 | 81.60 | 82.40 | 82.60 |
| haberman | 69.05 | 70.44 | 62.66 | 69.72 | 73.15 |
| hayes-roth | 80.31 | 80.25 | 78.19 | 79.44 | 79.81 |
| heart-statlog | 79.7 | 79.26 | 79.37 | 79.78 | 80 |
| hepatitis | 80.63 | 81.53 | 82.78 | 82.14 | 83.35 |
| horse-colic | 84.55 | 83.71 | 83.58 | 83.31 | 84.29 |
| hungarian-heart-dis | 78.96 | 79.46 | 79.56 | 79.77 | 81.12 |
| hypothyroid | 99.3 | 99.48 | 99.22 | 99.48 | 99.57 |
| ionosphere | 91.8 | 90.58 | 92.31 | 92.42 | 92.51 |
| iris | 93.8 | 94.2 | 90.07 | 92.33 | 93.47 |
| kr-vs-kp | 98.02 | 98.72 | 96.57 | 98.09 | 98.54 |
| labor | 81.57 | 81.87 | 86.90 | 86.07 | 86.20 |
| leaf | 67.97 | 69.09 | 75.38 | 74.65 | 74.85 |
| letter | 93.56 | 92.56 | 94.04 | 95.87 | 95.8 |
| leukemia-haslinger | 78.30 | 77.60 | 85.30 | 85.30 | 85.80 |
| liver-disorders | 70.47 | 69.43 | 69.38 | 69.87 | 70.22 |
| lsvt-voice-rehab | 76.54 | 80.21 | 81.03 | 80.33 | 80.88 |
| lymphography | 79.58 | 77.02 | 83.09 | 82.62 | 81.87 |
| mfeat-morphological | 70.15 | 72.83 | 68.83 | 70.38 | 71.95 |
| mfeat-pixel | 83.09 | 86.71 | 95.82 | 96.33 | 96.37 |
| mol-biology-promoters | 80.39 | 80.43 | 85.95 | 85.75 | 88.56 |
| mol-promotor-gene | 80.39 | 80.73 | 86.19 | 87.28 | 87.51 |

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| mol-splice-junction | 93.68 | 93.05 | 94.06 | 94.72 | 95.49 |
| monks1 | 98.58 | 95.29 | 92.70 | 95.34 | 96.28 |
| monks2 | 62.08 | 62.61 | 65.12 | 67.48 | 68.60 |
| monks3 | 98.29 | 98.60 | 93.24 | 97.93 | 98.58 |
| mushroom | 99.98 | 99.97 | 99.68 | 99.97 | 99.99 |
| nursery | 96.27 | 97.11 | 97.55 | 97.55 | 96.2 |
| optdigits | 95.7 | 95.81 | 98.26 | 98.34 | 98.33 |
| page-blocks | 97.11 | 97.2 | 96.49 | 97.15 | 97.35 |
| parkinsons | 87.33 | 87.27 | 89.76 | 90.02 | 89.67 |
| pendigits | 98.43 | 98.43 | 99.08 | 99.08 | 99.04 |
| phoneme | 86.93 | 87.22 | 88.39 | 88.83 | 89.08 |
| pima-diabetes | 75.59 | 74.48 | 74.24 | 74.16 | 74.86 |
| postoperative-patient | 64.22 | 70.22 | 57.67 | 66.67 | 69.44 |
| primary-tumor | 41.62 | 43.06 | 42.15 | 43.36 | 44.22 |
| qsar-biodegradation | 84.21 | 84.48 | 84.54 | 84.81 | 85.28 |
| qualitative-bankruptcy | 98.60 | 98.36 | 94.16 | 98.48 | 99.08 |
| robot-failure-lp1 | 74.19 | 75.81 | 86.83 | 86.49 | 85.79 |
| robot-failure-lp2 | 58.75 | 52.90 | 62.75 | 63.35 | 64.25 |
| robot-failure-lp3 | 57.85 | 55.35 | 70.40 | 69.55 | 70.20 |
| robot-failure-lp4 | 84.56 | 83.35 | 90.08 | 90.65 | 90.39 |
| robot-failure-lp5 | 66.13 | 65.35 | 73.09 | 73.40 | 72.92 |
| saheart | 67.48 | 67.98 | 66.29 | 66.36 | 66.68 |
| seeds | 91.05 | 90.86 | 91.00 | 91.29 | 91.62 |
| segment | 96.75 | 97.08 | 95.92 | 96.34 | 97.24 |
| seismic-bumps | 92.14 | 93.25 | 91.97 | 92.39 | 92.92 |
| sick | 98.08 | 98.47 | 98.18 | 98.28 | 98.32 |
| solar-flare2 | 98.58 | 99.47 | 97.56 | 99.46 | 99.52 |
| sonar | 77.45 | 79.47 | 81.61 | 82.08 | 81.35 |
| soybean | 91.22 | 90.25 | 90.41 | 94.03 | 94.42 |
| spambase | 93.23 | 93.32 | 93.13 | 93.33 | 93.73 |
| spect | 81.44 | 83.06 | 80.16 | 81.66 | 82.60 |
| spectf | 84.13 | 82.55 | 87.36 | 86.73 | 87.30 |
| spectrometer | 55.42 | 51.85 | 56.39 | 56.41 | 55.71 |
| splice | 93.11 | 93.54 | 93.98 | 94.73 | 95.52 |
| sponge | 91.39 | 92.68 | 92.98 | 93.52 | 93.93 |
| synthetic-control | 95.52 | 80.73 | 98.05 | 98.72 | 98.88 |
| tae | 56.17 | 57.15 | 61.69 | 60.28 | 59.82 |
| teaching-assistant-eval | 55.68 | 50.82 | 61.69 | 60.28 | 59.82 |
| thoracic-surgery | 82.70 | 84.77 | 81.13 | 82.19 | 83.57 |
| tic-tac-toe | 88.25 | 87.52 | 92.45 | 93.34 | 92.07 |
| trains | 72.00 | 85.00 | 76.00 | 79.00 | 79.00 |
| turkiye-student | 36.02 | 38.41 | 35.55 | 36.71 | 37.98 |
| user-knowledge | 89.28 | 89.48 | 89.75 | 89.70 | 89.73 |
| vehicle | 73.88 | 73.54 | 74.48 | 74.49 | 74.48 |
| vote | 95.22 | 95.35 | 94.11 | 95.28 | 95.54 |
| vowel | 92.73 | 90.74 | 92.18 | 93.28 | 94.67 |
| waveform | 83.16 | 83.16 | 84.94 | 84.9 | 84.8 |
| wine | 94.44 | 94.5 | 96.86 | 96.19 | 96.13 |
| wisconsin-breast-cancer | 95.49 | 95.75 | 94.64 | 96.01 | 96.32 |
| zoo | 93.66 | 93.37 | 92.97 | 95.86 | 95.47 |

**Table 9.** Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 20%.

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| acute-inflamm-nephritis | 95.75 | 95.92 | 80.33 | 91.92 | 94.67 |
| acute-inflamm-urinary | 95.42 | 94.17 | 82.00 | 90.83 | 94.00 |
| anneal | 95.34 | 97.42 | 91.16 | 95.51 | 97.55 |
| appendicitis | 81.90 | 83.05 | 76.84 | 77.00 | 80.25 |
| arrhythmia | 73.87 | 72.84 | 66.75 | 66.73 | 68.74 |
| audiology | 76.25 | 75.57 | 71.28 | 75.23 | 76.62 |
| autos | 73.34 | 69.8 | 70.63 | 73.4 | 73.7 |
| balance-scale | 79.26 | 80.97 | 75.28 | 80.38 | 81.93 |
| bank-marketing | 86.06 | 88.65 | 85.60 | 88.00 | 88.77 |
| banknote-auth | 96.38 | 97.51 | 91.04 | 93.51 | 96.34 |
| blogger | 71.40 | 70.30 | 73.70 | 74.80 | 74.70 |
| breast-cancer | 63.4 | 66.2 | 62.02 | 66.79 | 70.78 |
| bridges-version1 | 52.11 | 35.82 | 44.79 | 51.73 | 62.30 |
| bridges-version2 | 49.85 | 36.13 | 44.85 | 51.43 | 61.50 |
| bupa | 64.72 | 66.53 | 65.84 | 65.70 | 66.50 |
| car | 85.43 | 89.72 | 90.48 | 91.53 | 91.41 |
| cleveland-heart-dis | 79.02 | 79.15 | 79.48 | 79.91 | 80.17 |
| cmc | 48.38 | 50.14 | 46.58 | 48.7 | 50.6 |
| credit-rating | 79.41 | 82.67 | 80 | 82.77 | 84.03 |
| crx | 81.87 | 84.58 | 80.00 | 82.77 | 84.03 |
| cylinder-bands | 58.13 | 60.46 | 67.20 | 69.37 | 73.76 |
| dermatology | 92.73 | 93.52 | 94.86 | 95.76 | 96.57 |
| dresses-sales | 53.32 | 56.78 | 53.56 | 54.18 | 55.34 |
| ecoli | 82.56 | 82.91 | 80.74 | 81.48 | 82.82 |
| fertility-diagnosis | 77.80 | 86.60 | 78.60 | 79.70 | 80.60 |
| flags | 36.77 | 55.99 | 56.61 | 59.05 | 59.69 |
| german-credit | 69.91 | 71.38 | 71.8 | 72.71 | 73.11 |
| glass | 70.61 | 72.67 | 72.72 | 71.94 | 72.12 |
| glioma16 | 76.60 | 76.20 | 80.60 | 80.80 | 81.40 |
| haberman | 66.55 | 66.33 | 59.43 | 64.05 | 70.53 |
| hayes-roth | 78.00 | 76.63 | 72.38 | 75.75 | 76.94 |
| heart-statlog | 76.93 | 76.81 | 76.93 | 77 | 77.52 |
| hepatitis | 79.35 | 79.95 | 79.69 | 79.56 | 80.5 |
| horse-colic | 81.46 | 80.73 | 80.7 | 81.3 | 81.71 |
| hungarian-heart-dis | 78.14 | 78.36 | 77.81 | 78.75 | 80.97 |
| hypothyroid | 98.34 | 99.37 | 98.65 | 99.14 | 99.3 |
| ionosphere | 87.84 | 86.7 | 88.39 | 88.15 | 89.01 |
| iris | 90.07 | 90.93 | 82.8 | 88.87 | 90.67 |
| kr-vs-kp | 92.68 | 95.63 | 90.37 | 93.27 | 95.71 |
| labor | 77.13 | 80.60 | 80.53 | 82.57 | 82.37 |
| leaf | 65.97 | 66.00 | 72.15 | 72.71 | 74.15 |
| letter | 92.57 | 92.32 | 90.57 | 94.6 | 95.16 |
| leukemia-haslinger | 73.80 | 76.40 | 81.70 | 82.50 | 82.50 |
| liver-disorders | 67.08 | 66.45 | 65.84 | 65.7 | 66.5 |
| lsvt-voice-rehab | 74.15 | 78.07 | 77.38 | 77.69 | 77.62 |
| lymphography | 75.99 | 76 | 78.08 | 80.58 | 80.12 |
| mfeat-morphological | 67.98 | 72.24 | 66.05 | 67.68 | 70.37 |
| mfeat-pixel | 82.19 | 86.6 | 95.32 | 95.85 | 95.96 |
| mol-biology-promoters | 72.65 | 72.19 | 76.93 | 77.52 | 79.34 |
| mol-promotor-gene | 72.65 | 73.03 | 77.15 | 78.24 | 77.06 |

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| mol-splice-junction | 91.53 | 92.08 | 91.54 | 92.25 | 93.47 |
| monks1 | 90.37 | 89.69 | 83.27 | 86.58 | 89.12 |
| monks2 | 61.71 | 61.41 | 62.24 | 64.21 | 64.79 |
| monks3 | 95.20 | 97.30 | 85.15 | 91.14 | 95.62 |
| mushroom | 99.75 | 99.83 | 96.76 | 99.08 | 99.85 |
| nursery | 90.42 | 96.5 | 93.74 | 97.2 | 96.4 |
| optdigits | 95.73 | 96.07 | 98.01 | 98.08 | 98.19 |
| page-blocks | 96.33 | 96.79 | 94.68 | 95.97 | 96.94 |
| parkinsons | 80.63 | 84.02 | 84.13 | 84.94 | 84.59 |
| pendigits | 98.08 | 98.19 | 98.75 | 98.83 | 98.91 |
| phoneme | 84.50 | 85.41 | 83.48 | 84.08 | 85.95 |
| pima-diabetes | 74.62 | 72.6 | 71.85 | 72.68 | 72.72 |
| postoperative-patient | 62.56 | 68.67 | 55.56 | 59.44 | 66.22 |
| primary-tumor | 40.2 | 41.03 | 40.53 | 41.8 | 42.83 |
| qsar-biodegradation | 79.59 | 82.39 | 80.72 | 80.83 | 81.76 |
| qualitative-bankruptcy | 97.84 | 98.08 | 87.08 | 93.92 | 97.32 |
| robot-failure-lp1 | 74.42 | 73.89 | 83.57 | 83.28 | 83.33 |
| robot-failure-lp2 | 54.55 | 51.35 | 60.80 | 61.40 | 62.05 |
| robot-failure-lp3 | 51.55 | 51.10 | 65.50 | 65.75 | 65.75 |
| robot-failure-lp4 | 80.52 | 79.34 | 88.10 | 88.42 | 88.70 |
| robot-failure-lp5 | 63.36 | 63.05 | 70.03 | 71.03 | 70.49 |
| saheart | 65.99 | 66.21 | 64.84 | 64.77 | 65.62 |
| seeds | 86.95 | 89.52 | 86.38 | 87.67 | 88.14 |
| segment | 94.29 | 95.83 | 93.48 | 93.92 | 95.64 |
| seismic-bumps | 89.37 | 92.48 | 88.18 | 89.31 | 91.57 |
| sick | 96.14 | 97.87 | 96.82 | 97.3 | 97.82 |
| solar-flare2 | 96.45 | 99.23 | 94.76 | 98.99 | 99.48 |
| sonar | 74.77 | 76.27 | 78.54 | 79 | 79.33 |
| soybean | 88.07 | 87.7 | 84.83 | 92.33 | 93.6 |
| spambase | 90.39 | 89.95 | 89.33 | 89.63 | 90.3 |
| spect | 78.44 | 81.57 | 77.56 | 79.55 | 81.69 |
| spectf | 80.23 | 79.84 | 82.09 | 81.72 | 82.20 |
| spectrometer | 54.15 | 49.97 | 55.86 | 55.03 | 54.98 |
| splice | 90.87 | 91.5 | 91.52 | 92.22 | 93.43 |
| sponge | 87.89 | 90.57 | 89.45 | 90.82 | 91.66 |
| synthetic-control | 93.25 | 16.67 | 97.57 | 98.12 | 98.30 |
| tae | 53.13 | 54.8 | 54.87 | 54.82 | 55.27 |
| teaching-assistant-eval | 50.94 | 48.95 | 54.87 | 54.82 | 55.27 |
| thoracic-surgery | 79.02 | 83.49 | 76.62 | 78.55 | 80.81 |
| tic-tac-toe | 82.18 | 81.94 | 85.33 | 86.69 | 87.52 |
| trains | 50.00 | 45.00 | 48.00 | 49.00 | 50.00 |
| turkiye-student | 34.99 | 37.77 | 34.58 | 35.65 | 36.84 |
| user-knowledge | 86.76 | 88.86 | 86.76 | 86.98 | 88.39 |
| vehicle | 72.59 | 72.41 | 72.52 | 72.87 | 72.98 |
| vote | 92.59 | 93.93 | 90.55 | 93.26 | 94.34 |
| vowel | 88.88 | 84.42 | 84.23 | 85.79 | 88.84 |
| waveform | 82.7 | 82.8 | 84.46 | 84.41 | 84.3 |
| wine | 91.35 | 90.68 | 93.61 | 92.6 | 92.22 |
| wisconsin-breast-cancer | 93.41 | 94 | 90.83 | 93.48 | 95.12 |
| zoo | 93.5 | 93.27 | 87.83 | 93.1 | 94.3 |

**Table 10.** Accuracy results of the ensemble methods when they are used on data sets with a percentage of added label noise equal to 30%.

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| acute-inflamm-nephritis | 82.17 | 86.83 | 69.75 | 79.00 | 84.50 |
| acute-inflamm-urinary | 85.17 | 84.00 | 72.50 | 79.67 | 85.67 |
| anneal | 89.44 | 93.97 | 83.29 | 88.62 | 93.75 |
| appendicitis | 75.23 | 79.03 | 71.05 | 70.30 | 73.40 |
| arrhythmia | 72.86 | 71.64 | 65.58 | 66.48 | 67.57 |
| audiology | 73.37 | 71.51 | 66.02 | 70.54 | 72.71 |
| autos | 64.32 | 62.54 | 61.73 | 64.57 | 65.83 |
| balance-scale | 74.95 | 77.1 | 68.62 | 76.02 | 79.33 |
| bank-marketing | 76.46 | 85.00 | 78.23 | 82.37 | 85.34 |
| banknote-auth | 93.45 | 92.19 | 80.91 | 82.79 | 88.00 |
| blogger | 69.50 | 67.60 | 68.70 | 70.70 | 70.40 |
| breast-cancer | 59.83 | 61.24 | 59.1 | 61.34 | 65.02 |
| bridges-version1 | 38.71 | 33.30 | 41.55 | 46.00 | 56.45 |
| bridges-version2 | 37.85 | 33.09 | 41.29 | 45.31 | 54.35 |
| bupa | 60.50 | 61.60 | 60.26 | 60.60 | 60.34 |
| car | 78.65 | 84.87 | 85.41 | 87.54 | 88.92 |
| cleveland-heart-dis | 75.6 | 76.57 | 75.82 | 76.5 | 78.22 |
| cmc | 45.41 | 47.51 | 43.54 | 45.93 | 47.81 |
| credit-rating | 71.61 | 75.3 | 71.72 | 75.07 | 77.65 |
| crx | 72.16 | 79.49 | 71.72 | 75.07 | 77.65 |
| cylinder-bands | 58.33 | 54.24 | 61.35 | 63.24 | 65.94 |
| dermatology | 88.71 | 91.04 | 92.84 | 93.79 | 94.94 |
| dresses-sales | 51.48 | 53.64 | 52.02 | 53.16 | 53.36 |
| ecoli | 79.88 | 80.86 | 77.34 | 78.79 | 81.03 |
| fertility-diagnosis | 71.60 | 81.80 | 69.50 | 71.40 | 72.70 |
| flags | 35.22 | 47.49 | 54.00 | 55.18 | 56.84 |
| german-credit | 65.07 | 67.19 | 66.93 | 67.75 | 69.09 |
| glass | 66.9 | 68.39 | 67.69 | 68.48 | 68.02 |
| glioma16 | 63.80 | 72.20 | 69.00 | 71.00 | 72.60 |
| haberman | 62.34 | 60.46 | 56.03 | 59.26 | 63.94 |
| hayes-roth | 73.56 | 72.44 | 66.81 | 71.75 | 73.25 |
| heart-statlog | 69.52 | 69.89 | 70.96 | 70.93 | 72.67 |
| hepatitis | 73.24 | 75.51 | 75.24 | 75.95 | 77.31 |
| horse-colic | 76.04 | 75.16 | 74.34 | 75.49 | 75.95 |
| hungarian-heart-dis | 75.96 | 76.09 | 74.1 | 75.76 | 78.55 |
| hypothyroid | 95.9 | 98.82 | 97.31 | 97.94 | 98.63 |
| ionosphere | 79.86 | 78.38 | 81.01 | 80.98 | 81.49 |
| iris | 81.73 | 84.13 | 73.47 | 80.73 | 85.13 |
| kr-vs-kp | 82.68 | 86.36 | 79.88 | 82.81 | 87.57 |
| labor | 75.77 | 77.53 | 74.37 | 76.77 | 78.97 |
| leaf | 61.56 | 63.09 | 66.74 | 68.09 | 70.53 |
| letter | 90.29 | 91.3 | 85.85 | 92.15 | 94.02 |
| leukemia-haslinger | 68.10 | 71.00 | 75.40 | 77.00 | 78.00 |
| liver-disorders | 61.66 | 61.44 | 60.26 | 60.6 | 60.34 |
| lsvt-voice-rehab | 66.95 | 70.91 | 70.97 | 72.89 | 73.13 |
| lymphography | 73.02 | 73.14 | 72.06 | 76.72 | 79.15 |
| mfeat-morphological | 64.51 | 71.00 | 62.14 | 64.68 | 67.91 |
| mfeat-pixel | 81.81 | 87.03 | 94.35 | 95.46 | 95.57 |
| mol-biology-promoters | 64.43 | 63.85 | 69.20 | 69.34 | 72.07 |
| mol-promotor-gene | 64.23 | 64.69 | 69.76 | 70.36 | 71.30 |

| Data set | BA-C4.5 | BA-CDT | RF | CRF | RCRF |
|---|---|---|---|---|---|
| mol-splice-junction | 85.51 | 89.76 | 87.49 | 88.31 | 89.86 |
| monks1 | 77.81 | 80.09 | 73.40 | 75.20 | 78.62 |
| monks2 | 58.85 | 58.69 | 57.17 | 58.57 | 60.77 |
| monks3 | 85.45 | 90.14 | 75.16 | 78.58 | 85.57 |
| mushroom | 94.23 | 97.31 | 87.93 | 90.99 | 97.66 |
| nursery | 81.74 | 93.42 | 87.09 | 94.79 | 96.12 |
| optdigits | 95.08 | 95.9 | 97.73 | 97.69 | 97.67 |
| page-blocks | 94.22 | 95.73 | 91.53 | 93.28 | 95.73 |
| parkinsons | 75.07 | 78.36 | 75.58 | 75.23 | 75.42 |
| pendigits | 97.39 | 97.76 | 98.04 | 98.22 | 98.51 |
| phoneme | 81.12 | 80.50 | 75.28 | 75.69 | 78.64 |
| pima-diabetes | 70.8 | 67.5 | 67.04 | 66.92 | 68.2 |
| postoperative-patient | 57.44 | 66.78 | 54.00 | 55.56 | 59.67 |
| primary-tumor | 37.61 | 39.73 | 37.14 | 38.7 | 40.53 |
| qsar-biodegradation | 71.68 | 77.09 | 73.54 | 73.97 | 74.69 |
| qualitative-bankruptcy | 91.80 | 94.00 | 75.44 | 80.88 | 88.52 |
| robot-failure-lp1 | 69.18 | 69.10 | 80.94 | 80.38 | 81.03 |
| robot-failure-lp2 | 50.15 | 49.30 | 61.05 | 60.15 | 61.25 |
| robot-failure-lp3 | 49.70 | 48.85 | 59.80 | 60.75 | 59.95 |
| robot-failure-lp4 | 72.47 | 73.64 | 83.14 | 83.98 | 84.60 |
| robot-failure-lp5 | 58.02 | 58.79 | 66.97 | 66.00 | 66.19 |
| saheart | 63.17 | 62.99 | 61.29 | 60.96 | 62.23 |
| seeds | 81.14 | 87.71 | 80.33 | 81.05 | 83.52 |
| segment | 90.5 | 93.15 | 90.13 | 90.49 | 92.33 |
| seismic-bumps | 83.82 | 88.67 | 80.70 | 81.97 | 85.44 |
| sick | 90.34 | 94.64 | 91.44 | 92.46 | 94.43 |
| solar-flare2 | 92.24 | 97.11 | 90.19 | 95.52 | 99.15 |
| sonar | 69.52 | 71.75 | 72.75 | 72.42 | 72.48 |
| soybean | 83.45 | 81.65 | 79.31 | 89.47 | 92.09 |
| spambase | 86.06 | 83.51 | 83.02 | 83.21 | 83.87 |
| spect | 71.82 | 77.95 | 71.13 | 73.22 | 76.34 |
| spectf | 71.34 | 74.76 | 75.76 | 76.32 | 75.81 |
| spectrometer | 51.62 | 47.97 | 53.58 | 53.94 | 53.6 |
| splice | 87.83 | 88.76 | 87.55 | 88.33 | 89.93 |
| sponge | 77.45 | 84.05 | 81.07 | 84.34 | 86.38 |
| synthetic-control | 86.10 | 16.67 | 97.07 | 97.48 | 97.72 |
| tae | 49.83 | 49.2 | 51.38 | 51.4 | 50.35 |
| teaching-assistant-eval | 48.42 | 46.90 | 51.38 | 51.40 | 50.35 |
| thoracic-surgery | 71.34 | 78.66 | 69.38 | 72.02 | 74.09 |
| tic-tac-toe | 72.91 | 73.99 | 75.37 | 76.20 | 78.16 |
| trains | 73.00 | 76.00 | 74.00 | 77.00 | 76.00 |
| turkiye-student | 33.80 | 37.23 | 32.97 | 34.31 | 35.59 |
| user-knowledge | 81.03 | 87.54 | 82.56 | 83.60 | 85.66 |
| vehicle | 70.13 | 70.36 | 69.86 | 69.83 | 70.58 |
| vote | 86.25 | 88.87 | 83.33 | 87 | 89.97 |
| vowel | 81.63 | 74.76 | 75.21 | 76.72 | 80.63 |
| waveform | 81.82 | 82.14 | 83.6 | 83.57 | 83.57 |
| wine | 85.63 | 85.69 | 88.9 | 89.02 | 88.73 |
| wisconsin-breast-cancer | 87.78 | 88.35 | 82.88 | 86.32 | 90.28 |
| zoo | 89.71 | 90.71 | 80.5 | 87.56 | 89.93 |

**Table 11.** p-values of the Nemenyi test about the accuracy on data sets without added noise.

| $i$ | algorithms | $p$ |
|---|---|---|
| 10 | BA-CDT vs. **RCRF** | 0 |
| 9 | BA-CDT vs. **CRF** | 0 |
| 8 | BA-C4.5 vs. **RCRF** | 0 |
| 7 | BA-C4.5 vs. **CRF** | 0 |
| 6 | BA-CDT vs. **RF** | 0.000003 |
| 5 | BA-C4.5 vs. **RF** | 0.000574 |
| 4 | RF vs. **RCRF** | 0.001013 |
| 3 | RF vs. CRF | 0.01673 |
| 2 | BA-C4.5 vs. BA-CDT | 0.210498 |
| 1 | CRF vs. RCRF | 0.371093 |

**Table 12.** p-values of the Nemenyi test about the accuracy on data sets with 5% of added noise.

| $i$ | algorithms | $p$ |
|---|---|---|
| 10 | BA-CDT vs. **RCRF** | 0 |
| 9 | BA-C4.5 vs. **RCRF** | 0 |
| 8 | BA-CDT vs. **CRF** | 0 |
| 7 | RF vs. **RCRF** | 0 |
| 6 | BA-C4.5 vs. **CRF** | 0 |
| 5 | RF vs. **CRF** | 0 |
| 4 | BA-CDT vs. RF | 0.066717 |
| 3 | CRF vs. RCRF | 0.070108 |
| 2 | BA-C4.5 vs. RF | 0.158917 |
| 1 | BA-C4.5 vs. BA-CDT | 0.670944 |

**Table 13.** p-values of the Nemenyi test about the accuracy on data sets with 10% of added noise.

| $i$ | algorithms | $p$ |
|---|---|---|
| 10 | BA-C4.5 vs. **RCRF** | 0 |
| 9 | RF vs. **RCRF** | 0 |
| 8 | BA-CDT vs. **RCRF** | 0 |
| 7 | BA-C4.5 vs. **CRF** | 0 |
| 6 | RF vs. **CRF** | 0.000004 |
| 5 | BA-CDT vs. **CRF** | 0.000083 |
| 4 | CRF vs. **RCRF** | 0.001745 |
| 3 | BA-C4.5 vs. BA-CDT | 0.107405 |
| 2 | BA-C4.5 vs. RF | 0.347654 |
| 1 | BA-CDT vs. RF | 0.502335 |

**Table 14.** p-values of the Nemenyi test about the accuracy on data sets with 20% of added noise.

| $i$ | algorithms | $p$ |
|---|---|---|
| 10 | RF vs.**RCRF** | 0 |
| 9 | BA-C4.5 vs. **RCRF** | 0 |
| 8 | BA-CDT vs. **RCRF** | 0 |
| 7 | CRF vs. **RCRF** | 0 |
| 6 | RF vs. **CRF** | 0.000001 |
| 5 | BA-C4.5 vs.**CRF** | 0.000035 |
| 4 | **BA-CDT** vs. RF | 0.000083 |
| 3 | BA-C4.5 vs. **BA-CDT** | 0.001883 |
| 2 | BA-CDT vs. CRF | 0.303672 |
| 1 | BA-C4.5 vs. RF | 0.408041 |

**Table 15.** p-values of the Nemenyi test about the accuracy on data sets with 30% of added noise.

| $i$ | algorithms | $p$ |
|---|---|---|
| 10 | RF vs. **RCRF** | 0 |
| 9 | BA-C4.5 vs. **RCRF** | 0 |
| 8 | **BA-CDT** vs. RF | 0 |
| 7 | CRF vs. **RCRF** | 0 |
| 6 | RF vs. **CRF** | 0.000001 |
| 5 | BA-CDT vs.**RCRF** | 0.000001 |
| 4 | BA-C4.5 vs. **BA-CDT** | 0.000014 |
| 3 | BA-C4.5 vs. **CRF** | 0.000268 |
| 2 | BA-C4.5 vs. RF | 0.194659 |
| 1 | BA-CDT vs. CRF | 0.488196 |