Application of E^2M decision trees to rubber quality prediction.

Nicolas Sutton-Charani[1,2], Sébastien Destercke[1], and Thierry Denœux[1]

[1] Université Technologie de Compiègne, UMR 7253 Heudiasyc
60203 COMPIEGNE Cedex FRANCE
name.surname@hds.utc.fr
http://www.hds.utc.fr/
[2] CIRAD IATE, 2, place Pierre Viala F-34060 Montpellier Cedex 02 FRANCE
F-34392 Montpellier, France

Abstract. In many applications, data are often imperfect, incomplete or more generally uncertain. This imperfection has to be integrated into the learning process as an information in itself. The *E^2M decision trees* is a methodology that provides predictions from uncertain data modelled by belief functions. In this paper, the problem of rubber quality prediction is presented with a belief function modelling of some data uncertainties. Some resulting *E^2M decision trees* are presented in order to improve the interpretation of the tree compared to standard decision trees.

Keywords: classification; decision trees; rubber quality; hevea; belief functions; algorithm EM.

1 Introduction

Learning a classifier from uncertain data necessitates an adequate modelling of this uncertainty, however learning with uncertain data is rarely straightforward. As data uncertainty is of epistemic nature, the standard probabilistic framework is not necessarily the best framework to deal with it. More general frameworks have therefore been proposed [1–3] that provide more adequate model for this type of uncertainty. Different classifier learning techniques [4–6] using these models have then been developed.

In this paper, our goal is to learn a model from agronomic data. More precisely, we want to predict natural rubber quality from data concerning latex culture and natural rubber maturation. Generally speaking, uncertain measurements and expert assessments are common in agronomy and life science, mainly due to field and economy constraints. They are therefore domains where data uncertainty happens a lot. We retain the belief-function theory [2, 7], as it is flexible enough to model a large variety of data uncertainties. The chosen classifier is the E^2M decision tree [8], for it is usually efficient and interpretable (an essential feature for agronomic experts).

After a short overview on the necessary background in Section 2, we detail in Section 3 the application context as well as the uncertainty models we used. We conclude Section 3 by comparing the results of the obtained E^2M decision trees with classical ones.

2 Background

We briefly recall the elements needed in the application.

2.1 Formalism

As in any classification problem, the aim is to predict a class label Y from a set of attributes (or features) X. The classifier is learnt on a learning dataset LD containing samples of (X,Y). The classifier is then evaluated by measuring its accuracy on a test dataset TD, comparing the predicted class labels with the real ones.

The attributes $X = (X^1,\dots,X^J)$ take their values on $\Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^J}$, the class Y on $\Omega_Y = \{\omega_1,\dots,\omega_K\}$. That is, K different classes are predicted using J different attributes (either categorical or real-valued).

A *precise* dataset containing N samples is a set of observations of (X,Y) and is denoted by

$$D = \begin{pmatrix} x_1, y_1 \\ \vdots \\ x_N, y_N \end{pmatrix} = \begin{pmatrix} x_1^1,\dots,x_1^J, y_1 \\ \vdots \\ x_N^1,\dots,x_N^J, y_N \end{pmatrix}.$$

Samples are here assumed to be i.i.d (independant and identically distributed).

2.2 Belief-function theory

The theory of belief functions (TBF), also called evidence theory or Dempster-Shafer theory was first presented by Dempster [2] in a statistical approach. The very basis of the TBF is here presented, with a special focus on the evidential likelihood proposed by Denoeux [9].

Generalities Assume we have an uncertain observation of a variable W defined on a finite space Ω_W. We model this observation by a *belief mass* $m^W : 2^{\Omega_W} \to [0,1]$ verifying $\sum_{B \in 2^{\Omega_W}} m^W(B) = 1$. We assume here that $m^W(\emptyset) = 0$. A *focal element* $A \in 2^{\Omega_W}$ is a set such that $m^W(A) > 0$. From this mass, the belief and plausibility functions are defined by:

$$Bel^W(A) = \sum_{B \subseteq A} m^W(B), \qquad Pl^W(A) = \sum_{B \cap A \neq \emptyset} m^W(B)$$

$Bel^W(A)$ measures the amount of information that implies $W \in A$, and is a measure of certainty, while $Pl^W(A)$ measures the amount of information that does not conflict with $W \in A$, and is a measure of plausibility. We naturally have $Bel^W(A) \leq Pl^W(A)$ with the two being equal in the specific case of probabilities.

The particular cases of precise data, imprecise data, missing data, probabilities and possibilities can all be modelled by belief functions:

$$\begin{aligned}
\text{precise data} &: \quad m^W(\{w\}) = 1 \\
\text{imprecise data} &: \quad m^W(A) = 1 \\
\text{missing data} &: \quad m^W(\Omega_W) = 1 \quad \textit{(complete ignorance)} \\
\text{probabilities} &: \quad m^W(A) > 0 \quad \textit{if} \quad |A| = 1 \\
\text{consonnant mass functions} &: \quad m^W(A) > 0 \text{ and } m^W(B) > 0 \quad \textit{only if} \quad A \subset B \text{ or } B \subset A
\end{aligned}$$

In our classification context, an evidential dataset ED will be of the form

$$ED = m^{X,Y} = \begin{pmatrix} m_1^{X,Y} \\ \vdots \\ m_N^{X,Y} \end{pmatrix} = \begin{pmatrix} m_1^{X^1} & \cdots & m_1^{X^J} & m_1^Y \\ \vdots & \ddots & \vdots & \vdots \\ m_N^{X^1} & \cdots & m_N^{X^J} & m_N^Y \end{pmatrix}$$

where $m_i^{X,Y}$ describes the i^{th} sample with its uncertainty.

Evidential Likelihood Assume now we want to fit a parametric model with parameter θ to the data. Likelihood maximisation often provides a good estimator $\hat{\theta}$ of the unknown parameter θ. When data are uncertain, the likelihood can be re-written in the following way:

$$\begin{aligned}
\textit{precise likelihood:} \quad L(\theta; w) &= P_\theta(W = w) \\
\textit{imprecise likelihood:} \quad L(\theta; A) &= \sum_{w \in A} L(\theta; w) \\
\textit{evidential likelihood:} \quad L(\theta; m^W) &= \sum_{A \subseteq \Omega_W} m^W(A_i) L(\theta; A_i)
\end{aligned} \tag{1}$$

As shown in [9], this evidential likelihood can be maximised by the adaptation of the EM algorithm to belief functions: the E^2M algorithm. This algorithm is quite similar to the EM and is guaranteed to converge towards a local maximum. The main difference is at the Expectation step (E), where the measure used to compute the expectation is the conjunctive combination of P_θ and m^W.

2.3 Decision trees

Decision trees are basic classifiers widely used in many areas such as machine learning, statistics, data mining etc. They are usually built from precise datasets by partitioning the attribute space in a set of leaves, each leave being thus attached to some conditions on the attribute values and to a predicted class.

As each leave is characterized by the proportion of the population ”falling” into it as well as the frequencies of the class within this population, a decision tree can be viewed as a multinomial mixture whose parameters are the leaves probabilities (corresponding to the mixture coefficients) and the class probabilities inside the leaves (multinomial).

Example 1. Let us consider a data set with $N = 12$ data items, $J = 2$ attributes and $K = 3$ classes. Spaces are described as follows:

$$\Omega_{X^1} = [1,10], \quad \Omega_{X^2} = [13,150], \quad \Omega_Y = \{a,b,c\}.$$

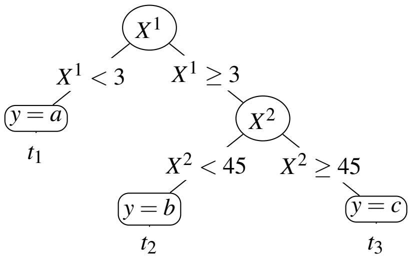

Fig. 1: Decision tree illustration

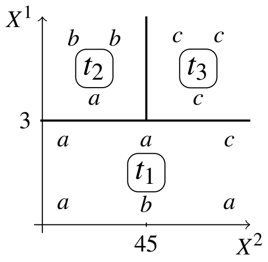

Fig. 2: Partitioned attribute space

Figures 1 and 2 respectively represent a possible decision tree and its corresponding partition of the attribute space with the learning dataset represented in it. To leaf t_2 are associated the values $A^1_2 \times A^2_2 = [3,10] \times [13,45[$, its estimated probability is $^3/_{12}$ and the class estimated probabilities are ($^1/_3$, $^2/_3$, 0), and the prediction for any test data item in t_2 (for example, $x^1 = 6$ and $x^2 = 30$) will be $y = b$.

The growing of a decision tree is recursive, and aims at separating classes inside leaves. This *separation* is usually measured by an impurity measure i computed from the leaves probabilities and the class distributions inside leaves. From a root node containing all the samples of the learning dataset, all possible splits (on all attributes and for all values) are explored, and the one with the highest purity gain (i.e., highest impurity reduction) is chosen. Then, for each newly created leave and the sample "falling" into it, the process is iterated until some stopping criteria is reached.

2.4 The E^2M decision trees

To learn decision trees from uncertain data (where potentially both attributes and classes can be uncertain), we proposed [8] to learn the multinomial mixture of the tree through the E^2M algorithm. Périnel [10] proposed a similar idea, yet only dealt with uncertain attributes and probabilistic data. We refer to [8] for technical details about the learning process of E^2M decision trees, as well as for some experiments on benchmark data comparing E^2M and *CART* decision trees and showing the potential interest of E^2M decision trees in terms of accuracy in noisy environments.

3 Application: the rubber quality problem

We first present the application and two uncertainty models we used on the data, before showing some interesting results. Note that the uncertainty models could be re-used in similar situations.

3.1 Problem description

Natural rubber is the result of a milky fluid transformation: the *latex* that is extracted from the *hevea* tree. Compared to synthetic rubber, natural rubber presents some unique physical properties but suffers from a pretty high variability in terms of quality, that experts are not fully able to explain and to control.

In order to better control this quality, one of the biggest rubber company has put some effort to study this variability in one of its plantations by collecting cultural data. This plantation is located in a part of Brazil where the climate has the particularity to be more *variable* than in the rest of Brazil (the natural origin of the *hevea* tree is in the Brazilian forests). However, due to the size of the plantation (approximatively 70 hectars) and various factors (e.g., untracked delay of collection/process due to weather conditions, tanks mixing productions of many parcels), some variables are subject to high uncertainties.

Data are constituted of many variables summarised in Table 1 (no Unit means a dimensionless variable). Meteorological data may influence the latex during three different periods: the latex fabrication by the tree (one week before tapping), the tapping day during which latex is collected, and the latex maturation in tanks (five days). For the temperature and the relative humidity, the minimum, median and maximum values are computed for each day.

The data set contains 3053 examples described by 106 attributes. The quality is measured by the P_{30} index which is an elasticity index. In order to use the E^2M decision trees methodology, the P_{30} was discretised into 5 equiprobable classes. This discretisation is presented in Table 2.

3.2 Data uncertainty modelling

Two types of uncertainty were modelled in this application: one relative to the rainfall, and one due to parcel mixture in tanks.

Rainfall uncertainty The rain is a phenomenon that is geographically very variable, especially in tropical areas. In the plantation, all the rain data come from a single meteorological station located inside the plantation. Since the plantation area is very large, it is sensible to make the hypothesis that the farther is located a parcel in the plantation from the meteorological station, the more uncertain is its rainfall data. This uncertainty is non-probabilistic and progressive, so we chose to model it with a consonant mass function. Moreover, as more rainfall implies more uncertainty, it is logical to assume that imprecision of focal elements increases multiplicatively (i.e. more rainfall measured by the station implies wider focal elements).

To keep the complexity of the uncertainty reasonable, we limited the mass to five focal elements of the form: $[w(1-\delta), w(1+\delta)]$ where w is the original precise rainfall data and where $\delta \in \Delta = \{0, 0.25, 0.5, 0.75, 1\}$. The proposed model is easy to expand to more than five focal elements and can therefore accommodate various levels of complexity (depending on the available computational power and on possible time constraints).

variable	type	category	unit
period	categorical	agronomical	
season	categorical	climatic	
weight	numerical	agronomical	kg
X (latitude)	numerical	geographical	km
Y (longitude)	numerical	geographical	km
clone	categorical	agronomical	
panel	categorical	agronomical	
tapping system	categorical	agronomical	
surface	numerical	agronomical	hectar
planting year	numerical	agronomical	year
first tapping year	numerical	agronomical	year
tapping age	numerical	agronomical	year
annual number of tapped trees	numerical	agronomical	
tapped tree per hectar	numerical	agronomical	
temperature	numerical	climatic	celcius degrees
relative humidity	numerical	climatic	
sun hours	numerical	climatic	hours
rainfall	numerical	climatic	mm
P30	numerical	agronomical	

Table 1: variables characteristics

class labels	P_{30} range
very bad	[1.87 ; 14.7[
bad	[14.7 ; 21.6[
medium	[21.6 ; 27.4[
good	[27.4 ; 32.9[
very good	[32.9 ; 49.5[

Table 2: P_{30} discretisation

We define a function $g : [0, d_{max}] \times \Delta \to [0,1]$ (d_{max} being the maximal distance between a location of interest and the measurement station) such that the rainfall w of a parcel located at a distance d from the meteorological is characterized by $g(d,\delta) = m^W([w(1-\delta), w(1+\delta)])$ for all $\delta \in \Delta$.

We distinguish two types of focal elements, the most precise ones ($\delta < 0.5$), and the most imprecise ones ($\delta \geq 0.5$). In the first case (precise ones), we assume that the farther was the parcel from the meteorological station (d increasing), the smaller had to be the masses assigned to those focal elements. In the second, we want to assign bigger masses to the farther parcels. Such assumptions can be translated in the following constraints:

$$\begin{cases} \delta < 0.5 \to \frac{\partial g}{\partial d} < 0 \\ \delta \geq 0.5 \to \frac{\partial g}{\partial d} > 0 \end{cases} \tag{2}$$

that merely translate the assumption into derivative behaviors. As function g is used to define mass functions, we must add the following constraints

$$\forall (d,\delta) \in [0,d_{max}] \times \Delta, \quad g(d,\delta) \geq 0 \tag{3}$$

$$\forall d \in [0,d_{max}], \quad \sum_{\delta \in \Delta} g(d,\delta) = 1 \tag{4}$$

that simply ensure one that the obtained mass functions will be well-defined, i.e. that it will be positive (constraint (3)) and will sum up to one (constraint (4)).

One simple solution of this problem is to use two linear functions, one increasing for the most precise focal elements ($\delta < 0.5$) and one decreasing for the most imprecise ones ($\delta \geq 0.5$), and to use a convex sum of those two functions. We obtain:

$$g(d,\delta) = \delta(\frac{2d}{5d_{max}}) + (1-\delta)(\frac{2}{5} - \frac{2d}{5d_{max}}) \tag{5}$$

Example 2. Consider three rainfall measurements $w_1 = 0$, $w_2 = 10$ and $w_3 = 30$ from the station, and for each of these measurements some corresponding parcels of interest respectively distant of $20km$, $50km$ and $2km$ from the station. Assuming that $d_{max} = 80$, we obtain $m^{w_1}(\{0\}) = 1$, given the multiplicative uncertainty, and

$$\begin{cases} m^{w_2}(\{10\}) &= 0.15 \\ m^{w_2}([7.5,12.5]) &= 0.175 \\ m^{w_2}([5,15]) &= 0.200 \\ m^{w_2}([2.5,17.5]) &= 0.225 \\ m^{w_2}([0,20]) &= 0.250 \end{cases} \quad \begin{cases} m^{w_3}(\{30\}) &= 0.39 \\ m^{w_3}([22.5,37.5]) &= 0.295 \\ m^{w_3}([15,45]) &= 0.2 \\ m^{w_3}([7.5,52.5]) &= 0.105 \\ m^{w_3}([0,60]) &= 0.01. \end{cases}$$

The absence of rain is thus considered certain (m^{w_1}) whereas positive rainfall data masses are concentrated on imprecise focal elements when coming from distant parcels (m^{w_2}) and on more precise ones when coming from parcels close from the meteorological station (m^{w_3}).

Parcelar mixtures uncertainty During the harvest, the latex coming from many parcels is usually mixed in some tank. All the parcel variables (i.e., most agronomical variables of Table 1) we have are therefore subject to uncertainty as the amount of latex coming from each parcel in tanks is not tracked. During a pre-treatment of the data, we therefore split all those parcel variables into rough and uncertain proportions (due to latex production high variability) computed from the weight of latex produced annually by each parcel (shifting from 18 original attributes to 106, with all the split ones being in range $[0,1]$).

For example, if 25% of a tank content comes from clone A parcels and 75% from clone B parcels, the actual amount of clones A and B latex in the tank may be quite different, as each clone has variable production capacities (that may depend differently on the weather, soil conditions, etc.). We model this uncertainty such that the more balanced are the parcel proportions in a tank, the more uncertain become those proportions: proportions of a tank with latex from only one pure parcel should remain certain, while proportions of a tank with equal amounts of latex coming from different parcels should be maximally uncertain.

To do that we used simple intervals around computed crisp proportions, with the interval width increasing as proportions become uniform. To measure this uniformity we simply used the Shannon entropy denoted *ent* computed on the set of parcel proportions of the tanks. The obtained model is the following: for each parcel variable X^j having r modalities with the positive proportions $\{p_1,\dots,p_r\}$,

$$\begin{cases} m^{j_1}([\max(p_1-\frac{ent(p_1,\dots,p_r)}{r},0),\min(p_1+\frac{ent(p_1,\dots,p_r)}{r},1)]) = 1 \\ \vdots \\ m^{j_r}([\max(p_r-\frac{ent(p_1,\dots,p_r)}{r},0),\min(p_r+\frac{ent(p_1,\dots,p_r)}{r},1)]) = 1 \end{cases} \qquad (6)$$

with m^{j_i} modelling the uncertainty about the j_ith proportion.

Example 3. Let us consider a tank containing 75% of clone A and 25% of clone B. The entropy on those proportions is equal to 0.8113. The obtained masses are therefore

$$\begin{cases} m^{cloneA}([34.43\%,100\%]) = 1 \\ m^{cloneB}([0\%,65.57\%]) = 1 \end{cases}$$

3.3 Experiments

In order to see the consequences of integrating data uncertainty the way we described in Section 3.2, we perform some experiments comparing standard *CART* trees and E^2M trees on the original precise dataset and its corresponding evidential dataset obtained with our uncertainty models, respectively.

For both methodologies, the stopping criteria is a maximum of 5 leaves (to preserve a high interpretability) and a relative purity gain of 0.05.The error rates were computed as the proportion of misclassified examples in the test dataset. Their means were computed from ten 3-fold cross validations. It is noticeable that we used standard (precise error rates) even for the E^2M decision trees for comparison purposes. Given the small tree size, no pruning is done.

methodology	mean error rates	95% confidence interval
CART	0.6607	[0.6314 ; 0.6900]
E^2M	0.6560	[0.6266 ; 0.6854]

Table 3: Results

As shown in Table 3, the accuracies of the two methodologies are quite similar, even if the E^2M accuracy is slightly better. Let us now shift to the main interesting part for the experts (and hence for the application goal): the interpretation of the trees.

3.4 Interpretation

In a pure knowledge discovery concern, we learnt *CART* and E^2M decision trees on the whole dataset in order to compare the informations they provide about the natural rubber quality explanation. Within such a goal, it should also be noted that integrating data uncertainty makes the result somehow more faithful to our actual information (and therefore more reliable for the experts). The learning parameters were exactly the same as in Section 3.3.

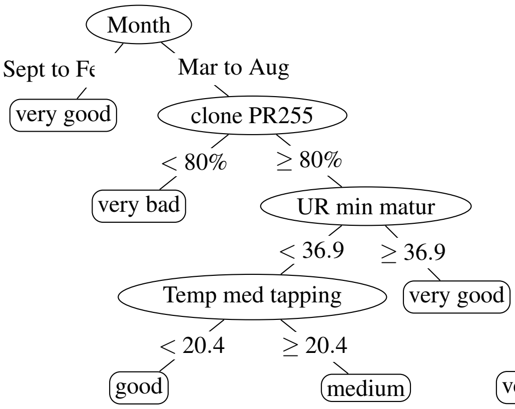

Fig. 3: *CART* decision tree

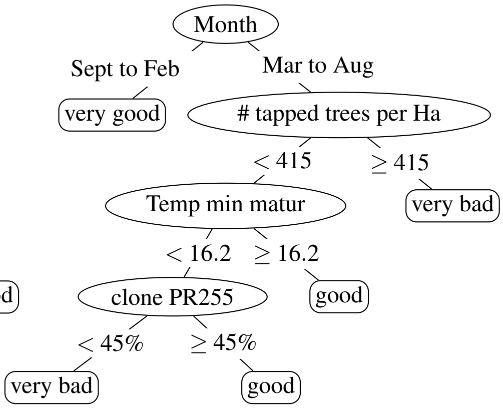

Fig. 4: E^2M decision tree

In Figures 3 and 4 that show the learning results, UR min matur is the minimum of relative humidity during the maturation of the latex in buckets, Temp med (resp. min) tapping is the medium (resp. minimum) temperature during the tapping day, # of tapped tree per Ha is the number of tapped trees per hectare.

As shown in those figures, integrating data uncertainty makes the number of tapped trees per hectare appear in the tree. Not only the E^2M decision tree suggests a high relation between quality and productivity, but it also provides a density acceptable bound inside the plantation. According to the experts, this issue points out the potential need to investigate further this path. We can also notice that this data uncertainty lessen the role of clone PR255 proportion in the rubber quality explanation.

4 Discussion

In the rubber application, the prediction results from *CART* and E^2M decision trees do not seem very different in term of accuracy (even if E^2M has a slight advantage), but the interpretation of the trees can be quite different; this latter interpretation being as important as accuracy in many applications. Indeed, we observe that some variables may play a significant role in rubber quality explanation once their uncertainty is modelled (here by belief functions). Modelling uncertainty also provides more trustful results for the experts.

Perspectives concerning the application include the modelling of additional uncertainties (in particular with respect to the tapping day) as well as additional studies involving more variables (e.g., tree diseases, soil conditions) or other prediction goal (e.g., quantity of production, which is also important from an economical viewpoint).

Finally, as the uncertainty models we have introduced may be useful in other areas (particularly the distance-based model of rainfall), we plan to generalize them and study in more details their properties.

Acknowledgments. This work has been done thanks to CIRAD fundings and LASB laboratory support. We also wish to greatly thank Dr. Patrice Loisel for his great help on the rainfall uncertainty model.

References

1. L. Zadeh, “Fuzzy sets as a basis for theory of possibility,” *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.
2. A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.
3. P. Walley, “Statistical reasoning with imprecise probabilities,” 1991.
4. S. Kanj, F. Abdallah, and T. Denœux, “Evidential multi-label classification using the random k-label sets approach,” pp. 21–28, 2012.
5. N. Sutton-Charani, S. Destercke, and T. Denœux, “Classification trees based on belief functions,” pp. 77–84, 2012.
6. C. Z. Janikow, “Fuzzy decision trees: issues and methods.” *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, vol. 28, pp. 1–14, 1998.
7. G. Shafer, “A mathematical theory of evidence.” 1976.
8. N. Sutton-Charani, S. Destercke, and T. Denœux, “Learning decision trees from uncertain data with an evidential em approach.”
9. T. Denœux, “Maximum likelihood estimation from uncertain data in the belief function framework,” *IEEE Trans. on Know. and Data Eng.*, vol. 25, pp. 119–130, 2011.
10. E. Périnel, “Construire un arbre de discrimination binaire à partir de données imprécises,” *Revue de statistique appliquée*, vol. 47, pp. 5–30, 1999.