Ensembles of decision trees based on imprecise probabilities and uncertainty measures

Joaquín Abellán
Department of Computer Science and Artificial Intelligence, ETS de Ingenierı́as Informática y de Telecomunicación, University of Granada, 18071 Granada, Spain

A R T I C L E   I N   F O

Article history:
Received 8 February 2011
Received in revised form 14 March 2012
Accepted 15 March 2012
Available online xxxx

Keywords:
Imprecise probabilities
Credal sets
Uncertainty measures
Supervised classification
Decision trees
Ensemble methods

A B S T R A C T

In this paper, we present an experimental comparison among different strategies for combining decision trees built by means of imprecise probabilities and uncertainty measures. It has been proven that the combination or fusion of the information obtained from several classifiers can improve the final process of the classification. We use previously developed schemes, known as Bagging and Boosting, along with a new one based on the variation of the root node via the information rank of each feature of the class variable. To this end, we applied two different approaches to deal with missing data and continuous variables. We use a set of tests on the performance of the methods analyzed here, to show that, with the appropriate approach, the Boosting scheme constitutes an excellent way to combine this type of decision tree. It should be noted that it provides good results, even compared with a standard Random Forest classifier, a successful procedure very commonly used in the literature.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

In the area of machine learning, supervised classification learning can be considered an important tool for decision support. Classification can be defined as a machine learning technique used to predict group membership for data instances. It can be applied to decision support in medicine, character recognition, astronomy, banking and other fields. A classifier may be represented using a Bayesian network, a neural network, a decision tree, etc. In the literature, it can be seen that the information obtained by a single classifier can be improved with the combination or fusion of several classifiers belonging to the same type or to different ones.

A decision tree, also known as a classification tree, is a simple structure that can be used as a classifier. An important aspect of decision trees is their inherent instability which means that they are sensitive to changes in the training examples, and significantly different hypotheses are therefore generated for different training datasets [1]. Compared with single classifiers, the strategies used to combine classifiers are known to increase classifier accuracy, providing them combine different models. Decision trees built from different training sub-samples from a given problem domain will produce quite different models. This characteristic is essential with regard to considering them as suitable classifiers in an ensemble scheme such as Bagging [2], Boosting [1] or Random Forest [3]. Ensembles of decision tree classifiers (also called forests in the literature), are well accepted in the data mining community. This is principally motivated by a significant improvement in accuracy when they are applied to artificial and real-world datasets.

Abellán and Moral [4] presented a method for building decision trees using the study on measures of information/uncertainty in general convex sets of probability distributions (see [5]). In this method, the split criterion used, called Imprecise Info-Gain (IIG) is different to the characteristics of classical split criteria [4,6,7]. An important characteristic of this criterion as opposed to the classic ones is that it can produce negative information values for the class variable, i.e. the value of the information on the class variable using a feature can be lower than the value obtained for the class variable alone. We label this type of trees credal decision trees1 or to simplify, credal trees (CTs).

In this paper, we perform a set of experiments on a wide range of datasets in order to check the performance of credal trees for different ensemble schemes: Bagging, Boosting, and IRN schemes. The latter scheme of IRN is the method recently presented in Abellán and Masegosa [8], which uses the IIG criterion in two important steps of the procedure: to select the set of informative features used as root nodes of the decision trees in the combination scheme;2 and as a split criterion for building each decision tree. The experimental study by Abellán and Masegosa [8] shows that

---

1 The term credal used here refers to the use of convex sets of probability distribution, i.e. credal sets, for building precise classifiers [see (6)]. It differs from its use in imprecise classification (see [9]).
2 Since the main characteristic of this method is the use of Informative Root Nodes, we have termed this method IRN-Ensemble or IRN to simplify.

---
IRN, with a standard previous preprocessing approach, performs well compared to other classic ensemble methods that use classical decision trees. The method does not need to set the number of trees to be used; rather, it uses a number of decision trees which depends on the number of informative features (using IIG criterion) in relation to the class variable. These informative features are used as root nodes.

We study the performance of the following methods: the original IRN method, Bagging CTs (noted as BAGCT), and Boosting CTs (noted as BOOC). To compare the results, we also use the following methods: the IRN method replacing its final procedure with a vote procedure (noted as IRNV), Bagging decision trees built with Quinlan’s classic Information Gain (Info-Gain) split criterion [10] (noted as BAGIG), Boosting decision trees built with the Info-Gain split criterion (noted as BOOC), and the Random Forest method.

To check the accuracy of the above-mentioned methods, we used two different approaches to deal with missing data and continuous variables. We show that for two of the classification methods used here, the type of approach can be strongly related to their accuracy. We compare the performance of the classification methods analyzed by conducting a set of tests, including a Friedman rank [11,12] on the methods used in each approach, and a set of known post hoc tests. The most important outcome of this experimental study is that the Boosting scheme, with the appropriate approach, is the best way to ensemble credal decision trees. It obtains excellent results compared with other ensemble procedures for this type of decision trees. Indeed, it should be noted that Boosting credal decision trees obtains a better Friedman rank score than a standard Random Forest method in the comparative study conducted among the seven methods presented in this paper. The results obtained with the set of tests carried out reinforce this statement.

The rest of the paper is organized as follows: in Section 2, we present previous knowledge on the method for building credal decision trees and the ensemble schemes utilized with credal decision trees in this paper. In Section 3, we check all the ensemble methods studied here, using two different approaches to deal with missing data and continuous variables, on a set of datasets widely used in classification. Section 4 is devoted to the conclusions.

2. Background

2.1. Method for building credal decision trees

A decision tree is a simple structure that can be used as a classifier. An important reference on its origin is the work Classification and Regression Trees by Breiman et al. [13]. Quinlan’s well-known ID3 algorithm for building decision trees, which was presented later, should also be mentioned.

In a decision tree, each internal (non-leaf) node represents an attribute variable (or predictive attribute or feature) and each branch represents one of the states of this variable. Each tree leaf specifies an expected value of the class variable (the variable to be predicted). Important aspects of the procedures for building decision trees are: the split criterion, i.e. the criterion used for branching; and the stop criterion, i.e. the criterion used to stop branching.

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

The split criterion employed to build credal decision trees [4] is based on the application of uncertainty measures on convex sets of probability distributions. Specifically, probability intervals are extracted from the dataset for each case of the class variable using Walley’s imprecise Dirichlet model (IDM) [14], which represents a specific kind of convex set of probability distributions (see [15]).

The IDM depends on a given hyperparameter \(s\) which does not depend on the sample space [14]. The IDM estimates that the probabilities for each value of the class variable \(C\) are within the interval:

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

with \(n_{c_j}\) as the frequency of the set of values \((C = c_j)\) in the dataset; and \(N\) the sample size. The value of parameter \(s\) determines the speed at which the upper and lower probability values converge when the sample size increases. Higher values of \(s\) give a more cautious inference. Walley [14] does not give a definitive recommendation for the value of this parameter but he suggests two candidates: \(s = 1\) or \(s = 2\).

For this type of probability intervals, the maximum entropy is estimated. This is a total uncertainty measure which is well known for this type of sets (see [16,17]).

The procedure for building credal trees is very close to the one used in Quinlan’s [10] well-known ID3 algorithm, replacing its Info-Gain split criterion with the Imprecise Info-Gain (IIG) split criterion. This criterion can be defined as follows: in a classification problem, let \(C\) be the class variable, \((Z_1, ..., Z_k)\) be the set of features, and \(Z\) be a feature; then

\[
IIG^C(Z, C) = S^C(K^C(C)) - \sum_i P(Z = z_i) S^C(K^C(C|Z = z_i)),
\]

where \(K^C(C)\) and \(K^C(C|Z = z_i)\) are the credal sets obtained using the IDM for the variables \(C\) and \((C|Z = z_i)\) respectively, for a partition \(D\) of the dataset (see [4]); and \(S^C\) is the maximum entropy function (see [15]).

The IIG criterion is different from the classical ones. It is based on the principle of maximum uncertainty (see [5]), broadly used in the classic information theory, in which it is known as the maximum entropy principle [18]. The use of the maximum entropy function in the procedure for building decision trees is justified in Abellán and Moral [6]. It is important to note that for a feature \(Z\) and a partition \(D\), \(IIG^C(Z, C)\) can be negative. This situation does not appear with classical split criteria such as the Info-Gain criterion used in ID3.\(^3\) This characteristic enables the IIG criterion to reveal features that worsen the information on the class variable.

Each node \(No\) in a decision tree produces a partition of the dataset (for the root node, \(D\) is considered to be the entire dataset). Furthermore, each node \(No\) has an associated list \(E\) of feature labels (that are not in the path from the root node to \(No\)). The procedure for building credal trees is explained in the algorithm in Fig. 1.

\(^3\) The Info-Gain criterion is actually a particular case of the IIG criterion using the parameter \(s = 0\).
Credal trees and the IIG criterion have been successfully used in other procedures and tools in data mining as a part of a procedure for selecting variables [20] and on datasets with classification noise [7]. An extended version of the IIG criterion has been used to define a semi-naïve Bayes classifier [21].

2.2. Ensemble methods using credal decision trees

The general idea normally used in the procedures for combining decision trees is based on the generation of a set of different decision trees combined with a majority vote criterion. When a new unclassified instance arises, each single decision tree makes a prediction and the instance is assigned to the class value with the highest number of votes. In an ensemble scheme, if all decision trees are quite similar, ensemble performance will not be much better than a single decision tree. However, if the ensemble comprises a broad set of different decision trees exhibiting good individual performance, the ensemble will become more robust, with a better prediction capacity. There are many different approaches to this problem, but Bagging [2], Random Forests [3], and AdaBoost [1] stand out as the best known and most competitive ones.

The schemes used here with credal trees can be briefly described as follows:

(i) The Bootstrap Aggregating ensemble method, known as Breiman’s Bagging [2] is an intuitive and simple method that performs very well. Diversity in Bagging is obtained by using bootstrap replicates of the original training dataset: different training datasets are randomly drawn with replacement. Subsequently, a single decision tree is built with each instance of the training dataset using the standard approach [13]. Thus, each tree can be defined by a different set of variables, nodes and leaves. Finally, their predictions are combined by a majority vote.

(ii) Boosting is an ensemble method based on the work of Kearns [22]. This procedure can be described in the following steps: (i) applying a weak classifier (such as a decision tree) to the learning data, where each observation is assigned an initially equal weight; (ii) applying weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification of the computed predicted classifications; (iii) going to the step (i) M times (M previously fixed); and (iv) combining predictions from individual models (weighted by accuracy of the models). In the Boosting procedure: for misclassified samples the weights are increased, while for correctly classified samples the weights are decreased. Thus, the principal idea of this method is that it uses a sequence of successive classifiers where each one depends upon its predecessors; and that each classifier considers the error of the previous classifier in order to decide what to focus on during the next iteration of the data. Boosting assumes training dataset to be perfect, so this procedure does not perform as well as other strategies with noisy training datasets. A special type of Boosting is the algorithm of Adaboost [1] which has demonstrated excellent experimental performance on benchmark datasets and real applications. AdaBoost is an adaptive algorithm in the sense that classifiers built subsequently are tweaked in favor of instances misclassified by previous classifiers. It is a feature selector with a principled strategy: minimization of upper bound on empirical error.

(iii) The IIG criterion is used in two important ways in the recent procedure IRN-Ensemble method of Abellán and Masegosa [8]. As a split criterion, a single credal tree needs to be built and used to obtain the set of features that will make up the root node of each tree used in the ensemble scheme. As stated in the preceding sections, this criterion can use certain features to provide negative values of information for the class variable. Therefore, all the features can be considered as potential root nodes or, to the contrary, none of them can be considered as root nodes. The method is not based upon a final voting procedure, and can be expressed as follows:

- Use IIG rank to obtain the set of features \( \{Z_1, \ldots, Z_m\} \), where \( IIG(C|Z_j) > 0 \), \( \forall j \in \{1, \ldots, m\} \).
- Build \( m \) credal trees \( \{T_1, \ldots, T_m\} \). Each \( T_j \) is built using \( Z_j \) as the root node and the rest of nodes, for this tree, are selected following the procedure described in the previous section, using the IIG criterion.
- In order to classify a new observation, apply the new case to each of the \( m \) credal trees and consider the frequency (number of cases) of each state of the class variable in each of the leaf nodes. For a tree \( T_j \), a new observation is associated with a leaf node, and this also has an associated partition of the dataset. Then \( n_{c_j} \) is the frequency (number of cases) with which \( c_j \) appears in the partition generated by the leaf node of \( T_j \). Calculate the following relative frequencies (probabilities):

\[
q_j = \frac{\sum_{i \in \text{partition}} n_{c_j}}{\sum_{j \in \{1, \ldots, m\}} n_{c_j}}.
\]

for each \( c_j \).

- For a new observation, the value obtained by the IRN-Ensemble classification method is \( c_n \), where: \( c_n = \arg(\max(q_j)) \).

As can be inferred from the above explanation, the IRN method has an important characteristic with respect to the rest: it does not fix the number of trees (or iterations) to be used. It depends on the number of informative features with respect to the class variable (see [8]). This implies that it is possible to use a number of trees equal to the number of features in the dataset, or even zero. As a very low number of trees can be used, the authors of the IRN method do not consider a final voting procedure appropriate to classify a new observation.

3. Experimentation

Our aim is to study the performance of the methods: the original IRN method (noted as IRN); the IRN method replacing its final

\[\text{Fig. 1. Procedure to build a credal decision tree.}\]
procedure with a voting procedure (noted as IRNV); Bagging CTs (noted as BAGCT); Bagging decision trees built with Quinlan's classic Info-Gain split criterion [10] (noted as BAGIG); Boosting CTs (noted as BOOCT); Boosting decision trees built with the Info-Gain split criterion (noted as BOOIG); and the Random Forest method (noted as RF).

In order to check the above procedures, we used a broad and diverse set of 25 known datasets, obtained from the UCI repository of machine learning datasets which can be downloaded directly from http://archive.ics.uci.edu/ml/. We took that datasets that are different with respect to the number of cases of the variable to be classified, dataset size, feature types (discrete or continuous) and number of cases of the features. A brief description of the datasets can be found in Table 1.

For our experimentation, we used Weka software [23] on Java 1.5, and we added the methods required to build decision trees using the IIG split criterion. The ensembles of credal trees used here: IRN-Ensembles (IRN and IRNV), Bagging credal trees (BAGCT) and AdaBoost credal trees (BOOCT) were implemented using data structures of Weka. The parameter of the IDM was set to s = 1, i.e. the value used in the original method by Abellán and Moral [4], because the procedure to obtain the maximum entropy value (see [15]) reaches its lowest computational cost for this value. Bagging and Boosting ensembles were built with 100 decision trees. Although the number of trees can strongly affect ensemble performance, this is a reasonable number of trees for the low-medium size of the datasets used in this study, and moreover it was used in related research, such as Freund and Schapire [1]. We used the Random Forest method (RF) that can be found in Weka software as a benchmark to compare the results. We used it with its default configuration, except that the number of trees used for this method was 100, the same as for the above-mentioned ensemble methods.

We used the following approaches to deal with missing data and continuous variables in different implementations of the methods:

(A) For each feature, missing values within datasets were replaced with mean (for continuous features) and mode (for discrete features) values, with Weka’s own filters. In the same way, continuous features were discretized using the known Fayyad and Irani’s discretization method [24]. The approach was applied by means of the training dataset and it was then transferring to the test dataset, for each pair of training/test datasets.

(B) Assume that missing values are randomly distributed. In order to compute the scores, the instances are split into units. The initial weight of an instance is equal to the unit, but when it drops down a branch, it receives a weight equal to the proportion of instances belonging to this branch (weights sum to 1). For continuous variables, only binary split attributes are considered and each possible split point is evaluated and finally selected. The one which induced a partition of the samples with the highest split score, using the IIG criterion, is ultimately selected.

It should be noted that both approaches were applied by using the training dataset and then translated the result to the test dataset. For each dataset, we repeated a 10-fold cross validation procedure 10 times. Alternative methods of processing missing data have been presented in Pelckmans et al. [25].

Tables 2 and 3 show the accuracy of the methods with approaches (A) and (B), respectively.

Following the recommendations of Demsar [26], we used a series of tests to compare the methods. Also, we have considered it would be equally interesting to take into account the work by García and Herrera [27], where the work of Demsar [26] is completed. We used the following tests:

- To compare two classifiers on multiple datasets:
  **Wilcoxon’s test** (Wilcoxon [28]): a non-parametric test which ranks the differences in performance of two classifiers for each dataset, ignoring the signs, and compares the ranks for the positive and the negative differences.

- To compare multiple classifiers on multiple datasets:
  **Friedman’s test** (Friedman [11,12]): a non-parametric test that ranks the algorithms separately for each dataset, 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 **Holm’s test** (Holm [29]).

We have considered the above set of tests following a trade-off on the recommendations by Demsar [26] and by García and Herrera [27]. As suggested in those works, the Friedman test may report a significant difference but the post hoc test fails to detect it. This is due to the lower power of the post hoc test conducted. Nemenyi's test [30] is recommended by Demsar, but in some situations can be

---

5 In the event of ties, the ensemble procedures implemented in Weka (such as the ones used in this paper) select the first case of the class variable they find, considering their order in the dataset file.

6 The IRN-Ensemble method does not have a fixed number of trees to be used. It uses a number of trees less or equal than the number of features of each dataset. Always it is less than 70 decision trees in this experimental study.

7 This approach is the one presented for the C4.5 method in Weka, adapted for credal trees using the IIG criterion.

8 The dataset is separated into 10 subsets, and each one is used as a test dataset. The set obtained by joining the other 9 subsets is used as the training dataset. Thus, we have 10 training datasets and 10 test datasets. This procedure is repeated 10 times with prior random reordering. Finally, it produces 100 training datasets and 100 test datasets. The percentage of correct classifications for each dataset, given in tables, is the average of these 100 trials.

9 All the tests were carried out using Weel software, available at http://www.kel.ee.
a less sensitive test than others, as it is described by García and Herrera. With Nemenyi’s test we can encounter situations where the differences expressed by the Friedman test were not detected. Hence, as the latter authors recommended, we considered conducting a post hoc Holm’s test.

The study of the p-values is very interesting because, citing a paragraph by García and Herrera [27]: “A p-value provides information about whether a statistical hypothesis test is significant or not, and it also indicates something about how significant the result is: The smaller p-value, the stronger the evidence against the null hypothesis. Most important, it does this without committing to a particular level of significance”. When a p-value comes from a multiple comparison, it does not take into account the other comparisons in the multiple study. To solve this problem, we can use a study of the Adjusted P-values (APVs) [31]. Using the APVs in a statistical analysis gives us more information because it is not focused on a fixed level of significance. The work by García and Herrera [27] provides an explanation of how to obtain the values. In some situations, such as this one, certain methods are more similar in performance. Therefore, to make a clear distinction between them, it is very interesting the study of the APVs of the tests conducted in the experiments.

Table 4 gives the results of the Wilcoxon’s test carried out in each method separately in order to compare their performance by using approaches (A) and (B), respectively. The table shows the approach in which the method is better using Wilcoxon’s test.

Tables 5 and 6 show Friedman’s ranks of the methods with approaches (A) and (B), respectively (the null hypothesis is rejected in both cases).

Table 7 shows the p-values of the tests carried out on the methods when the approach (A) is used. The column p-valuec corresponds to the p-values of the Holm’s tests on the methods in each row. The column APVc corresponds to the adjusted p-values of each comparison obtained from the Holm’s tests. Table 8 shows the same results of the tests on the methods when the approach (B) is used.11

### 3.1. Comments on the results

From Tables 2 and 3 we can obtain the differences between the results of the methods using approaches (A) and (B). Some differences can be seen in favor of the results obtained from approach (B) compared to the results from approach (A). This is particularly evident for some small datasets such as glass2, hepatitis and sonar, and for medium ones such as vowel. These differences are more noticeable in methods BOOc and RF.

We conducted Wilcoxon’s tests on each separate model with each approach, i.e. each method using approach (A) against the same method using approach (B). The results are given in Table 4. In these tests, we obtained that the IRN method with the (A) approach exhibits no significant differences from the same method using approach (B). The same situation also arises for the IRN, BAGc, BAGic and BOOc methods. However, it is important to remark that the tests performed for the BOOc and RF methods express that there are significant positive differences in accuracy when the (B) approach is used.

The Friedman test carried out for each set of methods with each approach rejects the null hypothesis. Therefore, Holm’s tests were conducted. The Friedman’s ranks of the methods using approach (A), expressed in Table 5, show that the best methods are IRN and RF in this order; whereas the worst methods are those in which the Info-Gain criterion was used for the decision trees: BOOc and BAGic. The situation is different for the best methods when approach (B) is used (Table 6): the best methods here are BOOc and RF in this order, but the situation is similar for the worst methods, again the BOOic and BAGic methods have the worse results.

Focusing on the results of the Holm’s tests, Tables 7 and 8 express the p-values of each comparison between two methods with

---

11 In both tables, for a weaker level of significance \( \alpha = 0.1 \), Holm’s procedure rejects those hypotheses that have a p-value \( \leq 0.003846 \). With the Nemenyi’s test the results are very similar to the ones obtained with the Holm’s test.
Table 3
Percentage of correct classification of the methods with the (B) approach.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IRN</th>
<th>IRNV</th>
<th>BAGCT</th>
<th>BAGIG</th>
<th>BOOCT</th>
<th>BOOIG</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anneal</td>
<td>99.10</td>
<td>99.09</td>
<td>98.89</td>
<td>98.86</td>
<td>99.78</td>
<td>99.31</td>
<td>99.68</td>
</tr>
<tr>
<td>Audiology</td>
<td>78.41</td>
<td>75.49</td>
<td>80.41</td>
<td>77.67</td>
<td>81.91</td>
<td>72.15</td>
<td>80.32</td>
</tr>
<tr>
<td>Autos</td>
<td>82.37</td>
<td>79.98</td>
<td>80.32</td>
<td>67.77</td>
<td>82.99</td>
<td>61.04</td>
<td>84.19</td>
</tr>
<tr>
<td>Breast-cancer</td>
<td>71.31</td>
<td>70.57</td>
<td>70.35</td>
<td>69.84</td>
<td>68.58</td>
<td>67.00</td>
<td>70.12</td>
</tr>
<tr>
<td>cnc</td>
<td>53.03</td>
<td>52.69</td>
<td>53.23</td>
<td>54.64</td>
<td>50.46</td>
<td>52.46</td>
<td>50.50</td>
</tr>
<tr>
<td>Horse-colic</td>
<td>85.83</td>
<td>85.48</td>
<td>84.91</td>
<td>84.75</td>
<td>81.27</td>
<td>81.92</td>
<td>85.02</td>
</tr>
<tr>
<td>German-credit</td>
<td>74.27</td>
<td>74.81</td>
<td>74.65</td>
<td>74.46</td>
<td>74.95</td>
<td>72.06</td>
<td>76.10</td>
</tr>
<tr>
<td>Pima-diabetes</td>
<td>74.45</td>
<td>74.48</td>
<td>75.80</td>
<td>76.05</td>
<td>73.97</td>
<td>74.43</td>
<td>75.97</td>
</tr>
<tr>
<td>Glass2</td>
<td>81.75</td>
<td>80.21</td>
<td>83.00</td>
<td>81.05</td>
<td>88.05</td>
<td>82.41</td>
<td>87.90</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>81.66</td>
<td>81.76</td>
<td>80.99</td>
<td>81.43</td>
<td>85.20</td>
<td>79.97</td>
<td>83.58</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>91.91</td>
<td>91.77</td>
<td>91.23</td>
<td>91.37</td>
<td>93.79</td>
<td>90.89</td>
<td>93.45</td>
</tr>
<tr>
<td>Labor</td>
<td>87.40</td>
<td>88.23</td>
<td>86.53</td>
<td>86.33</td>
<td>88.73</td>
<td>82.60</td>
<td>87.27</td>
</tr>
<tr>
<td>Lymphography</td>
<td>77.55</td>
<td>77.48</td>
<td>76.24</td>
<td>77.44</td>
<td>83.57</td>
<td>77.53</td>
<td>83.15</td>
</tr>
<tr>
<td>Mushroom</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.98</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Segment</td>
<td>97.56</td>
<td>97.55</td>
<td>97.45</td>
<td>96.83</td>
<td>98.67</td>
<td>96.52</td>
<td>98.16</td>
</tr>
<tr>
<td>Sick</td>
<td>98.98</td>
<td>98.98</td>
<td>98.97</td>
<td>98.76</td>
<td>99.00</td>
<td>98.69</td>
<td>98.42</td>
</tr>
<tr>
<td>Solar-flare</td>
<td>97.10</td>
<td>97.10</td>
<td>97.31</td>
<td>97.84</td>
<td>95.86</td>
<td>97.44</td>
<td>96.91</td>
</tr>
<tr>
<td>Sonar</td>
<td>82.99</td>
<td>83.14</td>
<td>80.78</td>
<td>78.00</td>
<td>86.81</td>
<td>79.86</td>
<td>84.72</td>
</tr>
<tr>
<td>Soybean</td>
<td>90.91</td>
<td>89.82</td>
<td>90.47</td>
<td>87.60</td>
<td>91.86</td>
<td>89.12</td>
<td>93.37</td>
</tr>
<tr>
<td>Sponge</td>
<td>92.50</td>
<td>92.50</td>
<td>92.63</td>
<td>92.50</td>
<td>93.71</td>
<td>91.46</td>
<td>95.00</td>
</tr>
<tr>
<td>Vote</td>
<td>96.00</td>
<td>95.77</td>
<td>96.34</td>
<td>95.54</td>
<td>95.47</td>
<td>94.94</td>
<td>96.43</td>
</tr>
<tr>
<td>Vowel</td>
<td>91.38</td>
<td>91.26</td>
<td>91.13</td>
<td>88.08</td>
<td>95.80</td>
<td>81.70</td>
<td>96.68</td>
</tr>
<tr>
<td>Zoo</td>
<td>92.01</td>
<td>92.01</td>
<td>92.40</td>
<td>92.51</td>
<td>95.65</td>
<td>92.10</td>
<td>96.33</td>
</tr>
<tr>
<td>Average</td>
<td>87.10</td>
<td>86.76</td>
<td>86.92</td>
<td>85.92</td>
<td>88.21</td>
<td>84.54</td>
<td>88.51</td>
</tr>
</tbody>
</table>

Table 4
Results of the Wilcoxon’s test conducted with a level of significance of 0.05, performed in each method separately using approaches (A) and (B). The table shows the approach in which the method is better using the results of the Wilcoxon’s test. *-* indicates non significant statistical differences.

<table>
<thead>
<tr>
<th>Wilcoxon Test</th>
<th>IRN</th>
<th>IRNV</th>
<th>BAGCT</th>
<th>BAGIG</th>
<th>BOOCT</th>
<th>BOOIG</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)/(B)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5
Friedman’s ranks of the algorithms with the approach (A).

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRN</td>
<td>3.02</td>
</tr>
<tr>
<td>RF</td>
<td>3.04</td>
</tr>
<tr>
<td>BAGCT</td>
<td>3.42</td>
</tr>
<tr>
<td>IRNV</td>
<td>3.5</td>
</tr>
<tr>
<td>BOOCT</td>
<td>4.42</td>
</tr>
<tr>
<td>BAGIG</td>
<td>5.16</td>
</tr>
<tr>
<td>BOOIG</td>
<td>5.44</td>
</tr>
</tbody>
</table>

each approach. Table 7 shows that with approach (A) the best methods are IRN and the RF. They obtain very similar results, i.e. similar number of wins in the tests carried out. The IRNV and BAGCT methods are also winner ones when they are compared with the worst method: BOOIG. Table 8 shows that with the approach (B) the best methods are BOOCT and the RF methods. The IRN and IRNV methods are also winners when they are compared with the worst method: BOOIG.

It is not easy to compare the best methods for each approach with the results obtained in this work, but as García and Herrera

12 IRN has one more win in the Holm’s test than the RF if a weaker level of significance of 0.1 is used.
13 We must remark that BOOCT has one more win in the Holm’s test than the RF if a weaker level of significance of 0.1 is used.

Please cite this article in press as: J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Informat. Fusion (2012), http://dx.doi.org/10.1016/j.inffus.2012.03.003
recommend for this type of situations, we can use the study of the APVs to analyse the differences between the best methods in this study in greater depth. Table 7 shows that for approach (A) there exist very similar results in the strength of the rejection of the null hypothesis (values of the APVs) of the comparisons "IRN vs. M" and "RF vs. M", with M being any other method. However, in Table 8, about the tests on the methods with approach (B), we can observe that this situation is more clearly in favor of the BOOCT method when it is compared with the RF method. For example, the APV of the comparison "RF vs. BOOGC" is 2.5 times greater than the one of the comparison "BOOCT vs. BOOGC"; and the APV of the comparison "RF vs. BAGGc" is 1.75 times the one of the comparison "BOOCT vs. BAGGc".

We can use similar analysis to check the differences between the worst methods for each approach. We have the same situation with both approaches: comparisons "BOOGC vs. M" always give an APV that is notably lower than the one for "BAGGc vs. M", with M being any other method. Thus, we can say that clearly the worst method of our study with the two approaches is BOOGC.

Using the analysis reported in this paper, we can summarize the following: the methods using approach (A) are worse than or equal to the equivalent using approach (B); the RF method obtains the best average value with both approaches but it cannot be considered the best one for any approach; with approach (B), the RF and BOOCT methods can be considered the best methods, with BOOCT being better than the RF in the comparisons with the other methods.

4. Conclusions

In this paper, we analyzed different strategies for combining credal decision trees, i.e., decision trees built using imprecise probabilities (using the IDM); and uncertainty measures (using the maximum entropy method). We used previously created Bagging and Boosting schemes and a new one specifically developed for credal decision trees. We experimentally showed that different approaches of dealing with missing data and continuous variables can significantly vary the results of some methods for creating ensemble credal decision trees. We can conclude that the Boosting scheme with the appropriate approach is the best way to combine credal decision trees. Using the best approach for each method (B approach), we used a set of tests carried out on the result obtained from the seven methods analyzed here to show that Boosting credal trees and the Random Forest methods are the best ones, with Boosting credal trees being better than the Random Forest method in comparison with any other method.

Acknowledgments

This work has been supported by the Spanish “Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía” under Project TIC-06016.

I’m very grateful to the anonymous reviewers of this paper for their valuable suggestions and comments to improve its quality.

References


Please cite this article in press as: J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Informat. Fusion (2012), http://dx.doi.org/10.1016/j.inffus.2012.03.003

