Expert Systems with Applications 41 (2014) 4625–4637

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data

Carlos J. Mantas, Joaquín Abellán *

Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Imprecise probabilities
Imprecise Dirichlet Model
Uncertainty measures
Credal decision trees
C4.5 algorithm
Noisy data

ABSTRACT

In the area of classification, C4.5 is a known algorithm widely used to design decision trees. In this algorithm, a pruning process is carried out to solve the problem of the over-fitting. A modification of C4.5, called *Credal-C4.5*, is presented in this paper. This new procedure uses a mathematical theory based on imprecise probabilities, and uncertainty measures. In this way, Credal-C4.5 estimates the probabilities of the features and the class variable by using imprecise probabilities. Besides it uses a new split criterion, called Imprecise Information Gain Ratio, applying uncertainty measures on convex sets of probability distributions (credal sets). In this manner, Credal-C4.5 builds trees for solving classification problems assuming that the training set is not fully reliable. We carried out several experimental studies comparing this new procedure with other ones and we obtain the following principal conclusion: in domains of class noise, Credal-C4.5 obtains smaller trees and better performance than classic C4.5.

1. Introduction and justification

A decision tree (DT) is a very useful tool for classification. Its structure is simple and easy to interpret. Moreover, to build the classification model normally requires a short time. When a DT is used for classification, a key question is the adjustment degree of the model to the training set. If the algorithm to build a DT employs a tight stopping criteria, then it tends to create small and underfitted DTs. On the other hand, if the algorithm uses a loose stopping criteria, then it tends to generate large DTs that over-fit the data of the training set. Pruning methods were developed for solving this dilemma. According to this methodology, a loosely stopping criterion is used, letting the DT to over-fit the training set. Then the over-fitted tree is cut back into a smaller tree by removing subbranches that are not contributing to the generalization accuracy (Rokach & Maimon, 2010). It has been shown in various studies that employing pruning methods can improve the general performance of a DT, especially in noisy domains.

The ID3 algorithm (Quinlan, 1986) and its extension C4.5 (Quinlan, 1993) are widely used for designing decision trees. C4.5 improves to ID3 algorithm with several characteristics: handling of continuous attributes, dealing training data with missing attribute values and a process for pruning a built tree.

There are different post-pruning processes for DTs (see Rokach & Maimon (2010) for a revision). They are based on estimating the generalization error and then removing useless sub-branches according this information. Usually, the basic idea of this estimation is that the ratio of error, calculated by using the training set, is not quite reliable. The training error is corrected in order to obtain a more realistic measure.

On the other hand, C4.5 algorithm uses a measure of information gain ratio for selecting an input variable in each node (split criterion). This variable selection process is based on the precise probabilities calculated from the training set. Therefore, C4.5 considers that the training set is reliable when the variable selection process is carried out, and it considers that the training set is not reliable when the pruning process is made. This situation can be unsuitable, specially when noisy data are classified. Let us see an example of this situation.

Example 1. Let us suppose a noisy data set composed by 15 instances, 9 instances of class A and 6 instances of class B. We consider that there are two binary feature variables X_1 and X_2. According with the values of these variables, the instances are organized in the following way:

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

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

http://dx.doi.org/10.1016/j.eswa.2014.01.017
0957-4174/

If this data set appears in the node of a tree, then the C4.5 algorithm chooses the variable X_1 for splitting the node (see Fig. 1).

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

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

If this data set is found in the node of a tree, then the C4.5 algorithm chooses the variable X_2 for splitting the node (see Fig. 2).

We can observe that C4.5 algorithm generates an incorrect subtree when noisy data are processed, because it considers that the data set is reliable. Later, the pruning process considers that the data set is not reliable in order to solve this problem. However, the pruning process can only delete the generated incorrect subtree. It can not make a detailed adjustment of the correct subtree illustrated in Fig. 2. The ideal situation is to carry out the branching shown in Fig. 2 and then to make the pruning process. This situation is achieved by using decision trees based on imprecise probabilities as it will be shown later.

In the last years, several formal theories for manipulation of imprecise probabilities have been developed (Walley, 1996; Wang, 2010; Weichselberger, 2000). By using the theory of imprecise probabilities presented in Walley (1996), known as the Imprecise Dirichlet Model (IDM), Abellán and Moral (2003) have developed an algorithm for designing decision trees, called *credal decision trees* (CDTs). The variable selection process for this algorithm (split criterion) is based on imprecise probabilities and uncertainty measures on credal sets, i.e. closed and convex sets of probability distributions. In particular, the CDT algorithm extends the measure of information gain used by ID3. The split criterion is called the Imprecise Info-Gain (IIG).

Recently, in Mantas and Abellán (2014), credal decision trees are built by using an extension of the IIG criterion. In this work, the probability values of the class variable and features are estimated via imprecise probabilities. The CDT algorithm obtains good experimental results (Abellán & Moral, 2005; Abellán & Masegosa, 2009). Besides, its use with bagging ensemble (Abellán & Masegosa, 2009, 2012; Abellán & Mantas, 2014) and its above mentioned extension (Mantas & Abellán, 2014) are especially suitable when noisy data are classified. A complete and recent revision of machine learning methods to manipulate label noise can be found in Frenay and Verleysen (in press). Here, the credal decision tree procedure is included as a *label noise-robust method*.

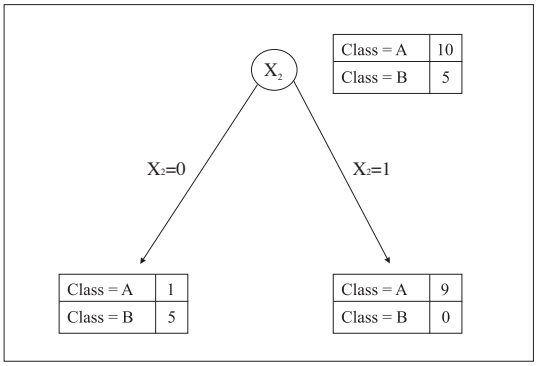

Fig. 2. Branching of a node with clean data produced by C4.5 algorithm.

According to the previous paragraphs, the CDT algorithm and its extensions consider that the training set is not reliable when the variable selection process is carried out. Hence, the problem shown in Example 1 is solved. If the noisy data set appears in the node of a credal tree, then the variable X_2 is chosen for splitting it (see Fig. 3).

Therefore, if we design a new credal tree algorithm inspired on C4.5 (with its improvements and advantages), then we can obtain an algorithm that considers the training set as unreliable when the processes of pruning and variable selection are made. This algorithm will be especially suitable for designing DTs in noisy domains.

Hence, C4.5 algorithm is redefined in this paper by using imprecise probabilities (Credal-C4.5). A new measure called Imprecise Information Gain Ratio (IIGR) is presented as split criterion. IIGR estimates the probability values of the class variable and features with imprecise probabilities as it is presented in Mantas and Abellán (2014). Besides, all the improvements of C4.5 are available: handling of continuous attributes, dealing of missing values, post-pruning process and so on. Credal-C4.5 and classic C4.5 are compared when they classify noisy data. It will be shown that Credal-C4.5 obtains smaller trees and better accuracy results than classic C4.5 with significant statistical difference.

Section 2 briefly describes the necessary previous knowledge about decision trees, C4.5 and credal decision trees. Section 3 presents Credal-C4.5 algorithm. Section 4 analyzes the differences between Credal-C4.5 and classic C4.5. Section 5 compares the action of Credal-C4.5 with the one performed by pessimistic pruning. In Section 6, we describe the experimentation carried out on a wide range of data sets and comments on the results. Finally, Section 7 is devoted to the conclusions and future works.

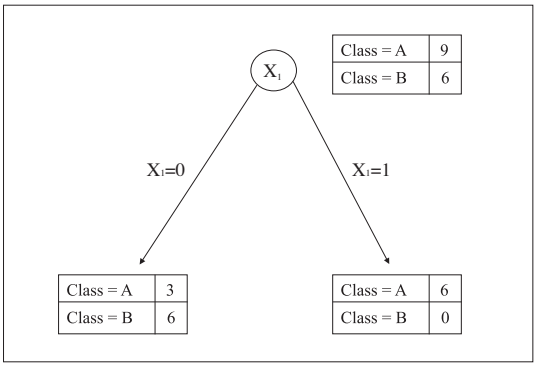

Fig. 1. Branching of a node with noisy data produced by C4.5 algorithm.

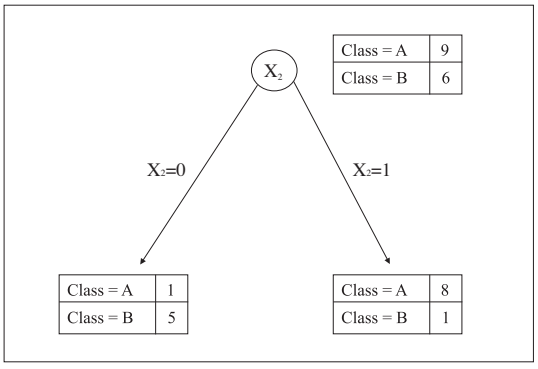

Fig. 3. Branching of a node with noisy data by a credal tree.

2. Previous knowledge

2.1. Decision trees

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

Decision trees are models based on a recursive partition method, the aim of which is to divide the data set using a single variable at each level. This variable is selected with a given criterion. Ideally, they define a set of cases in which all the cases belong to the same class.

Their knowledge representation has a simple tree structure. It can be interpreted as a compact set of rules in which each tree node is labeled with an attribute variable that produces branches for each value. The leaf nodes are labeled with a class label.

The process for inferring a decision tree is mainly determined by the followings aspects:

(i) The criteria used to select the attribute to insert in a node and branching (split criteria).
(ii) The criteria to stop the tree from branching.
(iii) The method for assigning a class label or a probability distribution at the leaf nodes.
(iv) The post-pruning process used to simplify the tree structure.

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

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

2.1.1. Split criteria

Let us suppose a classification problem. Let C be the class variable, $\{X_1,\ldots,X_n\}$ the set of features, and X a general feature. We can find the following split criteria to build a DT.

Info-Gain. This metric was introduced by Quinlan as the basis for his ID3 model (Quinlan, 1986). The model has the following main features: it was defined to obtain decision trees with discrete variables, it does not work with missing values, a pruning process is not carried out and it is based on Shannon's entropy H.

The split criterion of this model is *Info-Gain* (IG) which is defined as:

$$IG(C,X) = H(C) - \sum_i P(X = x_i)H(C|X = x_i), \tag{1}$$

where

$$H(C) = \sum_j P(C = c_j)\log P(C = c_j).$$

In a similar way is expressed $H(C|X = x_i)$.

Info-Gain ratio. In order to improve the ID3 model, Quinlan introduces the C4.5 model (Quinlan, 1993), where the Info-Gain split criterion (the split criterion of ID3) is replaced by an Info-Gain ratio criterion that penalizes variables with many states. The C4.5 model involves a more complete procedure defined to work with continuous variables and missing data. It has a complex subsequent pruning that is introduced to improve the results and obtain less complex structures.

The split criterion for this model is called *Info-Gain Ratio* (IGR) and it is defined as

$$IGR(C,X) = \frac{IG(C,X)}{H(X)}. \tag{2}$$

2.2. C4.5 Tree inducer

In this subsection we will give a brief explanation of the most important aspects of this well known tree inducer. We highlight the main ideas that were introduced in Quinlan (1993):

Split criteria: Information Gain (Quinlan, 1986) (see Eq. (1)) was firstly employed to select the split attribute at each branching node. But this measure is strongly affected by the number of states of the split attribute: attributes with a higher number of states were usually preferred. Quinlan introduced the *Information Gain Ratio* (IGR) criterion (see Eq. (2)) for this new tree inducer, which penalizes variables with many states. This score normalizes the information gain of an attribute X by its own entropy. It is selected the attribute with the highest Info-Gain Ratio score and whose Info-Gain score is higher than the average Info-Gain scores of the valid split attributes. These valid split attributes are those which are numeric or whose number of values is smaller than the thirty percent of the number of instances that are in this branch.

Stopping criteria: The branching of the decision tree is stopped when there is not attribute with a positive Info-Gain Ratio score or there are a minimum number of instances per leaf which is usually set to 2. But in addition to this, using the aforementioned condition in "Split Criteria" of valid split attributes, the branching of a decision tree is also stopped when there is not any valid split attribute.

Handling numeric attributes: This tree inducer manipulates numeric attributes with a very simple approach. Within this method, only binary split attributes are considered and each possible split point is evaluated. Finally, it is selected the point that induces a partition of the samples with the highest Information Gain based split score.

Dealing with missing values: It is assumed that missing values are randomly distributed (*Missing at Random Hypothesis*). In order to compute the scores, the instances are split into pieces. The initial weight of an instance is equal to the unit, but when it goes down a branch receives a weight equal to the proportion of instances that belongs to this branch (weights sum to 1). Information Gain based scores can work with this fractional instances using sum of weights instead of sum of counts.

When making predictions, C4.5 marginalize the missing variable by merging the predictions of all the possible branches that are consistent with the instance (there are several branches because it has a missing value) using their previously computed weights.

Post-pruning process: Although there are many different proposals to carry out a post-pruning process of a decision tree (see Rokach & Maimon (2010)), the technique employed by C4.5 is called *Pessimistic Error Pruning*. This method computes an upper bound of the estimated error rate of a given subtree employing a continuity correction of the Binomial distribution. When the upper bound of a subtree hanging from a given node is greater

than the upper bound of the errors produced by the estimations of this node supposing it acts as a leaf, then this subtree is pruned.

2.3. Credal decision trees

The original split criterion employed to build credal decision trees (CDTs) (Abellán & Moral, 2003) is based on imprecise probabilities and the application of uncertainty measures on credal sets. The mathematical basis of this theory is described below.

Let there be a variable Z whose values belong to $\{z_1,\ldots,z_k\}$. Let us suppose a probability distribution $p(z_j), j=1,\ldots,k$ defined for each value z_j from a data set.

A formal theory of imprecise probability called Walley's Imprecise Dirichlet Model (IDM) (Walley, 1996) is used to estimate probability intervals from the data set for each value of the variable Z. IDM estimates that the probabilities for each value z_j are within the interval:

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

with n_{z_j} as the frequency of the set of values $(Z=z_j)$ in the data set, N the sample size and s a given hyperparameter that does not depend on the sample space (*Representation Invariance Principle*, Walley (1996)). The value of parameter s determines the speed at which the values of probability upper and lower converge when sample size increases. Higher values of s give a more cautious inference. Walley (1996) does not give a definitive recommendation for the value of this parameter but he suggests two candidates: $s=1$ or $s=2$.

One important characteristic of this model is that intervals are wider if the sample size is smaller. Therefore, this method produces more precise intervals at the same time as N increases.

This representation gives rise to a specific kind of convex set of probability distributions on the variable $Z, K(Z)$ (Abellán, 2006). The set is defined as

$$K(Z)=\left\{p\,|\,p(z_j)\in\left[\frac{n_{z_j}}{N+s},\frac{n_{z_j}+s}{N+s}\right],\quad j=1,\ldots,k\right\}. \tag{3}$$

On this type of sets (really 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, a well established total uncertainty measure on credal sets (see Klir (2006)). This function, denoted as H^*, is defined as:

$$H^*(K(Z))=max\{H(p)\,|\,p\in K(Z)\} \tag{4}$$

where the function H is the Shannon's entropy function.

H^* is a total uncertainty measure which is well known for this type of set (see Abellán & Masegosa (2008)). H^* separates conflict and non-specificity (Abellán, Klir, & Moral, 2006), that is, H^* is a disaggregated measure of information that combines two elements:

(a) A conflict or randomness measure that indicates the arrangement of the samples of each class in the training set. This measure is related to the entropy of the probabilities in the convex set.
(b) A non-specificity measure that shows the uncertainty derived from the training set size. This measure is related to the size of the convex set.

The procedure for calculating H^* has a low computational cost for values $s\in(0,2]$ (see Abellán & Moral (2006)). The procedure for the IDM reaches the lowest cost with $s=1$ and it is simple (see Abellán (2006)). For this reason, we will use a value $s=1$ in the experimentation section. Firstly, this procedure consists in determining the set

$$A=\{z_j|n_{z_j}=\min_i\{n_{z_i}\}\} \tag{5}$$

then the distribution with maximum entropy is

$$p^*(z_i)=\begin{cases}\frac{n_{z_i}}{N+s} & \text{if } z_i\notin A\\ \frac{n_{z_i}+s/l}{N+s} & \text{if } z_i\in A\end{cases};\quad i=1,\ldots,k; \tag{6}$$

where l is the number of elements of A.

As the imprecise intervals are wider with smaller sample sizes, there is a tendency to obtain larger values for H^* with small sample sizes. This is due to that the non-specificity component of H^* will be higher in this case. This property will be important for distinguishing the action of Credal-C4.5 as opposed to the behavior of other classic algorithms.

3. Credal-C4.5

The method for building Credal-C4.5 trees is similar to the Quinlan's C4.5 algorithm (Quinlan, 1993). The main difference is that Credal-C4.5 estimates the probability values of the features and the class variable by using imprecise probabilities. As in the CDT procedure, an uncertainty measure on credal sets is used to define a new split criterion. In this way, Credal-C4.5 considers that the training set is not very reliable because it can be affected by class or attribute noise (see Mantas & Abellán (2014)). So, Credal-C4.5 can be considered as a proper method for noisy domains.

Credal-C4.5 is created by replacing the *Info-Gain Ratio* split criterion from C4.5 with the *Imprecise Info-Gain Ratio* (IIGR) split criterion. This criterion can be defined as follows: in a classification problem, let C be the class variable, $\{X_1,\ldots,X_m\}$ the set of features, and X a feature; then

$$IIGR^{\mathcal{D}}(C,X)=\frac{IIG^{\mathcal{D}}(C,X)}{H(X)}, \tag{7}$$

where *Imprecise Info-Gain* (IIG) is equal to:

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

with $K^{\mathcal{D}}(C)$ and $K^{\mathcal{D}}(C|X=x_i)$ are the credal sets obtained via the IDM for the C and $(C|X=x_i)$ variables respectively, for a partition $\mathcal{D}$ of the data set (see Abellán & Moral (2003)); $P^{\mathcal{D}}(X=x_i)(i=1,\ldots,n)$ is a probability distribution that belongs to the credal set $K^{\mathcal{D}}(X)$.

We choose the probability distribution $P^{\mathcal{D}}$ from $K^{\mathcal{D}}(X)$ that maximizes the following expression:

$$\sum_i P(X=x_i)H(C|X=x_i)).$$

It is simple to calculate this probability distribution. Let x_{j_0} be a value for X such that $H(C|X=x_i)$ is the maximum. Then the probability distribution $P^{\mathcal{D}}$ will be

$$P^{\mathcal{D}}(x_i)=\begin{cases}\frac{n_{x_i}}{N+s} & \text{if } i\neq j_0\\ \frac{n_{x_i}+s}{N+s} & \text{if } i=j_0\end{cases}. \tag{9}$$

The IIGR criterion is different from the classical criteria. It is based on the principle of maximum uncertainty (see Klir (2006)), widely used in classic information theory, where it is known as maximum entropy principle. This principle indicates that the probability distribution with the maximum entropy, compatible with available restrictions, must be chosen. Hence, the use of the maximum entropy function in the decision tree building procedure (see Abellán & Moral (2005)) and the definition of probability distribution $P^{\mathcal{D}}$ are justified.

Each node *No* in a decision tree causes a partition of the data set (for the root node, $\mathcal{D}$ is considered to be the entire data set). Furthermore, each *No* node 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 Credal-C4.5 trees is explained in the algorithm in Fig. 4.

We can summarize the main ideas of this procedure:

Split criteria: *Imprecise Info-Gain Ratio* (IIGR) is employed to select the split attribute at each branching node. In a similar way to the classic C4.5 algorithm, it is selected the attribute with the highest Imprecise Info-Gain Ratio score and whose Imprecise Info-Gain score is higher than the average Imprecise Info-Gain scores of the valid split attributes. These valid split attributes are those which are numeric or whose number of values is smaller than the thirty percent of the number of instances which are in this branch.

Labeling leaf node: The most probable value of the class variable in the partition associated with a leaf node is inserted as label, that is, the class label for the leaf node *No* associated with the partition $\mathcal{D}$ is:

$$Class(No,\mathcal{D}) = \max_{c_i \in C} |\{I_j \in \mathcal{D} / class(I_j) = c_i, j = 1,\ldots,|\mathcal{D}|\}|$$

where $class(I_j)$ is the class of the instance $I_j \in \mathcal{D}$ and $|\mathcal{D}|$ is the number of instances in $\mathcal{D}$.

Stopping criteria: The branching of the decision tree is stopped when the uncertainty measure is not reduced ($\alpha \leqslant 0$, step 6) or when there are no more features to insert in a node ($\mathcal{L} = \emptyset$, step 1) or when there are not a minimum number of instances per leaf (step 3). The branching of a decision tree is also stopped when there is not any valid split attribute using the aforementioned condition in "Split Criteria", like classic C4.5.

Handling numeric attributes: The numeric attributes are handled in the same way that C4.5, presented in Section 2.2. The only difference is the use of IIG instead of the IG measure.

Dealing with missing values: The missing values are manipulated in a similar way to C4.5, presented in Section 2.2. Again, the only difference is to use IIG instead of IG.

Post-pruning process: Like C4.5, *Pessimistic Error Pruning* is employed in order to prune a Credal-C4.5.

4. Credal-C4.5 versus classic C4.5

Next, it is commented the situations where Credal-C4.5 and classic C4.5 are different.

Procedure `BuildCredalC4.5Tree`(No, $\mathcal{L}$)

1. If $\mathcal{L} = \emptyset$, then `Exit`.
2. Let $\mathcal{D}$ be the partition associated with node No
3. If $|\mathcal{D}|$ < minimum number of instances, then `Exit`.
4. Calculate $P^{\mathcal{D}}(X = x_i)$ $(i = 1, ..., n)$ on the convex set $K^{\mathcal{D}}(X)$
5. Compute the value
   $\alpha = \max_{X_j \in \mathcal{M}} \{IIGR^{\mathcal{D}}(C, X_j)\}$
   with $\mathcal{M} = \{X_j \in \mathcal{L} / IIG^{\mathcal{D}}(C, X_j) > avg_{X_j \in \mathcal{L}} \{IIG^{\mathcal{D}}(C, X_j)\}\}$
6. If $\alpha \leq 0$ then `Exit`
7. Else
   8. Let X_l be the variable for which the maximum α 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 `BuilCredalC4.5Tree`(No_l, $\mathcal{L}$)

Fig. 4. Procedure to build a Credal-C4.5 decision tree.

(a) **Small data sets**. According Eq. (3), when imprecise probabilities are used to estimate values of a variable, the size of the obtained credal set is proportional to the parameter *s*. Hence, if $s = 0$ the credal set contains only one probability distribution and, in this case, IIGR measure is equal to IGR.[1] If $s > 0$ the size of the credal set is inversely proportional to the data set size *N*. If *N* is very high then the effect of the parameter *s* can be ignored and so the measures IIGR and IGR can be considered equivalent. That is, Credal-C4.5 and classic C4.5 have a similar behavior in the nodes with high associated data set, usually in the upper levels of the tree.

On the other hand, if *N* is small, then the parameter *s* produces credal sets with many probability distributions (big size of the convex set). Hence, the measures IIGR and IGR can be different (maximum entropy of a set H^* can be distinct from classic entropy *H*). That is, Credal-C4.5 and C4.5 have a different behavior in the nodes with small data set, usually in the lower levels of the tree.

(b) **Features with many states**. The IG measure used by ID3 is biased in favor of feature variables with a large number of values. The IGR measure used by C4.5 was created to compensate for this bias. As the IIG measure (Eq. (8)) also penalizes feature variables with many states (see Mantas & Abellán (2014)) and the difference between IGR (Eq. (2)) and IIGR (Eq. (7)) is the use of IIG measure instead of IG, we can conclude that Credal-C4.5 penalizes the features with many values in a higher degree than classic C4.5. As we have said in the previous paragraph, this fact happens in the nodes with a associated small data set.

(c) **Negative split criterion values**. It is important to note that for a feature *X* and a partition $\mathcal{D}$, $IIGR^{\mathcal{D}}(C,X)$ can be negative. This situation does not appear with classical split criteria, such as the IGR criterion used in C4.5. This characteristic enables the IIGR criterion to reveal features that worsen the information on the class variable. Hence, a new stopping criterion is defined for Credal-C4.5 (Step 6 in Fig. 4) that is not available for classic C4.5. In this way, IIGR provides a trade-off between stopping criteria tight and loose, that is, it offers a trade-off between trees small underfitted and large over-fitted. Hence, we can hope that Credal-C4.5 procedure produces smaller trees than the classic C4.5 procedure. In the experimental section we will show that it is so, and also that the accuracy results of the new method are similar or better than the ones of the classic C4.5 procedure before and after pruning.

5. Credal-C4.5 versus pessimistic pruning

Pessimistic pruning is based on estimating the generalization error of a data set. This estimation consists on increasing the training error, that is,

$$e_{gen}(N) = e_{tr}(N) + e_{inc}(N), \tag{10}$$

where $e_{gen}(N)$ is the generalization error of the node N, e_{tr} is the training error and e_{inc} is the increment of the training error.

Next, if the generalization error of the descendant nodes is greater than the one of the parent node, this node is pruned. For example, let us suppose a node N_0 with two descendant nodes N_1 and N_2, then the node N_0 is pruned if the following condition is fulfilled:

$$e_{gen}(N_0) \leqslant e_{gen}(N_1) + e_{gen}(N_2).$$

On the other hand, when we are using the Credal-C4.5 algorithm, a node N_0 is not expanded into the descendant nodes N_1 and N_2 if the following condition is fulfilled for all the available variables X_j and the class variable *C*:

$$IIGR^{\mathcal{D}_0}(C, X_j) \leqslant 0.$$

[1] The Info-Gain ratio criterion is actually a specific case of the IIGR criterion using the parameter $s = 0$.

This condition means that the features worsen the information on the class variable, that is, the maximum of entropy H^* on the partition $\mathcal{D}_0$ is less or equal than the proportional aggregation of H^* on the partitions $\mathcal{D}_1$ and $\mathcal{D}_2$ associated with the descendant nodes, noting as $\mathcal{D}_i$ to the partition associated with the node N_i.

According to Eqs. (4) and (6), the use of H^* is equivalent to work with the classic entropy function H and a probability distribution p^*. This distribution p^* assumes that the information about a data set is imprecise, there is a number s of instances that are unknown. According Eqs. (5) and (6), these unknown instances are assigned to less frequent class for obtaining the maximum of entropy of a credal set. Hence, as the label of a node is equal to the more frequent class, the unknown instances are added into the error of the node, that is, we have the following estimated error for a node N when imprecise probabilities are used:

$$e_{est}(N) = e_{tr}(N) + s. \tag{11}$$

If we compare the generalization error of the pessimistic pruning (Eq. (10)) and the assumed error with the use of imprecise probabilities (Eq. (11)), we can observe that they estimate the real error of a data set by increasing the training error.

The difference is that pessimistic pruning labels a node as leaf if the sum of the estimated error for the descendant nodes is greater than for parent node, whereas Credal-C4.5 labels a node as leaf if the obtained data sets with the error estimation do not provide information gain when the node is split. Let us see an example of this difference.

Example 2. Let us suppose the subtree illustrated in Fig. 5 where a parent node is split in terms of the values of the variable X, the increment of error with the pessimistic pruning is equal to 0.5 and the value of the parameter s is also 0.5. With these conditions, if classic C4.5 is used, this subtree is initially created when there is information gain, that is,

$$IGR(Class, X) = \frac{IG(Class, X)}{H(X)} > 0.$$

This condition is fulfilled for this example because $H(X) = 1.915$ and

$$\begin{aligned} IG(Class, X) &= H(Class) - \sum_{i=1}^{4} P(X = x_i) H(Class|X = x_i) \\ &= 0.918 \\ &\quad - \left(\frac{12}{30}0.918 + \frac{7}{30}0.985 + \frac{3}{30}0.918 + \frac{8}{30}0.811\right) \\ &= 0.918 - 0.9051 > 0. \end{aligned}$$

After the creation of this subtree, the parent node is pruned by the pessimistic pruning because

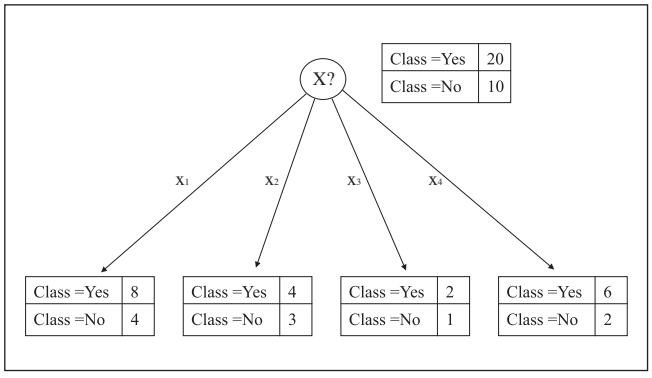

Fig. 5. Branching of a node in a decision tree.

$$e_{gen}(parent\_node) \leqslant \sum_{i=1}^{4} e_{gen}(descendant_i\_node),$$

that is,

$$(10 + 0.5) \leqslant (4 + 0.5) + (3 + 0.5) + (1 + 0.5) + (2 + 0.5).$$

On the other hand, this parent node is not expanded by using Credal-C4.5 when

$$IIGR^{\mathcal{D}}(Class, X) = \frac{IIG^{\mathcal{D}}(Class, X)}{H(X)} < 0.$$

This condition is fulfilled for this example because $H(X) = 1.915$ and

$$\begin{aligned} IIG^{\mathcal{D}}(Class, X) &= H^*(K^{\mathcal{D}}(Class)) - \sum_{i=1}^{4} P^{\mathcal{D}}(X = x_i) H^*(K^{\mathcal{D}}(Class|X = x_i)) \\ &= 0.928 \\ &\quad - \left(\frac{12}{30.5}0.942 + \frac{7.5}{30.5}0.996 + \frac{3}{30.5}0.985 + \frac{8}{30.5}0.873\right) \\ &= 0.928 - 0.941 < 0, \end{aligned}$$

where $\mathcal{D}$ is the data set composed by the 30 examples of Fig. 5 and $P^{\mathcal{D}}$ is the probability distribution chosen from Eq. (9), that is,

$$P^{\mathcal{D}} = \left(\frac{12}{30.5}, \frac{7.5}{30.5}, \frac{3}{30.5}, \frac{8}{30.5}\right).$$

We can observe with the above example that the processes of branching and pruning of a tree can be reduced by using the IIGR split criterion. We can also see that pessimistic pruning and Credal-C4.5 work with the data sets obtained after estimating the generalization error. Pessimistic pruning considers the sum of errors in order to prune a node, whereas Credal-C4.5 takes into account the information gain to label a node as leaf and to avoid one possible step of branching and pruning. Besides, this stopping criterion is not tight because Credal-C4.5 achieves good accuracy results. This fact will be shown in experimental section.

6. Experimental analysis

We present two experiments in this section.

(1) The aim of the first experiment is show that Credal-C4.5 improves to the best previously published credal decision tree, Complete Credal Decision Tree (CCDT) (Mantas & Abellán, 2014). CCDT is different to Credal-C4.5 because it uses the IIG measure for the split criterion instead of IIGR measure, it has not pruning process and the data sets are preprocessed before using CCDT (the missing values are replaced and the continuous variables are discretized).
(2) The second experiment studies the performance of Credal-C4.5 as opposed to classic C4.5. An algorithm similar to ID3, that we have called as MID3, is also implemented in order to carry out a more complete comparison. This MID3 algorithm is equal to classic C4.5 by replacing IGR measure by IG. Besides, all the variables available in a node are used for the split criterion.

CCDT was defined as algorithm without pruning in Mantas and Abellán (2014). For this reason, the results provided by C4.5, Credal-C4.5 and MID3 without pruning are used in the first experiment. In this way, all the algorithms are compared with the same experimental conditions. The second experiment is only focused on the methods with a pruning process. The trees built with the above mentioned algorithms will be referenced as C4.5, Credal-C4.5, MID3 and CCDT in this section.

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.

Data set	*N*	Feat	Num	Nom	*k*	Range
Anneal	898	38	6	32	6	2–10
Arrhythmia	452	279	206	73	16	2
Audiology	226	69	0	69	24	2–6
Autos	205	25	15	10	7	2–22
Balance-scale	625	4	4	0	3	–
Breast-cancer	286	9	0	9	2	2–13
Wisconsin-breast-cancer	699	9	9	0	2	–
Car	1728	6	0	6	4	3–4
CMC	1473	9	2	7	3	2–4
Horse-colic	368	22	7	15	2	2–6
Credit-rating	690	15	6	9	2	2–14
German-credit	1000	20	7	13	2	2–11
Dermatology	366	34	1	33	6	2–4
Pima-diabetes	768	8	8	0	2	–
Ecoli	366	7	7	0	7	–
Glass	214	9	9	0	7	–
Haberman	306	3	2	1	2	12
Cleveland-14-heart-disease	303	13	6	7	5	2–14
Hungarian-14-heart-disease	294	13	6	7	5	2–14
Heart-statlog	270	13	13	0	2	–
Hepatitis	155	19	4	15	2	2
Hypothyroid	3772	30	7	23	4	2–4
Ionosphere	351	35	35	0	2	–
Iris	150	4	4	0	3	–
kr-vs-kp	3196	36	0	36	2	2–3
Letter	20000	16	16	0	26	–
Liver-disorders	345	6	6	0	2	–
Lymphography	146	18	3	15	4	2–8
mfeat-pixel	2000	240	0	240	10	4–6
Nursery	12960	8	0	8	4	2–4
Optdigits	5620	64	64	0	10	–
Page-blocks	5473	10	10	0	5	–
Pendigits	10992	16	16	0	10	–
Primary-tumor	339	17	0	17	21	2–3
Segment	2310	19	16	0	7	–
Sick	3772	29	7	22	2	2
Solar-flare2	1066	12	0	6	3	2–8
Sonar	208	60	60	0	2	–
Soybean	683	35	0	35	19	2–7
Spambase	4601	57	57	0	2	–
Spectrometer	531	101	100	1	48	4
Splice	3190	60	0	60	3	4–6
Sponge	76	44	0	44	3	2–9
Tae	151	5	3	2	3	2
Vehicle	946	18	18	0	4	–
Vote	435	16	0	16	2	2
Vowel	990	11	10	1	11	2
Waveform	5000	40	40	0	3	–
Wine	178	13	13	0	3	–
Zoo	101	16	1	16	7	2

In order to check the above procedures, we used a broad and diverse set of 50 known data sets, obtained from the *UCI repository of machine learning data sets* which can be directly downloaded from http://archive.ics.uci.edu/ml. We took data sets that are different with respect to the number of cases of the variable to be classified, data set size, feature types (discrete or continuous) and number of cases of the features. A brief description of these can be found in Table 1.

We used *Weka* software (Witten & Frank, 2005) on Java 1.5 for our experimentation. The implementation of C4.5 algorithm provided by *Weka* software, called *J48*, was employed with its default configuration. We added the necessary methods to build Credal-C4.5 and MID3 trees with the same experimental conditions. The parameter of the IDM for the Credal-C4.5 algorithm was set to $s = 1.0$ (see Section 2.3). The minimum number of instances per leaf for branching was fixed to 2 for C4.5 and Credal-C4–5, as it is appears in the configuration by default for C4.5 (J48 in Weka). Data sets with missing values or continuous variables were processed according the procedures described in Section 2.2. We use the procedure for pruning that appears by default for C4.5 in Weka: the pessimistic pruning procedure.

On the other hand, by using *Weka's* filters, we added the following percentages of random noise to the class variable: 0%, 10% and 30%, only in the training data set. The procedure to introduce noise was the following for a variable: a given percentage of instances of the training data set was randomly selected, and then their current variable values were randomly changed to other possible values. The instances belonging to the test data set were left unmodified.

We repeated 10 times a 10-fold cross validation procedure for each data set.[2]

[2] The data set is separated in 10 subsets. Each one is used as a test set and the set obtained by joining the other 9 subsets is used as training set. So, we have 10 training sets and 10 test sets. This procedure is repeated 10 times with a previous random reordering. Finally, it produces 100 training sets and 100 test sets. The percentage of correct classifications for each data set, presented in tables, is the average of these 100 trials.

Table 2
Accuracy results of C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when are applied on data sets with percentage of random noise equal to 0%.

Dataset	C4.5	Credal-C4.5	MID3	CCDT
Anneal	98.57	98.19	98.99	99.34
Arrhythmia	64.05	67.55	64.89	67.08
Audiology	76.48	78.58	76.02	80.94
Autos	82.40	74.52	77.46	78.27
Balance-scale	79.44	78.00	79.45	69.59
Breast-cancer	68.15	71.44	68.34	72.03
Wisconsin-breast-cancer	94.37	95.08	94.56	94.74
Car	93.74	91.42	93.98	90.28
CMC	49.19	52.01	49.58	48.70
Horse-colic	82.09	84.64	82.91	83.17
Credit-rating	82.17	85.46	81.26	84.06
German-credit	68.11	70.17	69.68	69.53
Dermatology	94.04	93.82	92.20	94.58
Pima-diabetes	73.87	73.19	73.79	74.22
Ecoli	82.53	81.90	83.07	80.03
Glass	67.76	63.66	67.90	68.83
Haberman	70.52	73.89	70.62	73.59
Cleveland-14-heart-disease	76.44	76.60	78.62	76.23
Hungarian-14-heart-disease	78.55	82.54	76.12	78.62
Heart-statlog	76.78	80.04	77.93	82.11
Hepatitis	78.59	79.84	78.82	80.32
Hypothyroid	99.51	99.53	99.56	99.37
Ionosphere	89.83	88.35	88.15	89.75
Iris	94.80	94.73	94.80	93.73
kr-vs-kp	99.44	99.40	99.42	99.49
Letter	88.02	87.57	88.00	77.55
Liver-disorders	65.37	64.18	65.75	56.85
Lymphography	75.42	78.51	73.42	74.50
mfeat-pixel	78.42	79.58	75.66	80.31
Nursery	98.69	96.30	98.64	96.31
Optdigits	90.48	90.77	91.10	79.33
Page-blocks	96.78	96.72	96.92	96.26
Pendigits	96.54	96.39	96.39	88.87
Primary-tumor	42.60	42.19	38.59	38.73
Segment	96.80	96.04	96.77	94.18
Sick	98.77	98.77	98.82	97.80
Solar-flare2	99.49	99.53	99.38	99.46
Sonar	73.42	71.47	73.53	73.92
Soybean	90.69	92.50	86.76	92.21
Spambase	92.42	92.61	92.83	91.85
Spectrometer	47.31	45.52	43.49	45.22
Splice	92.16	93.81	91.37	93.17
Sponge	91.68	94.11	92.70	94.63
Tae	58.60	53.20	58.21	46.78
Vehicle	72.18	72.84	72.67	69.53
Vote	95.76	96.04	95.56	96.18
Vowel	81.63	77.87	84.09	75.60
Waveform	75.12	76.05	75.70	74.44
Wine	93.20	92.13	93.83	92.08
Zoo	93.41	92.42	92.01	95.83
Average	82.13	82.23	81.81	81.00

Table 3
Accuracy results of C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when are applied on data sets with percentage of random noise equal to 10%.

Dataset	C4.5	Credal-C4.5	MID3	CCDT
Anneal	96.05	97.84	96.31	97.87
Arrhythmia	59.81	64.77	57.65	63.90
Audiology	75.06	76.98	71.43	76.40
Autos	74.73	71.57	68.88	74.50
Balance-scale	76.56	78.46	76.35	71.90
Breast-cancer	66.31	68.68	64.85	68.98
Wisconsin-breast-cancer	92.00	94.29	92.20	92.93
Car	88.80	90.90	88.47	90.41
CMC	47.25	50.07	47.30	47.66
Horse-colic	79.74	83.52	78.93	79.24
Credit-rating	76.80	85.04	75.80	82.87
German-credit	65.23	69.01	65.94	68.69
Dermatology	88.00	92.66	85.73	92.10
Pima-diabetes	72.04	73.58	72.20	73.35
Ecoli	77.77	81.52	77.89	79.67
Glass	64.07	65.29	63.40	67.43
Haberman	68.86	73.40	69.02	72.74
Cleveland-14-heart-disease	72.93	76.31	73.83	75.87
Hungarian-14-heart-disease	75.73	80.73	73.73	77.70
Heart-statlog	72.33	77.67	72.48	79.70
Hepatitis	73.84	78.17	75.27	77.37
Hypothyroid	95.55	99.38	95.74	99.05
Ionosphere	86.30	87.30	85.31	85.95
Iris	89.73	93.60	89.20	93.67
kr-vs-kp	93.51	98.04	93.05	97.39
Letter	85.01	86.27	84.45	76.45
Liver-disorders	62.15	61.11	62.27	56.85
Lymphography	71.31	74.10	69.59	73.57
mfeat-pixel	71.79	76.88	67.51	76.43
Nursery	91.59	96.23	91.39	96.13
Optdigits	82.49	87.49	83.01	75.14
Page-blocks	93.95	96.67	94.06	96.18
Pendigits	89.66	95.19	89.62	87.80
Primary-tumor	38.70	39.53	37.85	36.99
Segment	90.09	95.02	90.31	93.58
Sick	95.68	98.07	95.53	97.49
Solar-flare2	98.22	99.50	98.32	99.44
Sonar	67.56	70.53	69.34	71.51
Soybean	86.85	91.70	79.55	90.04
Spambase	89.87	91.56	89.38	88.20
Spectrometer	42.63	43.01	38.78	42.43
Splice	81.23	90.87	80.30	90.18
Sponge	83.00	89.07	84.80	90.73
Tae	52.23	49.01	52.55	45.13
Vehicle	66.17	69.63	65.72	67.77
Vote	92.20	94.39	92.52	94.57
Vowel	78.19	75.15	78.42	74.12
Waveform	69.10	74.96	69.07	70.81
Wine	86.17	89.22	86.28	90.64
Zoo	91.99	92.10	91.49	92.07
Average	77.74	80.72	77.06	79.23

Following the recommendation of Demsar (2006), we used a series of tests to compare the methods.[3] We used the following tests to compare multiple classifiers on multiple data sets, with a level of significance of $\alpha = 0.1$:

Friedman test (Friedman, 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, we can compare all the algorithms to each other using the **Nemenyi test** (Nemenyi, 1963).

[3] All the tests were carried out using *Keel* software (Alcalá-Fdez et al., 2009), available at www.keel.es.

6.1. Experiment 1: methods without pruning

6.1.1. Results

Next, it is shown the results obtained by C4.5, Credal-C4.5, MID3 and CCDT trees. Tables 2–4 present the accuracy results of each method without post-pruning procedure, applied on data sets with a percentage of random noise to the class variable equal to 0%, 10% and 30%, respectively.

Tables 5 and 6 present the average result of accuracy and tree size (number of nodes) for each method without pruning when is applied to data sets with percentages of random noise equal to 0%, 10% and 30%.

Table 7 shows Friedman's ranks obtained from the accuracy results of C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when they are applied on data sets with percentages of random noise equal to 0%, 10% and 30%. We remark that the null hypothesis is rejected in all the cases with noise.

Table 4
Accuracy results of C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when are applied on data sets with percentage of random noise equal to 30%.

Dataset	C4.5	Credal-C4.5	MID3	CCDT
Anneal	80.49	90.84	79.84	88.92
Arrhythmia	46.55	59.71	44.83	57.13
Audiology	63.93	69.45	53.39	64.90
Autos	56.70	58.40	52.19	62.23
Balance-scale	64.80	73.61	64.40	73.21
Breast-cancer	58.76	62.71	58.58	60.94
Wisconsin-breast-cancer	84.59	91.45	84.57	86.51
Car	73.92	82.80	73.42	82.88
CMC	42.96	45.23	43.16	43.94
Horse-colic	68.39	73.91	66.06	63.56
Credit-rating	64.43	73.03	63.61	68.41
German-credit	57.98	60.57	59.41	61.47
Dermatology	69.16	79.93	67.52	77.02
Pima-diabetes	68.87	69.43	68.45	69.31
Ecoli	64.14	77.71	64.14	78.28
Glass	52.51	59.82	52.29	66.31
Haberman	63.93	66.44	63.89	67.90
Cleveland-14-heart-disease	59.97	69.09	60.00	69.39
Hungarian-14-heart-disease	70.80	79.94	66.22	75.27
Heart-statlog	63.07	71.19	62.56	72.74
Hepatitis	62.19	70.41	62.17	67.10
Hypothyroid	79.54	97.45	80.49	96.12
Ionosphere	77.87	79.84	77.13	71.17
Iris	78.60	88.27	78.00	89.87
kr-vs-kp	72.22	81.16	72.38	80.22
Letter	71.74	77.44	71.36	68.92
Liver-disorders	56.72	55.37	57.00	56.85
Lymphography	56.50	63.43	54.91	61.28
mfeat-pixel	57.78	64.57	53.55	62.63
Nursery	72.18	88.95	71.74	88.98
Optdigits	62.98	73.59	63.82	61.15
Page-blocks	83.80	96.04	83.06	94.01
Pendigits	69.88	86.66	70.11	77.42
Primary-tumor	34.69	34.90	33.66	31.92
Segment	70.66	89.86	71.52	83.44
Sick	87.65	94.96	87.82	92.41
Solar-flare2	91.35	96.90	91.94	96.56
Sonar	60.74	63.34	61.06	61.98
Soybean	73.03	86.32	58.82	79.76
Spambase	85.42	87.44	84.47	72.88
Spectrometer	31.85	35.40	29.12	35.51
Splice	62.46	68.00	61.85	67.62
Sponge	63.52	71.29	62.93	72.27
Tae	45.73	43.31	44.92	41.48
Vehicle	53.09	62.99	53.15	59.92
Vote	79.08	84.45	78.71	82.55
Vowel	64.68	65.00	64.13	65.72
Waveform	57.22	69.84	56.50	58.51
Wine	70.01	82.40	70.14	79.20
Zoo	84.06	85.49	85.81	82.50
Average	65.86	73.21	64.82	70.60

Table 5
Average result of accuracy for C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	82.13	77.74	65.86
Credal-C4.5	82.23	80.72	73.21
MID3	81.81	77.06	64.82
CCDT	81.00	79.23	70.60

Tables 8 and 9 show the p-values of the Nemenyi test obtained from the accuracy results of the methods C4.5, Credal-C4.5, MID3 and CCDT when they are applied on data sets with a percentage of random noise 10% and 30%. For 0% of noise, the null hypothesis is not rejected. In all the caseS, Nemenyi procedure rejects the hypotheses that have a p-value $\leqslant 0.016667$.

Table 6
Average result about tree size for C4.5, Credal-C4.5, MID3 and CCDT when are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	216.98	376.37	672.13
Credal-C4.5	138.78	167.09	317.92
MID3	216.15	373.97	662.42
CCDT	387.18	523.09	1033.15

Table 7
Friedman's ranks for $\alpha = 0.1$ obtained from the accuracy results of C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when they are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	2.52	3.10	3.11
Credal-C4.5	2.30	1.34	1.38
MID3	2.54	3.32	3.45
CCDT	2.64	2.24	2.06

Table 8
p-Values of the Nemenyi test with $\alpha = 0.1$ obtained from the accuracy results of the methods C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when they are applied on data sets with percentage of random noise equal to 10%. Nemenyi procedure rejects those hypotheses that have a p-value $\leqslant 0.016667$.

i	algorithms	p-Values
6	Credal-C4.5 vs. MID3	0
5	C4.5 vs. Credal-C4.5	0
4	MID3 vs. CCDT	0.000029
3	Credal-C4.5 vs. CCDT	0.000491
2	C4.5 vs. CCDT	0.000866
1	C4.5 vs. MID3	0.394183

Table 9
p-Values of the Nemenyi test with $\alpha = 0.1$ obtained from the accuracy results of the methods C4.5, Credal-C4.5, MID3 and CCDT (without pruning) when they are applied on data sets with percentage of random noise equal to 30%. Nemenyi procedure rejects those hypotheses that have a p-value $\leqslant 0.016667$.

i	algorithms	p-Values
6	Credal-C4.5 vs. MID3	0
5	C4.5 vs. Credal-C4.5	0
4	MID3 vs. CCDT	0
3	C4.5 vs. CCDT	0.000048
2	Credal-C4.5 vs. CCDT	0.008448
1	C4.5 vs. MID3	0.187901

6.1.2. Comments about the results

The objective of this section is to compare CCDT algorithm with the rest of algorithms without pruning. In particular, it is important the comparison between Credal-C4.5 without pruning and CCDT in order to show the improvement of the credal trees. The results of the previous section are analyzed according the following aspects: Average accuracy, Tree size, Friedman's ranking and Nemenyi test.

- **Average accuracy**: According to this factor, we can say that without noise all the tree methods are nearly equivalent, being some better the performance of the Credal-C4.5 without pruning. When noise is added there is a notable difference in favor of Credal-C4.5 without pruning respect to the rest ones. This difference is important when the level of noise is 30% On the other hand, CCDT presents the worst result for data without noise. However, CCDT has better average accuracy results than C4.5 and MID3 (without pruning) for data with noise.

Table 10
Accuracy results of C4.5, Credal-C4.5 and MID3 (with pruning) when are applied on data sets with percentage of random noise equal to 0%.

Dataset	C4.5	Credal-C4.5	MID3
Anneal	98.57	98.36	98.99
Arrhythmia	65.65	67.68	65.15
Audiology	77.26	78.94	76.91
Autos	81.77	74.57	78.24
Balance-scale	77.82	77.33	77.69
Breast-cancer	74.28	74.84	71.75
Wisconsin-breast-cancer	95.01	95.12	95.35
Car	92.22	91.16	93.02
CMC	51.44	52.80	52.06
Horse-colic	85.16	85.18	84.34
Credit-rating	85.57	85.43	84.03
German-credit	71.25	71.34	71.98
Dermatology	94.10	94.26	93.49
Pima-diabetes	74.49	74.15	74.39
Ecoli	82.83	81.60	83.61
Glass	67.63	63.61	67.67
Haberman	72.16	71.18	72.03
Cleveland-14-heart-disease	76.94	76.53	79.30
Hungarian-14-heart-disease	80.22	82.33	76.77
Heart-statlog	78.15	80.33	78.81
Hepatitis	79.22	79.79	80.33
Hypothyroid	99.54	99.52	99.56
Ionosphere	89.74	88.18	88.04
Iris	94.73	94.73	94.73
kr-vs-kp	99.44	99.45	99.42
Letter	88.03	87.58	87.97
Liver-disorders	65.84	64.53	66.16
Lymphography	75.84	78.31	75.01
mfeat-pixel	78.66	79.76	77.12
Nursery	97.18	96.30	97.10
Optdigits	90.52	90.83	91.10
Page-blocks	96.99	96.69	97.09
Pendigits	96.54	96.42	96.39
Primary-tumor	41.39	42.33	39.92
Segment	96.79	96.04	96.74
Sick	98.72	98.79	98.85
Solar-flare2	99.53	99.53	99.53
Sonar	73.61	71.37	73.53
Soybean	91.78	92.40	89.94
Spambase	92.68	92.56	93.11
Spectrometer	47.50	45.54	43.37
Splice	94.17	94.04	93.57
Sponge	92.50	92.50	92.50
Tae	57.41	53.26	57.62
Vehicle	72.28	72.78	72.71
Vote	96.57	96.59	96.11
Vowel	80.20	77.88	83.63
Waveform	75.25	76.07	75.83
Wine	93.20	92.13	93.83
Zoo	92.61	92.42	92.01
Average	82.62	82.30	82.37

Table 11
Accuracy results of C4.5, Credal-C4.5 and MID3 (with pruning) when are applied on data sets with percentage of random noise equal to 10%.

Dataset	C4.5	Credal-C4.5	MID3
Anneal	98.37	98.23	98.42
Arrhythmia	62.54	65.76	58.44
Audiology	77.53	77.39	72.70
Autos	74.72	71.65	69.61
Balance-scale	78.11	78.26	77.82
Breast-cancer	71.13	72.07	70.75
Wisconsin-breast-cancer	93.72	94.28	94.06
Car	90.92	90.53	90.74
CMC	49.95	51.36	50.36
Horse-colic	84.61	85.10	84.50
Credit-rating	84.78	85.23	84.22
German-credit	71.18	71.38	71.72
Dermatology	93.31	93.12	91.06
Pima-diabetes	72.37	73.83	72.56
Ecoli	81.87	81.49	82.04
Glass	65.37	65.57	64.55
Haberman	72.32	72.39	72.29
Cleveland-14-heart-disease	75.78	76.94	77.56
Hungarian-14-heart-disease	79.78	80.94	77.03
Heart-statlog	75.63	78.41	76.04
Hepatitis	77.88	80.19	78.62
Hypothyroid	99.40	99.44	99.43
Ionosphere	86.90	87.04	85.79
Iris	92.73	93.53	92.47
kr-vs-kp	98.97	98.95	98.80
Letter	86.74	86.67	86.38
Liver-disorders	62.38	61.69	62.73
Lymphography	75.11	74.78	76.53
mfeat-pixel	76.77	77.97	74.36
Nursery	96.29	96.08	96.00
Optdigits	88.47	88.94	88.86
Page-blocks	96.70	96.78	96.79
Pendigits	95.37	95.49	95.20
Primary-tumor	39.59	40.39	40.09
Segment	95.06	95.17	95.03
Sick	98.22	98.24	98.22
Solar-flare2	99.53	99.53	99.53
Sonar	67.56	70.39	69.34
Soybean	90.54	91.74	85.85
Spambase	90.96	91.52	90.57
Spectrometer	43.20	43.07	39.64
Splice	93.05	93.08	92.48
Sponge	91.80	91.66	92.50
Tae	50.77	49.01	51.61
Vehicle	68.51	69.99	68.26
Vote	95.74	95.45	95.28
Vowel	77.13	75.26	78.37
Waveform	69.51	75.13	69.50
Wine	87.35	89.39	87.36
Zoo	92.39	92.10	92.19
Average	80.77	81.25	80.29

- **Tree size**: We can observe that always (with an without noise) Credal-C4.5 built the smallest trees. The average number of nodes of this method is nearly the half of the number of nodes of the rest of methods. C4.5 and MID3 (without pruning) have a similar average tree size. Finally, CCDT presents the highest average tree size, because CCDT carries out a multi-interval discretization of the continuous variables (Fayyad & Irani, 1993) and the rest of methods make a discretization by using only two intervals.
- **Friedman's ranking**: According to this ranking, Credal-C4.5 without pruning is the best model to classify data sets in all the cases. For data with noise, this difference is important. In particular, Credal-C4.5 without pruning improves to CCDT for data with or without noise. Finally, MID3 and C4.5 without pruning have the worst rank for data with noise.
- **Nemenyi test**: This test is not carried out for the case of 0% of noise because the Friedman's test does not reject the null hypothesis, that is, the differences are not significant for data without noise. For data with noise (10% and 30%), this test indicates that Credal-C4.5 is the best model with significant statistical difference. This test also says that CCDT improves to MID3 and C4.5 (without pruning) for data without noise.

After this analysis we can conclude the following comments about comparison between algorithms:

- **Credal-C4.5 (without pruning) vs. CCDT**: Credal-C4.5 without pruning obtains better average accuracy results and Friedman's rank than CCDT. According Nemenyi test, this difference is significant for data with noise. Besides, Credal-C4.5 without pruning builds trees smaller than CCDT. Hence, we can conclude with this experiment that Credal-C4.5 algorithm without pruning improves to the best previously published credal decision tree.

Table 12
Accuracy results of C4.5, Credal-C4.5 and MID3 (with pruning) when are applied on data sets with percentage of random noise equal to 30%.

Dataset	C4.5	Credal-C4.5	MID3
Anneal	96.03	95.85	95.24
Arrhythmia	49.15	62.06	45.09
Audiology	70.88	70.68	60.25
Autos	57.92	60.35	53.81
Balance-scale	74.16	75.02	73.52
Breast-cancer	68.65	67.61	67.49
Wisconsin-breast-cancer	89.24	92.27	89.43
Car	86.00	85.97	85.89
CMC	46.39	47.70	45.59
Horse-colic	79.63	80.48	75.00
Credit-rating	74.58	81.41	71.77
German-credit	63.09	63.70	66.05
Dermatology	87.64	88.95	86.56
Pima-diabetes	69.39	69.67	68.93
Ecoli	75.27	79.78	73.63
Glass	55.23	60.49	54.69
Haberman	68.83	72.85	68.87
Cleveland-14-heart-disease	68.00	71.57	67.97
Hungarian-14-heart-disease	78.16	80.81	74.68
Heart-statlog	65.52	72.33	64.70
Hepatitis	68.15	73.36	68.63
Hypothyroid	98.59	98.96	98.41
Ionosphere	78.18	80.04	77.30
Iris	84.00	89.00	84.07
kr-vs-kp	91.13	90.97	90.53
Letter	82.13	82.54	81.62
Liver-disorders	56.83	55.45	57.06
Lymphography	66.33	68.11	68.59
mfeat-pixel	71.98	73.19	68.43
Nursery	93.99	94.30	93.46
Optdigits	76.91	80.77	70.24
Page-blocks	94.91	96.25	94.81
Pendigits	89.21	92.25	88.02
Primary-tumor	37.67	37.76	38.44
Segment	85.35	91.92	84.33
Sick	95.20	97.14	95.29
Solar-flare2	99.53	99.49	99.53
Sonar	60.84	63.34	61.10
Soybean	88.45	89.34	72.78
Spambase	86.07	87.69	85.32
Spectrometer	33.02	35.61	29.72
Splice	81.21	80.06	81.85
Sponge	88.84	86.71	92.50
Tae	45.86	43.64	45.26
Vehicle	56.06	63.50	55.56
Vote	90.99	91.55	91.38
Vowel	66.01	65.61	64.16
Waveform	57.32	70.08	56.59
Wine	71.02	82.91	70.98
Zoo	87.65	87.74	89.05
Average	74.14	76.58	72.88

Table 13
Average result of accuracy for C4.5, Credal-C4.5 and MID3 when are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	82.62	80.77	74.14
Credal-C4.5	82.30	81.25	76.58
MID3	82.37	80.29	72.88

Table 14
Average result about tree size for C4.5, Credal-C4.5 and MID3 when are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	156.54	170.02	244.05
Credal-C4.5	122.67	131.06	171.39
MID3	155.83	170.03	253.73

Table 15
Friedman's ranks for $\alpha = 0.1$ obtained from the accuracy results of C4.5, Credal-C4.5 and MID3 (with pruning) when they are applied on data sets with percentage of random noise equal to 0%, 10% and 30%.

Tree	Noise 0%	Noise 10%	Noise 30%
C4.5	1.90	2.07	2.07
Credal-C4.5	2.08	1.60	1.40
MID3	2.02	2.33	2.53

Table 16
p-Values of the Nemenyi test with $\alpha = 0.1$ obtained from the accuracy results of the methods C4.5, Credal-C4.5 and MID3 (with pruning) when they are applied on data sets with percentage of random noise equal to 10%. Nemenyi procedure rejects those hypotheses that have a p-value$\leqslant 0.033333$.

i	algorithms	p-Values
3	Credal-C4.5 vs. MID3	0.000262
2	C4.5 vs. Credal-C4.5	0.018773
1	C4.5 vs. MID3	0.193601

Table 17
p-Values of the Nemenyi test with $\alpha = 0.1$ obtained from the accuracy results of the methods C4.5, Credal-C4.5 and MID3 (with pruning) when they are applied on data sets with percentage of random noise equal to 30%. Nemenyi procedure rejects those hypotheses that have a p-value$\leqslant 0.033333$.

i	algorithms	p-Values
3	Credal-C4.5 vs. MID3	0
2	C4.5 vs. Credal-C4.5	0.000808
1	C4.5 vs. MID3	0.021448

- **Credal trees (CCDT and Credal-C4.5 without pruning) vs. Classic methods without pruning (C4.5 and MID3)**: We can observe with this experiment that credal trees improve the results obtained by the classic methods (without pruning) for data with noise. According Nemenyi test, this difference is significant. Hence, we can conclude that the pruning process is a fundamental part of the classic methods (C4.5 and MID3) to improve their accuracy. For this reason, we have focused on the methods with a pruning process in the next experiment.

6.2. Experiment 2: methods with pruning

6.2.1. Results

Next, it is shown the results obtained by C4.5, Credal-C4.5 and MID3 trees. Tables 10–12 present the accuracy results of each method with post-pruning procedure, applied on data sets with a percentage of random noise to the class variable equal to 0%, 10% and 30%, respectively.

Tables 13 and 14 present the average result of accuracy and tree size (number of nodes) for each method when is applied to data sets with percentages of random noise equal to 0%, 10% and 30%.

Table 15 shows Friedman's ranks obtained from the accuracy results of C4.5, Credal-C4.5 and MID3 (with pruning) when they are applied on data sets with percentages of random noise equal to 0%, 10% and 30%. We remark that the null hypothesis is rejected in all the cases with noise.

Tables 16, 17 show the p-values of the Nemenyi test obtained from the accuracy results of the methods C4.5, Credal-C4.5 and MID3 when they are applied on data sets with a percentage of random noise 10% and 30%. For 0% of noise, the null hypothesis is not rejected. In all the cases with noise, Nemenyi procedure rejects the hypotheses that have a p-value$\leqslant 0.033333$.

6.2.2. Comments about the results

The principal aim of this section is to compare the methods when a pruning process is used. In this case, all the methods improve their percentage of accuracy. The results shown in the previous section about the methods with pruning process are analyzed according the following aspects: Average accuracy, Tree size, Friedman's ranking and Nemenyi test.

- **Average accuracy**: According to this factor, we can observe that C4.5 obtains better average results of accuracy than Credal-C4.5 and MID3 for data without noise. On the other hand, Credal-C4.5 obtains the best average result of accuracy for data with noise. This difference is notable when a 30% of noise is added.
- **Tree size**: Credal-C4.5 obtains the smallest average tree size in all the cases (with and without noise). It can be remarked that always the Credal-C4.5 has an average of number of nodes that is less of the half of the average of the rest of methods.
- **Friedman's ranking**: According to this ranking, C4.5 obtains the lower rank for data without noise and Credal-C4.5 is the best model for classifying data sets with noise (lower rank value in each comparison). This test says that the null hypothesis can not be rejected without noise but with noise, always is rejected, i.e all the procedures performs similar without noise but not with noise. In the case of noise, always the Credal-C4.5 is better.
- **Nemenyi test**: This test is not carried out for the case of 0% of noise because the Friedman's test does not reject the null hypothesis. According to this test, we can observe that in the case of 10% of noise the Credal-C4.5 is statistically better than MID3 and C4.5 for a 0.1 level of significance.[4] C4.5 is not better than MID3 using this test with similar level of significance. When the level of noise is increased to 30% the test carried out expresses similar conclusions, but, in this case, the differences in favor of Credal-C4.5 are stronger (see the *p*-values of Table 17). For this level of noise, C4.5 is also statistically better than MID3 using the Nemenyi test.

We can point out the following comments:

- **C4.5 vs. MID3**: C4.5 obtains better average results of accuracy than MID3. Also, it has a Frieman's rank smaller than the one of MID3 in all the cases (data sets with or without noise, before or after pruning). However, according Nemenyi test, these differences are not statistically significant, except in the case of pruning and 30% of noise. On the other hand, the average tree size are very similar for these two methods.
- **Credal-C4.5 vs. classic methods (C4.5 and MID3)**: For data without noise: All the methods have a similar performance, with and without a pruning process. Only can be remarked the difference about the size of the trees built: Credal-C4.5 obtains the smallest average tree size.
  For data with noise: Credal-C4.5 obtains always (with and without pruning) better average results of accuracy than C4.5 and MID3; it obtains the lowest Friedman's rank; and, according to the Nemenyi test, these differences are statistically significant. Besides, Credal-C4.5 presents the smallest average tree size in all the cases of noise.

The above points allow us to remark the following conclusions about the experimental study:

(i) If we are interested to obtain smaller trees a with similar level of accuracy, Credal-C4.5 is more adequate than methods based on classic probabilities.

(ii) The use of Credal-C4.5 is especially suitable to be applied on data sets with noise. This conclusion is reasonable from the definition of Credal-C4.5. This method was defined with the assumption that the training set is not very reliable. Imprecise probabilities were used to estimate the values of the features and class variable. Hence, a very appropriate method is obtained for noisy data.

[4] Also for a stronger level of 0.075.

7. Conclusion and future works

We have presented a new model called Credal-C4.5, a modified version of the C4.5 algorithm. It has been defined by using a mathematical theory of imprecise probabilities and uncertainty measures on credal sets. Hence, the imprecision of the data is taken into account in the new method. With this modification, a data set is considered unreliable when the variable selection process is carried out. The pruning process of the C4.5 algorithm takes into account the same consideration. Hence, Credal-C4.5 as opposed to C4.5 assumes the same hypothesis on the data set when the tree is created and pruning. Credal-C4.5 with this new characteristic is especially suitable when noisy data sets are classified. Relevant differences in the performance of both methods are also shown.

In a first experimental study, we have compared a previous method which takes into account the imprecision of the data too with the new method Credal-C4.5, and we show that the new method beats to the previous one in all the comparative studies (with and without noise). In a second experimental study, we have compared Credal-C4.5, C4.5 and a modified and improved version of the known ID3, called MID3. We have showed that with no noise is added, all the methods are very similar in performance, and the only difference among them is that Credal-C4.5 presents trees with a notable lower number of nodes. When noise is added, Credal-C4.5 has a better performance than the one of the rest of methods, and, in this case, also the number of nodes of Credal-C4.5 is notably lower than the ones or the rest of methods.

Data sets obtained from real applications are not totally clean. Usually, they have some level of noise. We think that it can be very interesting to apply Credal-C4.5 algorithm on data sets of real applications, to analyze results and to extract knowledge about the application from the credal tree. New mathematical models, procedures and split criteria, as the ones of Abellán, Baker, Coolen, Crossman, and Masegosa (2014) and Abellán (2013a, 2013b) can be checked on data sets with noise. These tasks are proposed as future works.

Acknowledgment

This work has been supported by the Spanish "Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía" under Project TIC-6016 and Spanish MEC project TIN2012-38969.

References

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

Abellán, J., & Moral, S. (2005). Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning, 39*(2–3), 235–255.

Abellán, J. (2006). Uncertainty measures on probability intervals from Imprecise Dirichlet model. *International Journal of General Systems, 35*(5), 509–528.

Abellán, J., & Moral, S. (2006). An algorithm that computes the upper entropy for order-2 capacities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 14*(2), 141–154.

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

Abellán, J., & Masegosa, A. (2008). Requirements for total uncertainty measures in Dempster–Shafer theory of evidence. *International Journal of General Systems, 37*(6), 733–747.

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*(6), 833–854.

Abellán, J., & Masegosa, A. R. (2009). An experimental study about simple decision trees for Bagging ensemble on data sets with classification noise. In C. Sossai & G. Chemello (Eds.), *ECSQARU. LNCS* (Vol. 5590, pp. 446–456). Springer.

Abellán, J., & Masegosa, A. (2012). Bagging schemes on the presence of noise in classification. *Expert Systems with Applications, 39*(8), 6827–6837.

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

Abellán, J. (2013b). An application of non-parametric predictive inference on multi-class classification high-level-noise problems. *Expert Systems with Applications, 40*, 4585–4592.

Abellán, J., Baker, R. M., Coolen, F. P. A., Crossman, R., & Masegosa, A. (2014). Classification with decision trees from a nonparametric predictive inference perspective. *Computational Statistics and Data Analysis, 71*, 789–802.

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 41*, 3825–3830.

Demsar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Fayyad, U. M., & Irani, K. B. (1993). Multi-valued interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th international joint conference on artificial intelligence* (pp. 1022–1027). San Mateo: Morgan Kaufman.

Frenay, B., & Verleysen, M. (in press). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. <http://dx.doi.org/10.1109/TNNLS.2013.2292894>.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics, 11*, 86–92.

Alcalá-Fdez, J., Sánchez, L., García, S., Del Jesus, M. J., Ventura, S., Garrell, J. M., et al. (2009). KEEL: A software tool to assess evolutionary algorithms to data mining problems. *Soft Computing, 13*(3), 307–318.

Klir, G. J. (2006). *Uncertainty and information. Foundations of generalized information theory*. Hoboken, NJ: John Wiley.

Mantas, C. J., & Abellán, J. (2014). Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data. *Expert Systems with Applications, 41*, 2514–2525.

Nemenyi, P. B. (1963). *Distribution-free multiple comparison* (Ph.D. thesis). Princenton University.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Quinlan, J. R. (1999). Programs for machine learning. Morgan Kaufmann series in machine learning.

Rokach, L., & Maimon, O. (2010). Classification trees. Data mining and knowledge discovery handbook (pp. 149–174).

Walley, P. (1996). Inferences from multinomial data, learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B, 58*, 3–57.

Wang, Y. (2010). Imprecise probabilities based on generalised intervals for system reliability assessment. *International Journal of Reliability and Safety, 4*(30), 319–342.

Witten, I. H., & Frank, E. (2005). *Data mining, practical machine learning tools and techniques* (2nd edition.). San Francisco: Morgan Kaufman.

Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning, 24*(2-3), 149–170.