Aleatoric and Epistemic Uncertainty with Random Forests

Mohammad Hossein Shaker[1] and Eyke Hüllermeier[1]

Heinz Nixdorf Institute and Department of Computer Science
Paderborn University, Germany
{mhshaker,eyke}@upb.de

Abstract. Due to the steadily increasing relevance of machine learning for practical applications, many of which are coming with safety requirements, the notion of uncertainty has received increasing attention in machine learning research in the last couple of years. In particular, the idea of distinguishing between two important types of uncertainty, often refereed to as *aleatoric* and *epistemic*, has recently been studied in the setting of supervised learning. In this paper, we propose to quantify these uncertainties with random forests. More specifically, we show how two general approaches for measuring the learner's aleatoric and epistemic uncertainty in a prediction can be instantiated with decision trees and random forests as learning algorithms in a classification setting. In this regard, we also compare random forests with deep neural networks, which have been used for a similar purpose.

Keywords: machine learning · uncertainty · random forest

1 Introduction

The notion of uncertainty has received increasing attention in machine learning research in the last couple of years, especially due to the steadily increasing relevance of machine learning for practical applications. In fact, a trustworthy representation of uncertainty should be considered as a key feature of any machine learning method, all the more in safety-critical application domains such as medicine [22,9] or socio-technical systems [19,20].

In the general literature on uncertainty, a distinction is made between two inherently different sources of uncertainty, which are often referred to as *aleatoric* and *epistemic* [4]. Roughly speaking, aleatoric (*aka* statistical) uncertainty refers to the notion of randomness, that is, the variability in the outcome of an experiment which is due to inherently random effects. The prototypical example of aleatoric uncertainty is coin flipping. As opposed to this, epistemic (*aka* systematic) uncertainty refers to uncertainty caused by a lack of knowledge, i.e., it relates to the epistemic state of an agent or decision maker. This uncertainty can in principle be reduced on the basis of additional information. In other words, epistemic uncertainty refers to the *reducible* part of the (total) uncertainty, whereas aleatoric uncertainty refers to the *non-reducible* part.

More recently, this distinction has also received attention in machine learning, where the "agent" is a learning algorithm [18]. In particular, a distinction between aleatoric and epistemic uncertainty has been advocated in the literature on deep learning [6], where the limited awareness of neural networks of their own competence has been demonstrated quite nicely. For example, experiments on image classification have shown that a trained model does often fail on specific instances, despite being very confident in its prediction. Moreover, such models are often lacking robustness and can easily be fooled by "adversarial examples" [14]: Drastic changes of a prediction may already be

provoked by minor, actually unimportant changes of an object. This problem has not only been observed for images but also for other types of data, such as natural language text [17].

In this paper, we advocate the use of decision trees and random forests, not only as a powerful machine learning method with state-of-the-art predictive performance, but also for measuring and quantifying predictive uncertainty. More specifically, we show how two general approaches for measuring the learner's aleatoric and epistemic uncertainty in a prediction (recalled in Section 2) can be instantiated with decision trees and random forests as learning algorithms in a classification setting (Section 3). In an experimental study on uncertainty-based abstention (Section 4), we compare random forests with deep neural networks, which have been used for a similar purpose.

2 Epistemic and Aleatoric Uncertainty

We consider a standard setting of supervised learning, in which a learner is given access to a set of (i.i.d.) training data $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ the set of outcomes that can be associated with an instance. In particular, we focus on the classification scenario, where $\mathcal{Y} = \{y_1, \ldots, y_K\}$ consists of a finite set of class labels, with binary classification ($\mathcal{Y} = \{0, 1\}$) as an important special case.

Suppose a *hypothesis space* $\mathcal{H}$ to be given, where a hypothesis $h \in \mathcal{H}$ is a mapping $\mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$, i.e., a hypothesis maps instances $\boldsymbol{x} \in \mathcal{X}$ to probability distributions on outcomes. The goal of the learner is to induce a hypothesis $h^* \in \mathcal{H}$ with low risk (expected loss)

$$R(h) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(\boldsymbol{x}), y) \, d\, P(\boldsymbol{x}, y) \ , \tag{1}$$

where P is the (unknown) data-generating process (a probability distribution on $\mathcal{X} \times \mathcal{Y}$), and $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ a loss function. This choice of a hypothesis is commonly guided by the empirical risk

$$R_{emp}(h) := \frac{1}{N} \sum_{i=1}^{N} \ell(h(\boldsymbol{x}), y) \ , \tag{2}$$

i.e., the performance of a hypothesis on the training data. However, since $R_{emp}(h)$ is only an estimation of the true risk $R(h)$, the empirical risk minimizer (or any other predictor)

$$\widehat{h} := \operatorname*{argmin}_{h \in \mathcal{H}} R_{emp}(h) \tag{3}$$

favored by the learner will normally not coincide with the true risk minimizer (Bayes predictor)

$$h^* := \operatorname*{argmin}_{h \in \mathcal{H}} R(h) \, . \tag{4}$$

Correspondingly, there remains uncertainty regarding h^* as well as the approximation quality of $\widehat{h}$ (in the sense of its proximity to h^*) and its true risk $R(\widehat{h})$.

Eventually, one is often interested in the *predictive uncertainty*, i.e., the uncertainty related to the prediction $\widehat{y}_q$ for a concrete query instance $\boldsymbol{x}_q \in \mathcal{X}$. In other words, given a partial observation $(\boldsymbol{x}_q, \cdot)$, we are wondering what can be said about the missing outcome, especially about the uncertainty related to a prediction of that outcome. Indeed, estimating and quantifying uncertainty in a transductive way, in the sense of tailoring it to individual instances, is arguably important and practically more relevant than a kind of average accuracy or confidence, which is often reported in machine learning.

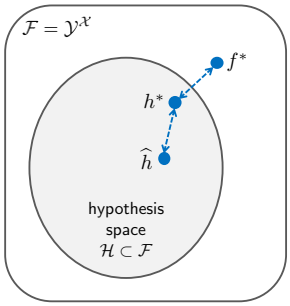


Fig. 1: Different types of uncertainties related to different types of discrepancies and approximation errors: f^* is the pointwise Bayes predictor, h^* is the best predictor within the hypothesis space, and $\widehat{h}$ the predictor produced by the learning algorithm.

As the prediction $\widehat{y}_q$ constitutes the end of a process that consists of different learning and approximation steps, all errors and uncertainties related to these steps may also contribute to the uncertainty about $\widehat{y}_q$ (cf. Fig. 1):

- Since the dependency between $\mathcal{X}$ and $\mathcal{Y}$ is typically non-deterministic, the description of a new prediction problem in the form of an instance $\boldsymbol{x}_q$ gives rise to a conditional probability distribution
$$p(y \,|\, \boldsymbol{x}_q) = \frac{p(\boldsymbol{x}_q, y)}{p(\boldsymbol{x}_q)} \tag{5}$$
on $\mathcal{Y}$, but it does normally not identify a single outcome y in a unique way. Thus, even given full information in the form of the measure P (and its density p), uncertainty about the actual outcome y remains. This uncertainty is of an *aleatoric* nature. In some cases, the distribution (5) itself (called the predictive posterior distribution in Bayesian inference) might be delivered as a prediction. Yet, when having to commit to a point estimate, the best prediction (in the sense of minimizing the expected loss) is prescribed by the pointwise Bayes predictor f^*, which is defined by
$$f^*(\boldsymbol{x}) := \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \int_{\mathcal{Y}} \ell(y, \widehat{y}) \, dP(y \,|\, \boldsymbol{x}) \tag{6}$$
for each $\boldsymbol{x} \in \mathcal{X}$.
- The Bayes predictor (4) does not necessarily coincide with the pointwise Bayes predictor (6). This discrepancy between h^* and f^* is connected to the uncertainty regarding the right type of model to be fit, and hence the choice of the hypothesis space $\mathcal{H}$. We refer to this uncertainty as *model uncertainty*. Thus, due to this uncertainty, one can not guarantee that $h^*(\boldsymbol{x}) = f^*(\boldsymbol{x})$, or, in case the hypothesis h^* delivers probabilistic predictions $p(y \,|\, h^*, \boldsymbol{x})$ instead of point predictions, that $p(\cdot \,|\, h^*, \boldsymbol{x}) = p(\cdot \,|\, \boldsymbol{x})$.
- The hypothesis $\widehat{h}$ produced by the learning algorithm, for example the empirical risk minimizer (3), is only an estimate of h^*, and the quality of this estimate strongly depends on the quality and the amount of training data. We refer to the discrepancy between $\widehat{h}$ and h^*, i.e., the uncertainty about how well the former approximates the latter, as *approximation uncertainty*.

As already said, aleatoric uncertainty is typically understood as uncertainty that is due to influences on the data-generating process that are inherently random, that is, due to the non-deterministic nature of the sought input/output dependency. This part of the uncertainty is irreducible, in the sense that the learner cannot get rid of it. Model uncertainty and approximation uncertainty, on

the other hand, are subsumed under the notion of epistemic uncertainty, that is, uncertainty due to a lack of knowledge about the perfect predictor (6). Obviously, this lack of knowledge will strongly depend on the underlying hypothesis space $\mathcal{H}$ as well as the amount of data seen so far: The larger the number $N = |\mathcal{D}|$ of observations, the less ignorant the learner will be when having to make a new prediction. In the limit, when $N \to \infty$, a consistent learner will be able to identify h^*. Moreover, the "larger" the hypothesis pace $\mathcal{H}$, i.e., the weaker the prior knowledge about the sought dependency, the higher the epistemic uncertainty will be, and the more data will be needed to resolve this uncertainty.

How to capture these intuitive notions of aleatoric and epistemic uncertainty in terms of quantitative measures? In the following, we briefly recall two proposals that have recently been made in the literature.

2.1 Entropy Measures

An attempt at measuring and separating aleatoric and epistemic uncertainty on the basis of classical information-theoretic measures of entropy is made in [2]. This approach is developed in the context of neural networks for regression, but the idea as such is more general and can also be applied to other settings. A similar approach was recently adopted in [10].

More specifically, given a query instance $\boldsymbol{x}$, the idea is to measure the total uncertainty in a prediction in terms of the (Shannon) entropy of the predictive posterior distribution, which, in the case of discrete $\mathcal{Y}$, is given as

$$H\big[\,p(y \,|\, \boldsymbol{x})\,\big] = \mathbf{E}_{p(y \,|\, \boldsymbol{x})}\big\{ -\log_2 p(y \,|\, \boldsymbol{x})\big\} = -\sum_{y \in \mathcal{Y}} p(y \,|\, \boldsymbol{x}) \log_2 p(y \,|\, \boldsymbol{x})\,. \tag{7}$$

Moreover, the epistemic uncertainty is measured in terms of the mutual information between hypotheses and outcomes (i.e., the Kullback-Leibler divergence between the joint distribution of outcomes and hypotheses and the product of their marginals):

$$I(y, h) = \mathbf{E}_{p(y,h)} \left\{ \log_2 \left(\frac{p(y,h)}{p(y)p(h)} \right) \right\}, \tag{8}$$

Finally, the aleatoric uncertainty is specified in terms of the difference between (7) and (8), which is given by

$$\mathbf{E}_{p(h \,|\, \mathcal{D})} H\big[p(y \,|\, h, \boldsymbol{x})\big] = -\int_{\mathcal{H}} p(h \,|\, \mathcal{D}) \left(\sum_{y \in \mathcal{Y}} p(y \,|\, h, \boldsymbol{x}) \log_2 p(y \,|\, h, \boldsymbol{x}) \right) d\, h \tag{9}$$

The idea underlying (9) is as follows: By fixing a hypothesis $h \in \mathcal{H}$, the epistemic uncertainty is essentially removed. Thus, the entropy $H[p(y \,|\, h, \boldsymbol{x})]$, i.e., the entropy of the conditional distribution on $\mathcal{Y}$ predicted by h for the query instance $\boldsymbol{x}$, is a natural measure of the aleatoric uncertainty. However, since h is not precisely known, aleatoric uncertainty is measured in terms of the expectation of this entropy with regard to the posterior probability $p(h \,|\, \mathcal{D})$.

The epistemic uncertainty (8) captures the dependency between the probability distribution on $\mathcal{Y}$ and the hypothesis h. Roughly speaking, (8) is high if the distribution $p(y \,|\, h, \boldsymbol{x})$ varies a lot for different hypotheses h with high probability. This is plausible, because the existence of different hypotheses, all considered (more or less) probable but leading to quite different predictions, can indeed be seen as a sign for high epistemic uncertainty.

Obviously, (8) and (9) cannot be computed efficiently, because they involve an integration over the hypothesis space $\mathcal{H}$. One idea, therefore, is to approximate these measures by means of ensemble techniques [10], that is, to represent the posterior distribution $p(h \mid \mathcal{D})$ by a finite ensemble of hypotheses $H = \{h_1, \ldots, h_M\}$. An approximation of (9) can then be obtained by

$$u_a(\boldsymbol{x}) := -\frac{1}{M} \sum_{i=1}^{M} \sum_{y \in \mathcal{Y}} p(y \mid h_i, \boldsymbol{x}) \log_2 p(y \mid h_i, \boldsymbol{x}) \,, \tag{10}$$

an approximation of (7) by

$$u_t(\boldsymbol{x}) := -\sum_{y \in \mathcal{Y}} \left(\frac{1}{M} \sum_{i=1}^{M} p(y \mid h_i, \boldsymbol{x}) \right) \log_2 \left(\frac{1}{M} \sum_{i=1}^{M} p(y \mid h_i, \boldsymbol{x}) \right) \,, \tag{11}$$

and finally and approximation of (8) by $u_e(\boldsymbol{x}) := u_t(\boldsymbol{x}) - u_a(\boldsymbol{x})$. For neural networks, it has been shown that techniques such as Dropout [3] and DropConnect [10] can be interpreted as (implicit) ensemble methods, and can hence be used to implement this approach.

2.2 Measures based on Relative Likelihood

Another approach, put forward in [18], is based on the use of relative likelihoods, historically proposed by [1] and then justified in other settings such as possibility theory [21]. Here, we briefly recall this approach for the case of binary classification, i.e., where $\mathcal{Y} = \{0, 1\}$; see [13] for an extension to the case of multinomial classification.

Given training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$, the normalized likelihood of $h \in \mathcal{H}$ is defined as

$$\pi_{\mathcal{H}}(h) := \frac{L(h)}{L(h^{ml})} = \frac{L(h)}{\max_{h' \in \mathcal{H}} L(h')} \,, \tag{12}$$

where $L(h) = \prod_{i=1}^{N} p(y_i \mid h, \boldsymbol{x}_i)$ is the likelihood of h, and $h^{ml} \in \mathcal{H}$ the maximum likelihood estimation. For a given instance $\boldsymbol{x}$, the degrees of support (plausibility) of the two classes are defined as follows:

$$\pi(1 \mid \boldsymbol{x}) = \sup_{h \in \mathcal{H}} \min \left[\pi_{\mathcal{H}}(h), p(1 \mid h, \boldsymbol{x}) - p(0 \mid h, \boldsymbol{x}) \right], \tag{13}$$

$$\pi(0 \mid \boldsymbol{x}) = \sup_{h \in \mathcal{H}} \min \left[\pi_{\mathcal{H}}(h), p(0 \mid h, \boldsymbol{x}) - p(1 \mid h, \boldsymbol{x}) \right]. \tag{14}$$

So, $\pi(1 \mid \boldsymbol{x})$ is high if and only if a highly plausible hypothesis supports the positive class much stronger (in terms of the assigned probability) than the negative class (and $\pi(0 \mid \boldsymbol{x})$ can be interpreted analogously). Given the above degrees of support, the degrees of epistemic and aleatoric uncertainty are defined as follows:

$$u_e(\boldsymbol{x}) = \min \left[\pi(1 \mid \boldsymbol{x}), \pi(0 \mid \boldsymbol{x}) \right] , \tag{15}$$

$$u_a(\boldsymbol{x}) = 1 - \max \left[\pi(1 \mid \boldsymbol{x}), \pi(0 \mid \boldsymbol{x}) \right] . \tag{16}$$

Thus, epistemic uncertainty refers to the case where both the positive and the negative class appear to be plausible, while the degree of aleatoric uncertainty (16) is the degree to which none of the classes is supported. More specifically, the above measures have the following properties:

- $u_e(\boldsymbol{x})$ will be high if class probabilities strongly vary within the set of plausible hypotheses, i.e., if we are unsure how to compare these probabilities. In particular, it will be 1 if and only if we have $h(\boldsymbol{x}) = 1$ and $h'(\boldsymbol{x}) = 0$ for two totally plausible hypotheses h and h';

- $u_a(\boldsymbol{x})$ will be high if class probabilities are similar for all plausible hypotheses, i.e., if there is strong evidence that $h(\boldsymbol{x}) \approx 0.5$. In particular, it will be close to 1 if all plausible hypotheses allocate their probability mass around $h(\boldsymbol{x}) = 0.5$.

As can be seen, the measures (15) and (16) are actually quite similar in spirit to the measures (8) and (9).

3 Random Forests

Our basic idea is to instantiate the (generic) uncertainty measures presented in the previous section by means of decision trees [15,16], that is, with decision trees as an underlying hypothesis space $\mathcal{H}$. This idea is motivated by the fact that, firstly, decision trees can naturally be seen as probabilistic predictors [7], and secondly, they can easily be used as an ensemble in the form of a random forest — recall that ensembling is needed for the (approximate) computation of the entropy-based measures in Section 2.1.

3.1 Entropy Measures

The approach in Section 2.1 can be realized with decision forests in a quite straightforward way. Let $H = \{h_1, \ldots, h_M\}$ be a classifier ensemble in the form of a random forest consisting of decision trees h_i. Moreover, recall that a decision tree h_i partitions the instance space $\mathcal{X}$ into (rectangular) regions $R_{i,1}, \ldots, R_{i,L_i}$ (i.e., $\bigcup_{l=1}^{L_i} R_{i,l} = \mathcal{X}$ and $R_{i,k} \cap R_{i,l} = \emptyset$ for $k \neq l$) associated with corresponding leafs of the tree (each leaf node defines a region R). Given a query instance $\boldsymbol{x}$, the probabilistic prediction produced by the tree h_i is specified by the Laplace-corrected relative frequencies of the classes $y \in \mathcal{Y}$ in the region $R_{i,j} \ni \boldsymbol{x}$:

$$p(y \,|\, h_i, \boldsymbol{x}) = \frac{n_{i,j}(y) + 1}{n_{i,j} + |\mathcal{Y}|} \,,$$

where $n_{i,j}$ is the number of training instances in the leaf node $R_{i,j}$, and $n_{i,j}(y)$ the number of instances with class y. With probabilities estimated in this way, the uncertainty degrees (10) and (11) can directly be derived.

3.2 Measures based on Relative Likelihood

Instantiating the approach in Section 2.2 essentially means computing the degrees of support (13–14), from which everything else can easily be derived.

As already said, a decision tree partitions the instance space into several regions, each of which can be associated with a constant predictor. More specifically, in the case of binary classification, the predictor is of the form h_θ, $\theta \in \Theta = [0, 1]$, where $h_\theta(\boldsymbol{x}) \equiv \theta$ is the (predicted) probability $p(1 \,|\, \boldsymbol{x} \in R)$ of the positive class in the region. If we restrict inference to a local region, the underlying hypothesis space is hence given by $\mathcal{H} = \{h_\theta \,|\, 0 \leq \theta \leq 1\}$.

With n and p the number of positive and negative instances, respectively, within a region R, the likelihood and the maximum likelihood estimate of θ are respectively given by

$$L(\theta) = \binom{n+p}{n} \theta^n (1-\theta)^p \text{ and } \theta^{ml} = \frac{n}{n+p} \,. \tag{17}$$

Therefore, the degrees of support for the positive and negative classes are

$$\pi(1 \mid \boldsymbol{x}) = \sup_{\theta \in [0,1]} \min \left(\frac{\theta^p (1-\theta)^n}{\left(\frac{p}{n+p}\right)^p \left(\frac{n}{n+p}\right)^n}, \, 2\theta - 1 \right) , \tag{18}$$

$$\pi(0 \mid \boldsymbol{x}) = \sup_{\theta \in [0,1]} \min \left(\frac{\theta^p (1-\theta)^n}{\left(\frac{p}{n+p}\right)^p \left(\frac{n}{n+p}\right)^n}, \, 1 - 2\theta \right) . \tag{19}$$

Solving (18) and (19) comes down to maximizing a scalar function over a bounded domain, for which standard solvers can be used. From (18–19), the epistemic and aleatoric uncertainty associated with the region R can be derived according to (15) and (16), respectively. For different combinations of n and p, these uncertainty degrees can be pre-computed.

Note that, for this approach, the uncertainty degrees (15) and (16) can be obtained for a single tree. To leverage the ensemble H, we average both uncertainties over all trees in the random forest.

4 Experiments

The empirical evaluation of methods for quantifying uncertainty is a non-trivial problem. In fact, unlike for the prediction of a target variable, the data does normally not contain information about any sort of "ground truth" uncertainty. What is often done, therefore, is to evaluate predicted uncertainties *indirectly*, that is, by assessing their usefulness for improved prediction and decision making. Adopting an approach of that kind, we produced *accuracy-rejection curves*, which depict the accuracy of a predictor as a function of the percentage of rejections [5]: A classifier, which is allowed to abstain on a certain percentage p of predictions, will predict on those $(1-p)\,\%$ on which it feels most certain. Being able to quantify its own uncertainty well, it should improve its accuracy with increasing p, hence the accuracy-rejection curve should be monotone increasing (unlike a flat curve obtained for random abstention).

4.1 Implementation Details

For this work, we used the Random Forest Classifier from SKlearn. The number of trees within the forest is set to 50, with the maximum level of tree grows set to 10. We use bootstrapping to create diversity between the trees of the forest.

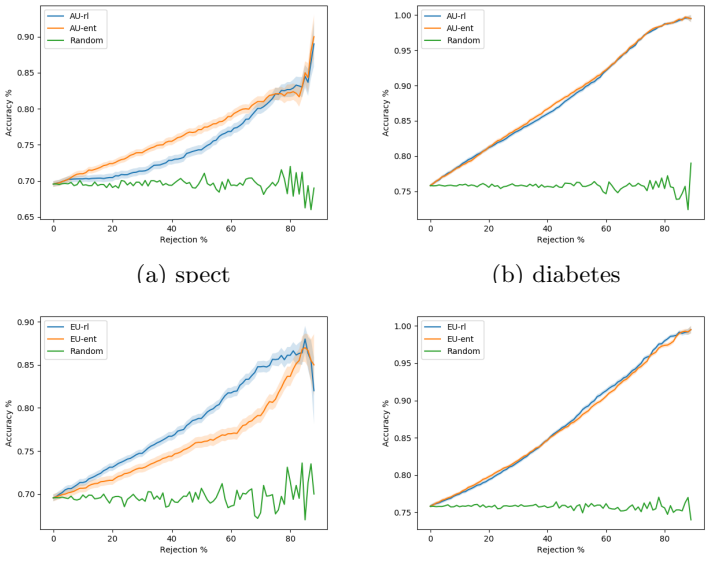
As a baseline to compare with, we used the DropConnect model for deep neural networks as introduced in [10]. The idea of DropConnect is similar to Dropout, but here, instead of randomly deleting neurons, we randomly delete the connections between neurons. In this model, the act of dropping the connections is also active in the test phase. In this way, the data passes through a different network on each iteration, and therefore we can compute Monte Carlo samples for each query instance. The DropConnect model is a feed forward neural network consisting of two DropConnect layers with 32 neurons and a final softmax layer for the output. The model is trained for 20 epochs with mini batch size of 32. After the training is done, we take 50 Monte Carlo samples to create an ensemble, from which the uncertainty values can be calculated.

4.2 Results

Due to space limitations, we show results in the form of accuracy-rejection curves for only two exemplary data sets from the UCI repository[1], spect and diabetes — yet, very similar results were

[1] `https://archive.ics.uci.edu/ml/datasets/`

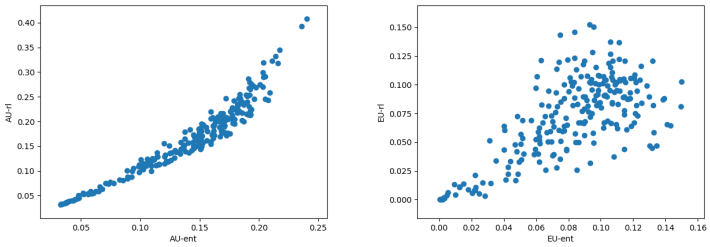
obtained for other data sets. The data is randomly split into 70% for training and 30% for testing, and accuracy-rejection curves are computed on the latter (the curves shown are averages over 100 repetitions). In the following, we abbreviate the aleatoric and epistemic uncertainty degrees produced by the entropy-based approach (Section 2.1) and the approach based on relative likelihood (Section 2.2) by AU-ent, EU-ent, AU-rl, and EU-rl, respectively.



(a) spect (b) diabetes

(c) spect (d) diabetes

Fig. 2: Accuracy-rejection curves for aleatoric (above) and epistemic (below) uncertainty using random forests. The curve for random rejection is included as a baseline.

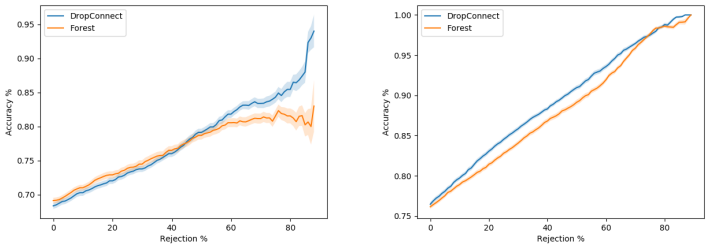


(a) spect (b) diabetes

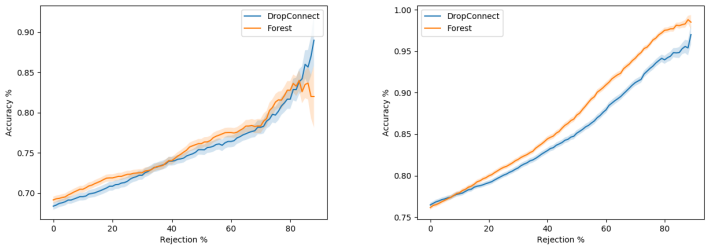
Fig. 3: Scatter plot for test set on diabetes data, showing the relationship between the uncertainty degrees (aleatoric left, epistemic right) estimated by the two approaches.

As can be seen from Figures 1–4, both approaches to measuring uncertainty are effective in the sense of producing monotone increasing accuracy-rejection curves, and on the data sets we analyzed so far, we could not detect any systematic differences in performance. Besides, rejection seems to work well on the basis of both criteria, aleatoric as well as epistemic uncertainty. This is plausible,

since both provide reasonable reasons for a learner to abstain from a prediction. Likewise, there are no big differences between random forests and neural networks, showing that the former are indeed a viable alternative to the latter — this was actually a major concern of our study.



(a) spect (b) diabetes



(c) spect (d) diabetes

Fig. 4: Comparison between random forests and neural networks (DropConnect) for aleatoric (above) and epistemic (below) uncertainty.

5 Conclusion

The distinction between aleatoric and epistemic uncertainty has recently received a lot of attention in machine learning, especially in the deep learning community [6]. Roughly speaking, the approaches in deep learning are either based on the idea of equipping networks with a probabilistic component, like in Bayesian deep learning [11], or on using ensemble techniques [8], which can be implemented (indirectly) through techniques such as Dropout [3] or DropConnect. The main purpose of this paper was to show that the use of decision trees and random forests is an interesting alternative to neural networks.

Indeed, as we have shown, the basic ideas underlying the estimation of aleatoric and epistemic uncertainty can be realized with random forests in a very natural way. In a sense, they even appear to be simpler and more flexible than neural networks. For example, while the approach based on relative likelihood (Section 2.2) could be realized efficiently for random forests, a neural network implementation is far from obvious (and was therefore not included in the experiments).

There are various directions for future work. For example, since the hyper-parameters of random forests have an influence on the hypothesis space we are (indirectly) working with, they also influence the estimation of uncertainty degrees. This relationship calls for a thorough investigation. Besides, going beyond a proof of principle with statistics such as accuracy-rejection curves, it would be interesting to make use of uncertainty quantification with random forests in applications such as active learning, as recently proposed in [12].

References

1. Birnbaum, A.: On the foundations of statistical inference. Journal of the American Statistical Association **57**(298), 269–306 (1962)
2. Depeweg, S., Hernandez-Lobato, J., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: Proc. ICML, 35th Int. Conf. on Machine Learning. Stockholm, Sweden (2018)
3. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with Bernoulli approximate variational inference. In: Proc. of the ICLR Workshop Track (2016)
4. Hora, S.: Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliability Engineering and System Safety **54**(2–3), 217–223 (1996)
5. Hühn, J., Hüllermeier, E.: FR3: A fuzzy rule learner for inducing reliable classifiers. IEEE Transactions on Fuzzy Systems **17**(1), 138–149 (2009)
6. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Proc. NIPS, pp. 5574–5584 (2017)
7. Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I., Malley, J., Ziegler, A.: Probability estimation with machine learning methods for dichotomous and multi-category outcome: Theory. Biometrical Journal **56**(4), 534–563 (2014)
8. Lakshminarayanan, B., Pritzel, A., C. Blundell: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proc. NeurIPS, 31st Conference on Neural Information Processing Systems. Long Beach, California, USA (2017)
9. Lambrou, A., Papadopoulos, H., Gammerman, A.: Reliable confidence measures for medical diagnosis with evolutionary algorithms. IEEE Trans. on Information Technology in Biomedicine **15**(1), 93–99 (2011)
10. Mobiny, A., Nguyen, H., Moulik, S., Garg, N., Wu, C.: DropConnect is effective in modeling uncertainty of Bayesian networks. CoRR **abs/1906.04569** (2017), `http://arxiv.org/abs/1906.04569`
11. Neal, R.: Bayesian learning for neural networks. Springer Science & Business Media **118** (2012)
12. Nguyen, V., Destercke, S., Hüllermeier, E.: Epistemic uncertainty sampling. In: Proc. DS 2019, 22nd Int. Conf. on Discovery Science. Split, Croatia (2019)
13. Nguyen, V.L., Destercke, S., Masson, M.H., Hüllermeier, E.: Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty. In: Proc. IJCAI, pp. 5089–5095. AAAI Press (2018)
14. Papernot, N., McDaniel, P.: Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. CoRR **abs/1803.04765v1** (2018), `http://arxiv.org/abs/1803.04765`
15. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
16. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics **21**(3), 660–674 (1991)
17. Sato, M., Suzuki, J., Shindo, H., Matsumoto, Y.: Interpretable adversarial perturbation in input embedding space for text. In: Proceedings IJCAI 2018, pp. 4323–4330. Stockholm, Sweden (2018)
18. Senge, R., Bösner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., Hüllermeier, E.: Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences **255**, 16–29 (2014)
19. Varshney, K.: Engineering safety in machine learning. In: Proc. Inf. Theory Appl. Workshop. La Jolla, CA (2016)
20. Varshney, K., Alemzadeh, H.: On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. CoRR **abs/1610.01256** (2016), `http://arxiv.org/abs/1610.01256`
21. Walley, P., Moral, S.: Upper probabilities based only on the likelihood function. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(4), 831–847 (1999)
22. Yang, F., Wanga, H.Z., Mi, H., de Lin, C., Cai, W.W.: Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics **10** (2009)