ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Credal decision trees in noisy domains

Carlos J. Mantas and Joaquín Abellán

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

Abstract. Credal Decision Trees (CDTs) are algorithms to design classifiers based on imprecise probabilities and uncertainty measures. In this paper, the C4.5 and CDT procedures are combined in a new one. This depends on a parameter s. Several experiments are carried out with different values for s. The new procedure obtains better performance than C4.5 on data sets with different noise levels.

1 Introduction

By using the theory of imprecise probabilities presented in Walley [10], known as the Imprecise Dirichlet Model (IDM), Abellán and Moral [1] have developed an algorithm for designing decision trees, called *credal decision trees* (CDTs). The variable selection process for this algorithm is based on imprecise probabilities and uncertainty measures on credal sets, i.e. closed and convex sets of probability distributions. In this manner, this algorithm considers that the training set is not reliable when the variable selection process is carried out. This method obtains good experimental results, especially when noisy data are classified [3, 6].

The theory of credal decision trees and the C4.5 are connected in this paper. So, Credal-C4.5 is presented. The performance of this algorithm depends on a parameter s. Credal-C4.5 has a low computational cost with $s \leq 1$. A trial-and-error process has been carried out in order to find the best value for s. We have compared Credal-C4.5 and classic C4.5 when they classify data sets with or without noise. Different optimal values for s are found in terms of the noise level of the data sets. These facts are analyzed in this work.

2 Credal Decision Trees

The split criterion employed to build Credal Decision Trees (CDTs) (Abellán and Moral [1]) is based on imprecise probabilities and the application of uncertainty measures on credal sets. The mathematical basis of this procedure can described as follows: Let Z be a variable with values in $\{z_1, \dots, z_k\}$. Let us suppose a probability distribution $p(z_j), j = 1, .., k$ defined for each value z_j from a data set.

Walley's Imprecise Dirichlet Model (IDM) [10] is used to estimate probability intervals from the data set for each value of the variable Z, in the following way

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

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.

This representation gives rise to a specific kind of credal set on the variable Z, $K(Z)$ (see Abellán [2]), defined as

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

On this type of sets (credal sets), uncertainty measures can be applied. The procedure to build CDTs uses the maximum of entropy function on the above defined credal set (see Klir [5]). This function, denoted as H^*, is defined as $H^*(K(Z)) = max\,\{H(p) \,|\, p \in K(Z)\}$, where the function H is the Shannon's entropy function [9]. H^* is a total uncertainty measure which is well known for this type of set [5]. The procedure for H^* in the IDM reaches its lowest cost with $s \leq 1$ and it is simple (see [2]). For this reason, we will use values $s \leq 1$ in the experimentation section.

3 Credal-C4.5

The method for building Credal-C4.5 trees is similar to the Quinlan‘s C4.5 algorithm [8]. The main difference is that Credal-C4.5 estimates the values of the features and class variable by using imprecise probabilities and uncertainty measures on credal sets. Credal-C4.5 considers that the training set is not very reliable because it can be affected by class or attribute noise. 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 $IIGR^{\mathcal{D}}(C, X) = \frac{IIG^{\mathcal{D}}(C,X)}{H(X)}$, where *Imprecise Info-Gain* (IIG) is equal to:

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

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 and Moral [1]); $P^{\mathcal{D}}(X = x_i)$ $(i = 1, ..., n)$ 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. From the set

$$B = \{x_j \in X \mid H(C|X = x_j) = \max_i\{H(C|X = x_i)\}\},$$

the probability distribution $P^{\mathcal{D}}$ will be equal to

$$P^{\mathcal{D}}(x_i) = \begin{cases} \frac{n_{x_i}}{N+s} & \text{if } x_i \notin B \\ \frac{n_{x_i}+s/m}{N+s} & \text{if } x_i \in B \end{cases}$$

where m is the number of elements of B. This expression shares out s among the values x_i with $H(C|X = x_i)$ maximum.

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 Figure 1 and its characteristics below:

```
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
        $\alpha = \max_{X_j \in \mathcal{M}} \left\{ IIGR^{\mathcal{D}}(C, X_j) \right\}$
   with $\mathcal{M} = \left\{ X_j \in \mathcal{L} \mid IIG^{\mathcal{D}}(C, X_j) > avg_{X_j \in \mathcal{L}} \left\{ IIG^{\mathcal{D}}(C, X_j) \right\} \right\}$
6. If $\alpha \leq 0$ then Exit
7. Else
   8. Let $X_l$ be the variable for which the maximum $\alpha$ is attained
9. Remove $X_l$ from $\mathcal{L}$
10. Assign $X_l$ to node $No$
11. For each possible value $x_l$ of $X_l$
   12. Add a node $No_l$
   13. Make $No_l$ a child of $No$
   14. Call BuilCredalC4.5Tree($No_l$,$\mathcal{L}$)
```

Fig. 1: Procedure to build a Credal-C4.5 decision tree.

Split Criteria: *Imprecise Info-Gain Ratio* is employed for branching. As in the 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 split attributes. **Labeling leaf node**: The most probable value of the class variable in the partition associated with a leaf node is inserted as label. **Stopping Criteria**: The branching is stopped when the uncertainty measure is not reduced ($\alpha \leq 0$, 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). **Handling Numeric Attributes and Missing Values**: Both are handled in the same way that classic C4.5 algorithm (using here the IIG criterion). **Post-Pruning Process**: Like C4.5, *Pessimistic Error Pruning* is employed in order to prune a Credal-C4.5.

3.1 The parameter s and its relation with the noise

The IDM considers that the data distributions are not precise. According with the model, there is a number of data s that is not present for a variable Z and the values for these data are not known. Hence, it is considered that the information s can take any value. In this way, we employ a credal set instead of only one probability distribution to estimate the values of a variable Z.

If our data set is noisy, we have two probability distributions for each variable: the one provided by the noisy data, called PD_{noise}, and the unknown distribution without noise, called PD_{clean}. If we design a classifier by using the IDM, we work with a credal set that has several probability distributions around PD_{noise}. Finally, we use the probability distribution selected by the maximum of the entropy function H^* according the principle of maximum uncertainty, called PD_{IDM}.

The distance between PD_{noise} and PD_{clean} depends on the level of noise. On the other hand, the size of the credal set depends on the value of the parameter s. According to these facts, the following situations can happen: (i) If the level of noise is low, PD_{noise} and PD_{clean} are close. In this case, it is possible that the credal set contains to PD_{clean} even with a low value for s. Besides, it is also feasible that PD_{IDM} is close to PD_{clean} when s is low; (ii) If the level of noise is high, PD_{noise} and PD_{clean} are far. In this way, PD_{IDM} is close to PD_{clean} when the credal set is high. This is achieved by using values of s closer to 1.

4 Experimental analysis

We used a broad and diverse set of 50 known data sets, obtained from the *UCI repository of machine learning data sets*. They are anneal, arrhythmia, audiology, autos, balance-scale, breast-cancer, wisconsin-breast-cancer, car, cmc, horse-colic, credit-rating, german-credit, dermatology, pima-diabetes, ecoli, Glass, haberman, cleveland-14-heart-disease, hungarian-14-heart-disease, heart-statlog, hepatitis, hypothyroid, ionosphere, iris, kr-vs-kp, letter, liver-disorders, lymphography, mfeat-pixel, nursery, optdigits, page-blocks, pendigits, primary-tumor, segment sick, solar-flare2, sonar, soybean, spambase, spectrometer, splice, Sponge, tae, vehicle, vote, vowel, waveform, wine and zoo.

We used *Weka* software [11] on Java 1.5 for our experimentation. We use C4.5 algorithm provided by *Weka* software, called *J48*, and added the methods to build Credal-C4.5 trees with the same experimental conditions. The parameter of the IDM for the Credal-C4.5 algorithm was set to the values $s = 0.25$, $s = 0.5$, $s = 0.75$ and $s = 1.0$ (for $s = 0.0$ Credal-C4.5 and C4.5 are equivalent). Using *Weka's* filters, we added the following percentages of random noise to the class variable: $0\%, 5\%, 10\%$ and 30%, only in the training data set. Finally, we repeated 10 times a 10-fold cross validation procedure for each data set.

Following the recommendation of Demsar [4], we used a series of tests to compare the methods. We used, for a level of significance of $\alpha = 0.05$: a **Friedman test** to check if all the procedures are equivalents and a pos-hoc **Nemenyi test** to compare all the algorithms to each other (see [4] for more references about the tests).

4.1 Results and comments

Table 1 presents the average result of accuracy and standard deviations for each method and each level of noise. More details are not shown by limitations of space. Table 2 shows Friedman's ranks. We remark that the null hypothesis is rejected in the cases with noise 10% and 30%. Tables 3, 4 show the p-values of the Nemenyi test for the methods C4.5 and Credal-C4.5 in the experimentation.

In all the cases, Nemenyi procedure rejects the hypotheses that have a p-value$\leq$ 0.005.

Tree	noise 0%	noise 5%	noise 10%	noise 30%
C4.5	82.62 (14.16)	81.77 (14.50)	80.77 (14.97)	74.14 (16.24)
Credal-C4.5$_{s=0.25}$	82.54 (14.23)	81.92 (14.44)	81.06 (14.77)	75.27 (15.66)
Credal-C4.5$_{s=0.5}$	82.50 (14.24)	81.92 (14.48)	81.16 (14.82)	75.85 (15.63)
Credal-C4.5$_{s=0.75}$	82.49 (14.30)	81.98 (14.46)	81.30 (14.72)	76.22 (15.67)
Credal-C4.5$_{s=1.0}$	82.35 (14.29)	81.90 (14.51)	81.27 (14.76)	76.65 (15.59)

Table 1: Average result of accuracy and standard deviations for C4.5 and Credal-C4.5 (varying s) on each level of noise

Tree	noise 0%	noise 5%	noise 10%	noise 30%
C4.5	2.95	3.17	3.79	4.04
Credal-C4.5$_{s=0.25}$	2.83	2.78	3.4	3.47
Credal-C4.5$_{s=0.5}$	3.08	3.0	2.94	2.97
Credal-C4.5$_{s=0.75}$	2.89	3.06	2.44	2.57
Credal-C4.5$_{s=1.0}$	3.25	2.99	2.43	1.95

Table 2: Friedman's ranks of C4.5 and Credal-C4.5 (varying s) on each level of noise

The results shown are analyzed as follows:

Average accuracy: According to this factor, C4.5 obtains the best result for data without noise. If the level of noise is increased, then the best results are achieved by Credal-C4.5 with values for s closer to 1. **Friedman's ranking**: According this ranking, Credal-C4.5 with the value $s = 0.25$ obtains the best results when data sets without noise are classified. This fact indicates that the original data sets are not fully clean, they are noisy with a very low level. For data sets with noise, classic C4.5 is the worst model. In particular, for noise 5% Credal-C4.5 with $s = 0.25$ achieves the best results. That is, credal sets with a low size are enough to estimate the values of variables with a small level of noise. For noise 10% and 30%, the value $s = 1.0$ builds the credal trees with the best results. **Nemenyi test**: According to this test, the differences between methods are statistically significant for noise 10% and 30%. In these cases, Credal-C4.5 with values of s closer to 1 (0.75 and 1.0) are better than classic C4.5 and Credal-C4.5 with low value for s (0.25).

5 Conclusion

Credal-C4.5 model has been presented. The performance of this model depends on the parameter s. The relation between s and noise has been exposed. We have concluded that Credal-C4.5 with a value of s close to 1 is the best model when data sets with a big noise level are classified.

Acknowledgments

This work has been supported by the Spanish "Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía" under Project TIC-6016, TIC-04813 and Spanish MEC project TIN2012-38969.

i	algorithms	$p - values$
10	C4.5 vs. Credal-C4.5$_{s=1.0}$	0.000017
9	C4.5 vs. Credal-C4.5$_{s=0.75}$	0.000020
8	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=1.0}$	0.002159
7	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=0.75}$	0.002399
6	C4.5 vs. Credal-C4.5$_{s=0.5}$	0.007190
5	Credal-C4.5$_{s=0.5}$ vs. Credal-C4.5$_{s=1.0}$	0.106796
4	Credal-C4.5$_{s=0.5}$ vs. Credal-C4.5$_{s=0.75}$	0.113846
3	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=0.5}$	0.145767
2	C4.5 vs. Credal-C4.5$_{s=0.25}$	0.217468
1	Credal-C4.5$_{s=0.75}$ vs. Credal-C4.5$_{s=1.0}$	0.974773

Table 3: p-values of the Nemenyi test with $\alpha = 0.05$ for the methods C4.5 and Credal-C4.5 (varying s) and 10% of noise

i	algorithms	$p - values$
10	C4.5 vs. Credal-C4.5$_{s=1.0}$	0
9	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=1.0}$	0.000002
8	C4.5 vs. Credal-C4.5$_{s=0.75}$	0.000003
7	C4.5 vs. Credal-C4.5$_{s=0.5}$	0.000715
6	Credal-C4.5$_{s=0.5}$ vs. Credal-C4.5$_{s=1.0}$	0.001257
5	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=0.75}$	0.004427
4	Credal-C4.5$_{s=0.75}$ vs. Credal-C4.5$_{s=1.0}$	0.049924
3	C4.5 vs. Credal-C4.5$_{s=0.25}$	0.071467
2	Credal-C4.5$_{s=0.25}$ vs. Credal-C4.5$_{s=0.5}$	0.113846
1	Credal-C4.5$_{s=0.5}$ vs. Credal-C4.5$_{s=0.75}$	0.205903

Table 4: p-values of the Nemenyi test with $\alpha = 0.05$ for the methods C4.5 and Credal-C4.5 (varying s) and 30% of noise

References

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

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

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

[4] J. Demsar, Statistical Comparison of Classifiers over Multiple Data Sets, *Journal of Machine Learning Research*, 7, 1–30, 2006.

[5] G.J. Klir, Uncertainty and Information, Foundations of Generalized Information Theory. John Wiley, Hoboken, NJ, 2006.

[6] C.J. Mantas and J. Abellán, Analysis and extension of decision trees based on imprecise probabilities: application on noisy data, *Expert Systems with Applications*, 41, 2514–2525, 2014.

[7] J.R. Quinlan, Induction of decision trees, *Machine Learning*, 1, 81–106, 1986.

[8] J.R. Quinlan, Programs for Machine Learning. Morgan Kaufmann series in Machine Learning, 1993.

[9] C.E. Shannon, A mathematical theory of communication, *The Bell System Technical Journal*, 27, 379–423, 623–656, 1948.

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

[11] I.H. Witten and E. Frank, Data Mining, Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.