Training and Evaluating Classifiers from Evidential Data: Application to E^2M Decision Tree Pruning

Nicolas Sutton-Charani, Sébastien Destercke, and Thierry Denœux

Université Technologie de Compiègne UMR 7253 Heudiasyc
60203 COMPIEGNE Cedex France
{nicolas.sutton-charani,sebastien.destercke,
t.denoeux}@hds.utc.fr
http://www.hds.utc.fr/

Abstract. In many application data are imperfect, imprecise or more generally uncertain. Many classification methods have been presented that can handle data in some parts of the learning or the inference process, yet seldom in the whole process. Also, most of the proposed approach still evaluate their results on precisely known data. However, there are no reason to assume the existence of such data in applications, hence the need for assessment method working for uncertain data. We propose such an approach here, and apply it to the pruning of E^2M decision trees. This results in an approach that can handle data uncertainty wherever it is, be it in input or output variables, in training or in test samples.

Keywords: classification, uncertain data, E^2M algorithm, error rate, belief functions, E^2M decision trees, pruning.

1 Introduction

Data uncertainty can have many origins: measurements approximations, sensor failures, subjective expert assessments, etc. Taking into account this uncertainty to learn a classifier is challenging because of the analytical and computational difficulties to extend standard statistical learning methodologies to uncertain data. However, in the past years, several approaches [6,3] have been proposed to learn model from uncertain data.

Once a classifier is built from a learning (uncertain) samples, it is usually evaluated by a *misclassification* or *error* rate which is computed from test samples. This error rate corresponds to the probability of misclassification and is estimated by the frequency of misclassified test samples. However, even in methods dealing with uncertain data, this misclassification rate is usually computed using precise test samples. This can be explained by the absence of genuine uncertain benchmark datasets, that remain to be built.

Yet, there is no reason to separate the training and the learning data by making only the former uncertain. In practice, one should be able to tackle uncertainty in all the data sets, without distinction. This is the main issue tackled in this paper, in which we propose a means to evaluate classifiers and models from uncertain test data. The uncertain data, from now on called *evidential data*, will be modelled by the means of belief functions, that offer a flexible framework to model epistemic uncertainty.

F. Cuzzolin (Ed.): BELIEF 2014, LNAI 8764, pp. 87–94, 2014.

We will use the evidential likelihood [3] as a generic tool to learn and to assess probabilistic models from such evidential data.

In addition, we will apply our proposition to the E^2M decision trees classification model [8], which is a decision tree methodology adapted to uncertain learning data modelled by belief functions. It is learned using the E^2M algorithm [3] which is an extension of the well known EM algorithm to evidential data. Our proposal will be applied in two different manners: to prune E^2M decision trees, and to evaluate the resulting classifiers. Indeed, pruning requires to evaluate the pruned trees performances, hence to potentially evaluate them on evidential data in the case of E^2M decision trees.

Section 2 introduces the problem of learning under evidential data, and recalls the evidential likelihood approach, together with the E^2M decision tree approach. In Section 3 we give the details of the evidential error rate estimations and in Section 4 a E^2M pruning procedure is proposed and some experiments are presented. Apart from solving the evaluation problem with evidential data, it will also provides us with a classification technique able to handle uncertain data at all levels, both in training and in test phases.

2 Background

This section briefly reminds required elements to understand the paper. Further details can be found in complementary papers [3,8]

2.1 Classification under Evidential Data

The goal of a classification technique is to learn a mapping $\mathscr{C}$ from J attributes $X = \{X_1,\ldots,X_J\} \in \Omega_1 \times \cdots \times \Omega_J = \Omega_X$ to a class $Y \in \Omega_Y = \{C_1,\ldots,C_K\}$. Classically, this is done using a set of n learning precise data (x,y). In this paper, we consider evidential data, meaning that each datum is modelled by a mass function on Ω_X (for the input uncertainty) and Ω_Y (for the class uncertainty). Recall that a mass function on a space Ω is a positive mass $m : 2^\Omega \to [0,1]$ defined on Ω power set such that $\sum_{E\subseteq\Omega, E\neq\emptyset} m(E) = 1$. The contour function[1] $pl : \Omega \to [0,1]$ induced by it is $pl(\omega) = \sum_{\omega\in E} m(E)$.

We consider that this classifier $\mathscr{C}$ is learned from an evidential learning set of n samples

$$(m^x_\ell, m^y_\ell) = \begin{pmatrix} m^x_{1,\ell} & m^y_{1,\ell} \\ \vdots & \vdots \\ m^x_{n,\ell} & m^y_{n,\ell} \end{pmatrix}$$

and is evaluated using an evidential test sample of n' samples

$$(m^x_t, m^y_t) = \begin{pmatrix} m^x_{1,t} & m^y_{1,t} \\ \vdots & \vdots \\ m^x_{n',t} & m^y_{n',t} \end{pmatrix}.$$

While data are assumed to be evidential, we want to learn a probabilistic parametric classifier with parameters θ providing for an (evidential) entry m^x a probability $P_\theta(Y|m^x)$, the decided class then corresponding to $\mathscr{C}(m^x) = \arg\max_{C_i\in\Omega_Y} P_\theta(C_i|m^x)$.

[1] No other notions will be needed in this paper.

2.2 Evidential Likelihood and E^2M Algorithm

In standard probability theory, the likelihood $L(\theta;w)$ of a parameter θ given a sample w corresponds to the probability $P_\theta(W=w)$ of observing this sample given that parameter. Maximising this likelihood provides good estimators of the parameter value. Denoeux [3] has extended this concept to evidential data.

When $w \in A$ is imprecisely observed, then an imprecise likelihood corresponding to the probability to pick a sample *inside* A in the population can be computed as

$$L(\theta;A) = \sum_{w \in A} L(\theta;w) = P_\theta(W \in A)$$

If our knowledge about w is not only imprecise but also uncertain and modelled by a mass function m^w having $A_1,\ldots,A_z$ as focal elements, the evidential likelihood of the parameter becomes

$$L(\theta;m^w) = \sum_{i=1}^{z} m^w(A_i) L(\theta;A_i). \tag{1}$$

In general, finding the (global) value θ maximizing Eq. (1) is difficult, as the function is non-convex and complex. However, the E^2M algorithm provides a means to obtain a local maximum of (1). It is an iterative algorithm very similar to the EM algorithm [2], the main difference is the measure used to compute expectations at the E step. In order to take into account both the available knowledge (represented by m^w) and the model aleatory uncertainty, the E step uses the conjunctive combination $P(.\mid\theta,m^w) := P_\theta \bigcirc\!\!\!\!\cap\, m^w$, which is a probability measure, to compute the expectation. Algorithm 1 summarizes the E^2M algorithm.

Algorithm 1. Estimation with the E^2M algorithm

Input: $\theta^{(0)}, \gamma$
Output: final θ

$r = 1$;
repeat
  E-step: $Q(\theta,\theta^{(r)}) = E[log(L(\theta;W)) \mid \theta^{(r)}, m^W]$;
  M-step: $\theta^{(r+1)} = \arg\max_{\theta \in \Theta} Q(\theta,\theta^{(r)})$;
  $r = r+1$;
until $\frac{L(\theta^{(r)};m^w) - L(\theta^{(r-1)};m^w)}{L(\theta^{(r-1)};m^w)} < \gamma$;
$\theta = \theta^{(r)}$;

2.3 E^2M Decision Trees

Decision trees or more precisely classification trees are famous classifiers that provide interpretable outputs [1]. They recursively partition the space Ω_X into leaves that contains probabilities over the classes Ω_Y.

The purity of a leaf t_h (defining a subset of Ω_X) is usually evaluated by some impurity measure such as Shanon entropy $i(t_h) = -\sum_{k=1}^{K} \alpha_h^k \log(\alpha_h^k)$ where $\alpha_h^k = P(Y = C_k \mid t_h)$. The purity gain obtained by splitting t_h into t_L and t_R is computed as $\delta i = i(t_h) - \pi_L i(t_L) - \pi_R i(t_R)$ where $\pi_L = P(t_L \mid t_h)$ and $\pi_R = P(t_R \mid t_h)$ are the probabilities of being in each children leaves. In usual approaches the leaves probabilities π_h and the class probabilities inside leaves α_h^k are estimated by frequencies of learning samples in leaves and of their different class labels inside the leaves:

$$\tilde{\pi_h} = \frac{n(t_h)}{n} \qquad \tilde{\alpha_h^k} = \frac{n_k(t_h)}{n(t_h)}$$

where n is the number of learning samples, $n(t_h)$ is the number of learning sample in the leaf t_h and $n_k(t_h)$ is the number of learning samples of class C_k inside the leaf t_h.

E^2M decision trees [8] are an extension of classical decision trees to evidential data. The main idea is to see the tree as a mixture (the leaves weights π_h) of multinomial distributions (the class probabilities α_h^k), and to learn this probabilistic model using the E^2M. We proposed to estimate the probabilities of leaves and of class in leaves by maximising their likelihood in regard to the uncertain learning sample (m_ℓ^x, m_ℓ^y). We have:

$$\{\hat{\pi_h}, \hat{\alpha_h^k}\}_{h,k} = \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; (m_\ell^x, m_\ell^y))$$

Within decision trees techniques, *pruning* is a classical way to avoid over-fitting and that are usually based on a compromise between interpretability and accuracy [1,4]. Most of them consider possible sub-trees of the initial learned tree, and pick one satisfying an evaluation criteria always based (at least partially) on classification accuracy. Yet, evidential data do not allow a straightforward computation of accuracy, hence a need of new evaluation techniques to be able to prune.

3 Uncertain Classifiers Evaluation with the E^2M Algorithm: Evidential Error Rates Estimation

While techniques introduced in the previous sections allow to learn from evidential data (see also [5]), the problem of evaluating classifiers with an evidential test data set (m_t^x, m_t^y) remains. This section proposes a solution also exploiting the evidential likelihood and E^2M algorithm.

Let $E \in \{0,1\}$ be an aleatory *Bernoulli* variable representing the *misclassification* of $\mathscr{C}$, equal to 1 in case of misclassification, 0 in case of good classification. We have $E \sim Ber(\varepsilon)$ where ε is the *misclassification* rate, i.e., $P(Y \neq \mathscr{C}(x)|x)$.

With precise data, ε, whose estimation $\tilde{\varepsilon}$ is the frequency of misclassified examples in the learning test and corresponds to maximising its likelihood $L(\theta; e = \{e_1, \ldots, e_{n'}\})$ with $e_i = 0$ if $\mathscr{C}(x_{i,t}) \neq y_{i,t}$, 1 otherwise. We therefore get $\tilde{\varepsilon} = {}^{n_1}/_{n'}$ where n_1 is and the number of 1 in e.

In practice, when one has evidential data, the E^2M model still provides a unique prediction $\mathscr{C}(m^x_{i,t})$, which has to be compared to an evidential output $m^y_{i,t}$. In practice, each $m^y_{i,t}$ can be mapped to a mass function m^e_i over $\{0,1\}$ such that

$$m^e_i(\{0\}) = m^y_{i,t}(\mathscr{C}(m^x_{i,t})) \tag{2}$$

$$m^e_i(\{1\}) = \sum_{E\subseteq\Omega_y, \mathscr{C}(m^x_{i,t})\notin E} m^y_{i,t}(E) \tag{3}$$

$$m^e_i(\{0,1\}) = \sum_{E\subseteq\Omega_y, \mathscr{C}(m^x_{i,t})\in E, |E|>1} m^y_{i,t}(E) \tag{4}$$

Given this sample, the evidential accuracy can be computed as follows:

$$L(\varepsilon; m^e) = \prod_{i=1}^{n}[(1-\varepsilon)pl_i(0) + \varepsilon pl_i(1)] \tag{5}$$

$$Q(\varepsilon; \hat{\varepsilon}^{(q)}) = nlog(1-\varepsilon) + log(\frac{\varepsilon}{1-\varepsilon})\sum_{i=1}^{n}\xi_i^{(q)} \tag{6}$$

$$\hat{\varepsilon}^{(r+1)} = \underset{\varepsilon\in[0,1]}{\text{argmax}}\, Q(\varepsilon; \hat{\varepsilon}^{(q)}) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{(q)} \tag{7}$$

where

$$\xi_i^{(q)} = E[E_i \,|\, \hat{\varepsilon}^{(q)}; m^e_i] = \frac{\hat{\varepsilon}^{(q)} pl_i(1)}{(1-\hat{\varepsilon}^{(q)})pl_i(0) + \hat{\varepsilon}^{(q)} pl_i(1)}$$

with $pl_i(0) = Pl(\{e_i = 0\}) = m^e_i(\{0\}) + m^e_i(\{1,0\})$ and $pl_i(1) = Pl(\{e_i = 1\}) = m^e_i(\{1\}) + m^e_i(\{1,0\})$

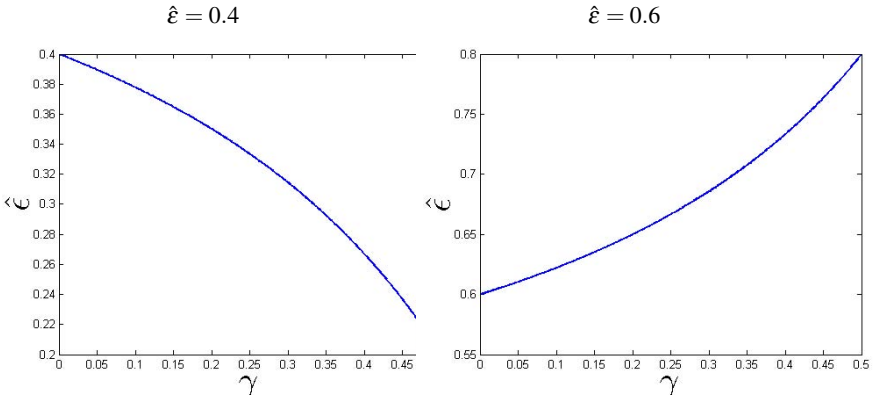

Fig. 1. Variations of the evidential error rate $\hat{\varepsilon}$ with the uncertainty level γ when $\tilde{\varepsilon} = 0.4$ and 0.6

As an illustration, Figure 1 represents the variation of the evidential error rate in function of $m^e_i(\{0,1\}) = \gamma$ for $n' = 100$ samples, and where the proportion of samples

where $m_i^e(\{1\}) = 1-\gamma$ versus samples where $m_i^e(\{0\}) = 1-\gamma$ is given by the precise error rates $\hat{\varepsilon}$ 0.4 and 0.6. Interestingly we can see that the estimation, by privileging the most present observation, tends to accentuate either the quality of accurate models ($\hat{\varepsilon} < 0.5$) or the unreliability of inaccurate ones. We can therefore expect this evidential accuracy to provide reliable choices.

4 Application: Pruning of E^2M Decision Trees

This section illustrates the evidential error rate to the pruning of E^2M decision trees. Considering the sequence of sub-trees induced by successive splits, we simply pick the one that obtains the smallest evidential error rate on a pruning sample (different from the initial learning sample). Indeed, our goal is not to define optimal pruning criterion, but simply to illustrate the use of evidential error rates.

Our experiments concern five precise benchmark datasets (coming from UCI) in which we artificially injected uncertainty. For each observation w_i (attribute and class) of the precise datasets, a noise level γ_i was uniformly sampled in $[0,1]$. A number u was then uniformly sampled on $[0,1]$, if $u < \gamma_i$ then the (noised) value w_i is replaced by another value w_i^* drawn uniformly from Ω_W (either attribute or class spaces), otherwise $w_i^* = w_i$. Obtained evidential data are $m(w_i^*) = 1-\gamma_i$ and $m(\Omega_W) = \gamma_i$.

We learnt simultaneously standard *CART* decision trees and E^2M ones and compared their error rates. For each test we learnt the trees on one third of the datasets, pruned them on another third and test them on the left third by computing standard error rates and evidential ones. All computations are achieved on noised data (considering crisp replacements for CART and evidential ones for E^2M) The stopping criteria were composed of a maximum of 10 leaves and a relative minimum purity gain of 5%.

Table 1. Comparison of *CART* and E^2M decision trees efficiency before and after pruning

data set	classical error rate					evidential error rate				
	naive	*CART*		E^2M		*naive*	*CART*		E^2M	
		unpruned	pruned	unpruned	pruned		unpruned	pruned	unpruned	pruned
iris	0.67	0.60	0.60	0.57	0.58	0.79	0.65	0.66	0.59	0.60
balance scale	0.60	0.60	0.60	0.58	0.58	0.63	0.62	0.62	0.51	0.51
wine	0.65	0.61	0.62	0.60	0.60	0.75	0.64	0.67	0.64	0.64
glass	0.68	0.69	0.68	0.68	0.67	0.74	0.73	0.73	0.67	0.67
e.coli	0.72	0.73	0.72	0.74	0.73	0.75	0.74	0.74	0.71	0.70

Table 1 summarizes the means of error rates obtained for 100 tests for each dataset. For each methodology the error rate are compared before the learning (the *naive* error rate is obtained by predicting systematicaly the class the most frequent in the learning sample), once the trees are learnt but before pruning and after pruning. The high error rates are due to noise both in the learning and in the testing phases.

Both evidential and classical error rates are slightly smaller for E^2M trees than for *CART* ones. If this is not surprising for the evidential error rate after pruning (as it is the minimized criterion), the other better scores confirm the interest of using evidential approaches. The pruning strategy also increases accuracy for the *balance* and *glass* datasets, despite the small size of the learnt trees. E^2M trees appear to be naturally smaller than *CART* ones but can still be pruned thank to the evidential error rates computations.

Table 2. Comparison of *CART* and E^2M decision trees sizes before and after pruning

data set	*CART*				E^2M			
	before pruning		after pruning		before pruning		after pruning	
	# *failures*	# leaves	# *failures*	# leaves	# *failures*	# leaves	# *failures*	# leaves
iris	3	9.57	13	4.57	0	4.36	6	3.39
balance scale	99	1.01	99	1.01	0	7.01	0	5.21
wine	0	10	15	4.79	0	4.05	10	3.06
glass	52	5.26	70	2.08	0	6.92	14	4.46
e.coli	52	5.26	70	2.08	0	6.92	14	4.46

Table 2 compares the size of the *CART* and E^2M trees before and after pruning. A learning failure occurs when the *noised* dataset does not enable any first split in regards to the stopping criteria. *CART* trees appears to be bigger than E^2M ones before and after pruning. We can interpret this as an impact of the data uncertainty on the complexity of the learnt model. In deed, it is not necessary to have a complex model with partially unreliable data.

5 Conclusions

We have introduced a way, through the notion of evidential likelihood, to evaluate classifier in presence of uncertain (evidential) data. Such a technique appears as essential and necessary if we want to fully tackle the problem of uncertain data, as assuming uncertain learning data and certain test data (at least in the output), if valid on benchmark data sets, seems unrealistic in practical applications. We have also tested our approach on the E^2M decision trees, and doing so have proposed, to our knowledge, the first method that is able to handle data uncertainty in attributes and classes, both in learning and testing.

As perspective, it would be interesting to compare our study to other approaches, both from a theoretical and practical standpoint. For example, we could compare ourselves to the strategy consisting of transforming evidential data into probabilistic one through the pignistic transform [7].

References

1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees (1984)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistial society, series B 39(1), 1–38 (1977)
3. Denœux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans. on Know. and Data Eng. (2011)
4. Esposito, F., Malerba, D., Semeraro, G., Kay, J.: A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(5), 476–491 (1997)
5. Masson, M.H., Denoeux, T.: Ecm: An evidential version of the fuzzy *c*-means algorithm. Pattern Recognition 41(4), 1384–1397 (2008)
6. Périnel, E.: Construire un arbre de discrimination binaire à partir de données imprécises. Revue de statistique appliquée 4747, 5–30 (1999)
7. Smets, P.: Belief induced by the partial knowledge of the probabilities. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, UAI 1994, pp. 523–530. Morgan Kaufmann Publishers Inc., San Francisco (1994)
8. Sutton-Charani, N., Destercke, S., Denœux, T.: Learning decision trees from uncertain data with an evidential em approach. In: 12th International Conference on Machine Learning and Applications, ICMLA (2013)