Machine Learning (2022) 111:89–122
https://doi.org/10.1007/s10994-021-06003-9

How to measure uncertainty in uncertainty sampling for active learning

Vu-Linh Nguyen[1] · **Mohammad Hossein Shaker**[2] · **Eyke Hüllermeier**[2]

Received: 29 February 2020 / Revised: 17 February 2021 / Accepted: 14 May 2021 /
Published online: 18 June 2021

Abstract
Various strategies for active learning have been proposed in the machine learning literature. In uncertainty sampling, which is among the most popular approaches, the active learner sequentially queries the label of those instances for which its current prediction is maximally uncertain. The predictions as well as the measures used to quantify the degree of uncertainty, such as entropy, are traditionally of a probabilistic nature. Yet, alternative approaches to capturing uncertainty in machine learning, alongside with corresponding uncertainty measures, have been proposed in recent years. In particular, some of these measures seek to distinguish different sources and to separate different types of uncertainty, such as the reducible (epistemic) and the irreducible (aleatoric) part of the total uncertainty in a prediction. The goal of this paper is to elaborate on the usefulness of such measures for uncertainty sampling, and to compare their performance in active learning. To this end, we instantiate uncertainty sampling with different measures, analyze the properties of the sampling strategies thus obtained, and compare them in an experimental study.

Keywords Active learning · Uncertainty sampling · Credal uncertainty · Epistemic uncertainty · Aleatoric uncertainty

Editors: Petra Kralj Novak, Tomislav Šmuc.

✉ Eyke Hüllermeier
eyke@upb.de

Vu-Linh Nguyen
v.l.nguyen@tue.nl

Mohammad Hossein Shaker
mhshaker@mail.upb.de

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[2] Heinz Nixdorf Institute and Department of Computer Science, Paderborn University, Paderborn, Germany

1 Introduction

The goal in standard supervised learning, such as binary or multi-class classification, is to learn models with high predictive accuracy from labelled training data (Hastie et al. 2005; Vapnik 1999). However, labelled data does not come for free. On the contrary, labelling can be expensive, time-consuming, and costly. The ambition of *active learning*, therefore, is to exploit labelled data in the most effective way. More specifically, the idea is to let the learning algorithm itself decide which examples it considers to be most informative. Compared to random sampling, the hope is to achieve better performance with the same amount of training data, or to reach the same performance with less data (Fu et al. 2013; Settles 2009).

The selection of training examples is often done in an iterative manner, i.e., the active learner alternates between re-training and selecting new examples. In each iteration, the usefulness of a candidate example is estimated in terms of a *utility score*, and the one with the highest score is queried. In this regard, the notion of utility typically refers to uncertainty reduction: To what extent will the knowledge about the label of a specific instance help to reduce the learner's uncertainty about the sough model? In *uncertainty sampling* (Settles 2009), which is among the most popular approaches, utility is quantified in terms of predictive uncertainty, i.e., the active learner selects those instances for which its current prediction is maximally uncertain. The predictions as well as the measures used to quantify the degree of uncertainty, such as entropy, are almost exclusively of a probabilistic nature. Such approaches indeed proved to be successful in many applications.

Yet, as pointed out by Sharma and Bilgic (2017), existing approaches can be criticized for not informing about the *reasons* for why an instance is considered uncertain, although this might be relevant for judging the potential usefulness of an example. They propose an evidence-based approach to active learning, in which *conflicting-evidence* uncertainty is distinguished from *insufficient-evidence* uncertainty. A similar distinction between two types of uncertainty, called *epistemic* and *aleatoric* uncertainty, has been made in the recent machine learning literature (Hüllermeier and Waegeman 2021; Kendall and Gal 2017; Senge et al. 2014). Roughly speaking, aleatoric uncertainty is due to inherent randomness, whereas epistemic uncertainty captures the lack of knowledge of the learner. Thus, the latter corresponds to the reducible and the former to the irreducible part of the total uncertainty in a prediction. Last but not least, measures of uncertainty are also discussed in connection with generalizations of standard probability theory, most notably imprecise probability (De Campos et al. 1994; Zaffalon 2002). Here, incomplete information is captured in the form of *credal sets*, that is, (convex) sets of probability distributions; correspondingly, standard uncertainty measures for single probability distributions (such as entropy) are generalized toward credal uncertainty measures.

This paper is an extension of (Nguyen et al. 2019), in which the authors conjecture that, in uncertainty sampling, the usefulness of an instance is better reflected by its epistemic than by its aleatoric uncertainty, and provide first evidence in favor of this conjecture. The goal of the current paper is to elaborate more broadly on the usefulness of different measures for uncertainty sampling, and to compare their performance in active learning. To this end, we instantiate uncertainty sampling with different measures, analyze the properties of the sampling strategies thus obtained, and compare them in an experimental study.

The rest of this paper is organized as follows. In the next section, we first recall the general framework of uncertainty sampling and provide a brief survey of related work on active learning. We present different approaches for measuring the learner's uncertainty

in a query instance and a comparison of the approaches in Sects. 3 and 4, respectively. Experimental evaluations for local learning (Parzen window classifier), decision trees and logistic regression are presented in Sect. 5, prior to concluding the paper in Sect. 6. Technical details for instantiations of aleatoric and epistemic uncertainty are deferred to the appendices.

2 Uncertainty sampling

In this section, we briefly recall the basic setting of uncertainty sampling. As usual in active learning, we assume to be given a labelled set of training data $\mathbf{D}$ and a pool of unlabeled instances $\mathbf{U}$ that can be queried by the learner:

$$\mathbf{D} = \left\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\right\}, \quad \mathbf{U} = \left\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J\right\}.$$

Instances are represented as features vectors $\boldsymbol{x}_i = \left(x_i^1, \ldots, x_i^d\right) \in \mathcal{X} = \mathbb{R}^d$. In this paper, we only consider the case of binary classification, where labels y_i are taken from $\mathcal{Y} = \{0, 1\}$, leaving the more general case of multi-class classification for future work. We denote by $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ the underlying hypothesis space, i.e., the class of candidate models $h : \mathcal{X} \longrightarrow \mathcal{Y}$ the learner can choose from. Often, hypotheses are parametrized by a parameter vector $\theta \in \Theta$; in this case, we equate a hypothesis $h = h_\theta \in \mathcal{H}$ with the parameter θ, and the model space $\mathcal{H}$ with the parameter space Θ.

In uncertainty sampling, instances are queried in a greedy fashion. Given the current model θ that has been trained on $\mathbf{D}$, each instance $\boldsymbol{x}_j$ in the current pool $\mathbf{U}$ is assigned a *utility* score $s(\theta, \boldsymbol{x}_j)$, and the next instance to be queried is the one with the highest score (Lewis and Gale 1994; Settles 2009; Settles and Craven 2008; Sharma and Bilgic 2017). The chosen instance is labelled (by an oracle or expert) and added to the training data $\mathbf{D}$, on which the model is then re-trained. The active learning process for a given budget B (i.e, the number of unlabelled instances to be queried) is summarized in Algorithm 1.

Algorithm 1: Uncertainty sampling

Input: $\mathbf{U}$, $\mathbf{D}$, θ- initial pool, training data, classifier, and B-budget
Output: $\mathbf{U}$, $\mathbf{D}$, θ - updated pool, training data, classifier

initialize $b = 0$;
while $b < B$ **do**
  foreach $\boldsymbol{x} \in \mathbf{U}$ **do**
    compute $s(\theta, \boldsymbol{x})$
  query the label of the optimal instance $\boldsymbol{x}^*$ with respect to $s(\theta, \boldsymbol{x})$
  $\mathbf{D} = \mathbf{D} \cup \{\boldsymbol{x}^*, y^*\}$;
  $\mathbf{U} = \mathbf{U} \setminus \{\boldsymbol{x}^*, y^*\}$;
  train θ from $\mathbf{D}$;
  $b = b + 1$;
Return $\mathbf{U}$, $\mathbf{D}$, θ;

Assuming a probabilistic model producing predictions in the form of probability distributions $p_\theta(\cdot \mid \boldsymbol{x})$ on $\mathcal{Y}$, the utility score is typically defined in terms of a measure of uncertainty. Thus, instances on which the current model is highly uncertain are supposed to be maximally informative (Settles 2009; Settles and Craven 2008; Sharma and Bilgic 2017). Popular examples of such measures include

– the entropy:

$$s(\theta, \boldsymbol{x}) = -\sum_{y \in \mathcal{Y}} p_\theta(y \mid \boldsymbol{x}) \log p_\theta(y \mid \boldsymbol{x}), \tag{1}$$

– the least confidence:

$$s(\theta, \boldsymbol{x}) = 1 - \max_{y \in \mathcal{Y}} p_\theta(y \mid \boldsymbol{x}), \tag{2}$$

– the smallest margin:

$$s(\theta, \boldsymbol{x}) = p_\theta(y_m \mid \boldsymbol{x}) - p_\theta(y_n \mid \boldsymbol{x}), \tag{3}$$

where $y_m = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid \boldsymbol{x})$ and $y_n = \arg\max_{y \in \mathcal{Y} \setminus y_m} p_\theta(y \mid \boldsymbol{x})$.

While the first two measures ought to be maximized, the last one has to be minimized. In the case of binary classification, i.e, $\mathcal{Y} = \{0, 1\}$, all these measures rank unlabelled instances in the same order and look for instances with small difference between $p_\theta(0 \mid \boldsymbol{x})$ and $p_\theta(1 \mid \boldsymbol{x})$.

3 Measures of uncertainty

In this section, we present different frameworks for measuring the learner's uncertainty in a query instance: evidence-based uncertainty (EBU), credal uncertainty (CU), and an approach focusing on a distinction between epistemic and aleatoric uncertainty (EAU). While the first one has been specifically developed for the purpose of active learning, the other two are more general approaches to uncertainty quantification in machine learning. Yet, their potential usefulness for active learning has been pointed out as well (Antonucci et al. 2012; Nguyen et al. 2019).

3.1 Evidence-based uncertainty (EBU)

In their evidence-based uncertainty sampling approach, Sharma and Bilgic (2013, 2017) propose to differentiate between uncertainty due to conflicting evidence and insufficient evidence. The corresponding measures of *conflicting-evidence uncertainty* and *insufficient-evidence uncertainty* are mainly motivated for the Naïve Bayes (NB) classifier as a learning algorithm. In the spirit of this classifier, evidence-based uncertainty sampling first looks at the influence of individual features x^m in the feature representation $\boldsymbol{x} = (x^1, \ldots, x^d)$ of instances. More specifically, given the current model θ, denote by $p_\theta(x^m \mid 0)$ and $p_\theta(x^m \mid 1)$ the class-conditional probabilities on the values of the m^{th} feature. For a given instance $\boldsymbol{x}$, the authors partition the set of features into those that provide evidence for the positive and for the negative class, respectively:

$$P_\theta(\boldsymbol{x}) = \left\{ x^m \,\middle|\, \frac{p_\theta(x^m \mid 1)}{p_\theta(x^m \mid 0)} > 1 \right\}, \tag{4}$$

$$N_\theta(\boldsymbol{x}) = \left\{ x^m \,\middle|\, \frac{p_\theta(x^m \mid 0)}{p_\theta(x^m \mid 1)} > 1 \right\}. \tag{5}$$

Then, the total evidence for the positive and the negative class is determined as follows:

$$E_1(\boldsymbol{x}) = \prod_{x^m \in P_\theta(\boldsymbol{x})} \frac{p_\theta(x^m \mid 1)}{p_\theta(x^m \mid 0)}\,, \tag{6}$$

$$E_0(\boldsymbol{x}) = \prod_{x^m \in N_\theta(\boldsymbol{x})} \frac{p_\theta(x^m \mid 0)}{p_\theta(x^m \mid 1)}\,. \tag{7}$$

The authors consider a situation as *conflicting* evidence if both $E_0(\boldsymbol{x})$ and $E_1(\boldsymbol{x})$ are high, because in such as situation, there is strong evidence in favor of the positive as well as strong evidence in favor of the negative class. Likewise, a situation in which both evidences are low is considered as *insufficient* evidence. Measuring these conditions in terms of the product[1] $E_1(\boldsymbol{x}) \times E_0(\boldsymbol{x})$, the conflicting evidence-based approach simply queries the instance with the highest conflicting evidence, while the insufficient evidence-based approach looks for the one with the highest insufficient evidence:

$$s^*_{conf} = \arg\max_{\boldsymbol{x} \in \mathbf{S}} E_1(\boldsymbol{x}) \times E_0(\boldsymbol{x})\,, \tag{8}$$

$$s^*_{insu} = \arg\min_{\boldsymbol{x} \in \mathbf{S}} E_1(\boldsymbol{x}) \times E_0(\boldsymbol{x})\,. \tag{9}$$

Note that the selection is restricted to the set $\mathbf{S}$ of instances $\boldsymbol{x}$ in the pool $\mathbf{U}$ having the highest scores $s(\theta, \boldsymbol{x})$ according to standard uncertainty sampling; the size of this set, $t = |\mathbf{S}|$, is a parameter of the method (and hence a hyper-parameter for the active learning algorithm). The restriction to the most uncertain cases puts evidence-based uncertainty sampling close to standard uncertainty sampling. Instead of using conflicting-evidence and insufficient-evidence uncertainties as selection criteria on their own, they are merely used for prioritizing cases that appear to be uncertain in the traditional sense.

3.1.1 A note on evidence-based uncertainty

Interestingly, to motivate their approach, Sharma and Bilgic (2017) note that "regardless of whether we want to maximize or minimize $E_1(\boldsymbol{x}) \times E_0(\boldsymbol{x})$, we want to guarantee that the underlying model is uncertain about the chosen instance", thereby suggesting that the evidence-based uncertainties alone do not necessarily inform about this uncertainty. Indeed, it is true that these uncertainties are not easy to interpret (see also our discussion in Sect. 4.1), and that their relationship to standard uncertainty measures is not fully obvious.

In particular, note that the latter also comprises the influence of the prior class probabilities, which is completely neglected by the evidence-based uncertainties (which only look at the likelihood). This is especially relevant in the case of imbalanced class distributions. In such cases, evidence-based uncertainty may strongly deviate from standard uncertainty, i.e., the entropy of the posterior distribution. For instance, $E_0(\boldsymbol{x})$ and $E_1(\boldsymbol{x})$ could both be very large, and $p_\theta(\boldsymbol{x} \mid 0) \approx p_\theta(\boldsymbol{x} \mid 1)$, although $p_\theta(0 \mid \boldsymbol{x})$ is very different from $p_\theta(1 \mid \boldsymbol{x})$ due to unequal prior odds, and hence the entropy small. Likewise, the entropy of the posterior can be large although both evidence-based uncertainties are small.

[1] This is the measure used in (Sharma and Bilgic 2013, 2017). The authors mention, however, that other measures could be used as well.

3.1.2 A note on uncertainty sampling for Naïve Bayes

The evidence-based approach to uncertainty sampling has been introduced with a focus on Naïve Bayes as a base learner. In this regard, we like to note that uncertainty sampling for this learner might be considered critical in general.

It is clear that active learning may always incorporate a bias, simply because the data is no longer produced by sampling independently according to the true underlying distribution. Thus, the data is no longer completely representative. While this may affect any learning algorithm, the effect appears to be especially strong for NB, so that uncertainty sampling for NB appears to be questionable in general. In fact, a sample bias has a very direct influence on the probabilities estimated by NB. In particular, the estimated class priors are strongly biased toward the conditional class probabilities of those instances with a high uncertainty, because these are sampled more often. This bias may in turn affect the classifier as a whole, and lead to suboptimal predictions.

As an illustration of the problem, let us consider a small example with only two binary attributes x^1 and x^2. This example may appear unrealistic, because the instance space is finite and actually quite small. Please note, however, that even in practice NB is typically applied to discrete attributes with finite domains (possibly after a discretization of numerical attributes in a pre-processing step).

Suppose the class priors to be given by $p(y = 0) = 0.3$ and $p(y = 1) = 0.7$, and the class-conditional probabilities as follows:

$$\begin{aligned} p(x^1 = 1 \,|\, y = 0) &= 0.4\,, \\ p(x^1 = 1 \,|\, y = 1) &= 0.2\,, \\ p(x^2 = 1 \,|\, y = 0) &= 0.8\,, \\ p(x^2 = 1 \,|\, y = 1) &= 0.4\,. \end{aligned}$$

From these, one derives the following posterior probabilities:

$$\begin{aligned} p(y = 0 \,|\, x^1 = 0, x^2 = 0) &\approx 0.10\,, \quad p(y = 1 \,|\, x^1 = 0, x^2 = 0) \approx 0.90\,, \\ p(y = 0 \,|\, x^1 = 0, x^2 = 1) &\approx 0.40\,, \quad p(y = 1 \,|\, x^1 = 0, x^2 = 1) \approx 0.60\,, \\ p(y = 0 \,|\, x^1 = 1, x^2 = 0) &\approx 0.22\,, \quad p(y = 1 \,|\, x^1 = 1, x^2 = 0) \approx 0.78\,, \\ p(y = 0 \,|\, x^1 = 1, x^2 = 1) &\approx 0.63\,, \quad p(y = 1 \,|\, x^1 = 1, x^2 = 1) \approx 0.37\,. \end{aligned}$$

Now, consider an active learner that can sample from a large (in principle infinite) pool of unlabeled data points (i.e., multiple copies of each of the four instances). Since the second instance $(x^1, x^2) = (0, 1)$ has the highest entropy, standard uncertainty sampling will sooner or later focus on this instance and sample it over and over again[2]. This of course has an influence on the estimation of priors and conditional probabilities by NB. In particular, the estimated class priors $\hat{p}(y = 0)$ and $\hat{p}(y = 1)$ will converge to the conditional posteriors, i.e., the posteriors of y given $(x^1, x^2) = (0, 1)$. Consequently, we will produce a bias in the estimates, and will obtain

[2] We confirmed this in a simulation study.

$$\hat{p}(y=0 \mid x^1=0, x^2=0) \approx 0.19\,, \quad \hat{p}(y=1 \mid x^1=0, x^2=0) \approx 0.81\,,$$
$$\hat{p}(y=0 \mid x^1=0, x^2=1) \approx 0.38\,, \quad \hat{p}(y=1 \mid x^1=0, x^2=1) \approx 0.62\,,$$
$$\hat{p}(y=0 \mid x^1=1, x^2=0) \approx 0.24\,, \quad \hat{p}(y=1 \mid x^1=1, x^2=0) \approx 0.76\,,$$
$$\hat{p}(y=0 \mid x^1=1, x^2=1) \approx 0.45\,, \quad \hat{p}(y=1 \mid x^1=1, x^2=1) \approx 0.56\,.$$

As one can see, this will even have an effect on the Bayes-optimal predictor: For $(x^1, x^2) = (1, 1)$, the prediction will be $\hat{y} = 1$ instead of the actually optimal prediction $\hat{y} = 0$. Similar effects can be found for the evidence-based approach. For example, when applying the insufficient evidence approach, it can happen that the active learner will completely focus on the third instance, which has the highest insufficient evidence, and then produce the following estimates:

$$\hat{p}(y=0 \mid x^1=0, x^2=0) \approx 0.14\,, \quad \hat{p}(y=1 \mid x^1=0, x^2=0) \approx 0.86\,,$$
$$\hat{p}(y=0 \mid x^1=0, x^2=1) \approx 0.36\,, \quad \hat{p}(y=1 \mid x^1=0, x^2=1) \approx 0.64\,,$$
$$\hat{p}(y=0 \mid x^1=1, x^2=0) \approx 0.23\,, \quad \hat{p}(y=1 \mid x^1=1, x^2=0) \approx 0.78\,,$$
$$\hat{p}(y=0 \mid x^1=1, x^2=1) \approx 0.49\,, \quad \hat{p}(y=1 \mid x^1=1, x^2=1) \approx 0.51\,.$$

So again, the prediction for $(x^1, x^2) = (1, 1)$ will be $\hat{y} = 1$ instead of $\hat{y} = 0$.

3.1.3 Extension to other learners

As explained above, the approach by Sharma and Bilgic (2017) is specifically tailored for Naïve Bayes as a learning algorithm. Yet, the authors also propose variants of their measures for logistic regression and support vector machines. For example, if the decision boundary obtained by fitting a logistic regression model is given by $h_\theta(\boldsymbol{x}) = \theta_0 + \sum_{i=1}^{d} \theta_i \cdot x^i$, the evidences for the positive and the negative class are defined, respectively, as follows (Sharma and Bilgic 2017):

$$E_1(\boldsymbol{x}) = \sum_{x^m \in P_\theta(\boldsymbol{x})} \theta_m \cdot x^m, \quad E_0(\boldsymbol{x}) = -\sum_{x^m \in N_\theta(\boldsymbol{x})} \theta_m \cdot x^m\,, \tag{10}$$

where

$$P_\theta(\boldsymbol{x}) = \left\{ x^m \,\middle|\, \theta_m \cdot x^m > 0 \right\}, \quad N_\theta(\boldsymbol{x}) = \left\{ x^m \,\middle|\, \theta_m \cdot x^m < 0 \right\}. \tag{11}$$

Obviously, evidence-based uncertainty measures can be derived is a quite natural way for models in which the features *independently* contribute to the prediction. However, the approach becomes much less straightforward in the case where features may interact with each other. In any case, new measures need to be derived for every model class separately. The approaches to be discussed next are more generic (and hence more principled) in the sense of being independent of the model class. That is, concrete measures of uncertainty can be derived for any model class in a generic way.

3.2 Credal uncertainty (CU)

Consider an instance space $\mathcal{X}$, output space $\mathcal{Y} = \{0, 1\}$, and a hypothesis space $\mathcal{H}$ consisting of probabilistic classifiers $h : \mathcal{X} \longrightarrow [0, 1]$. Assuming that each hypothesis $h = h_\theta$ is identified by a (unique) parameter vector $\theta \in \Theta$, we can equate $\mathcal{H}$ with the parameter space

Θ. We denote by $p_\theta(1 \mid \boldsymbol{x}) = h_\theta(\boldsymbol{x})$ and $p_\theta(0 \mid \boldsymbol{x}) = 1 - h_\theta(\boldsymbol{x})$ the (predicted) probability that instance $\boldsymbol{x} \in \mathcal{X}$ belongs to the positive and negative class, respectively.

Credal uncertainty sampling (Antonucci et al. 2012) seeks to differentiate between the *reducible* and *irreducible* part of the uncertainty in a prediction. Denote by $C \subseteq \Theta$ a *credal set* of models, i.e., a set of plausible candidate models. We say that a class y dominates another class y' if y is more probable than y' for each distribution in the credal set, that is

$$\gamma(y, y', \boldsymbol{x}) = \inf_{\theta \in C} \frac{p_\theta(y \mid \boldsymbol{x})}{p_\theta(y' \mid \boldsymbol{x})} > 1 \,. \tag{12}$$

The credal uncertainty sampling approach simply looks for the instance $\boldsymbol{x}$ with the highest uncertainty, i.e, the least evidence for the dominance of one of the classes. In the case of binary classification with $\mathcal{Y} = \{0, 1\}$, this is expressed by the score

$$s(\boldsymbol{x}) = -\max\left(\gamma(1, 0, \boldsymbol{x}), \gamma(0, 1, \boldsymbol{x})\right) . \tag{13}$$

Practically, the computations are based on the interval-valued probabilities

$$\underline{p}(y \mid \boldsymbol{x}), \overline{p}(y \mid \boldsymbol{x})] \,,$$

assigned to each class $y \in \{0, 1\}$, where

$$\underline{p}(y \mid \boldsymbol{x}) = \inf_{\theta \in C} p_\theta(y \mid \boldsymbol{x}) \,, \quad \overline{p}(y \mid \boldsymbol{x}) = \sup_{\theta \in C} p_\theta(y \mid \boldsymbol{x}) \,. \tag{14}$$

Such interval-valued probabilities can be produced within the framework of the Naïve credal classifier (Antonucci et al. 2012; Antonucci and Cuzzolin 2010; De Campos et al. 1994; Zaffalon 2002). In the case of binary classification, where $p_\theta(0 \mid \boldsymbol{x}) = 1 - p_\theta(1 \mid \boldsymbol{x})$, the score $\gamma(1, 0, \boldsymbol{x})$ can be rewritten as follows:

$$\gamma(1, 0, \boldsymbol{x}) = \inf_{\theta \in C} \frac{p_\theta(1 \mid \boldsymbol{x})}{p_\theta(0 \mid \boldsymbol{x})} = \inf_{\theta \in \Theta} \frac{p_\theta(1 \mid \boldsymbol{x})}{1 - p_\theta(1 \mid \boldsymbol{x})} = \frac{\underline{p}(1 \mid \boldsymbol{x})}{1 - \underline{p}(1 \mid \boldsymbol{x})} \,. \tag{15}$$

Likewise,

$$\gamma(0, 1, \boldsymbol{x}) = \inf_{\theta \in C} \frac{p_\theta(0 \mid \boldsymbol{x})}{p_\theta(1 \mid \boldsymbol{x})} = \inf_{\theta \in C} \frac{1 - p_\theta(1 \mid \boldsymbol{x})}{p_\theta(1 \mid \boldsymbol{x})} = \frac{1 - \overline{p}(1 \mid \boldsymbol{x})}{\overline{p}(1 \mid \boldsymbol{x})} \,. \tag{16}$$

Finally, the uncertainty score (13) can simply be expressed as follows:

$$s(\boldsymbol{x}) = -\max\left(\frac{\underline{p}(1 \mid \boldsymbol{x})}{1 - \underline{p}(1 \mid \boldsymbol{x})}, \frac{1 - \overline{p}(1 \mid \boldsymbol{x})}{\overline{p}(1 \mid \boldsymbol{x})}\right) . \tag{17}$$

3.3 Epistemic and aleatoric uncertainty (EAU)

A distinction between the *epistemic* and *aleatoric* uncertainty (Hora 1996) in a prediction for an instance $\boldsymbol{x}$ has been motivated by Senge et al. (2014)[3]. Their approach is based on

[3] More recently, this distinction has also attracted attention in the deep learning community (Kendall and Gal 2017).

the use of relative likelihoods, historically proposed by Birnbaum (1962) and then justified in other settings such as possibility theory (Walley and Moral 1999).

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

$$\pi_\Theta(\theta) = \frac{L(\theta)}{L(\theta^{ml})} = \frac{L(\theta)}{\max_{\theta' \in \Theta} L(\theta')}, \tag{18}$$

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

$$\pi(1 \mid \boldsymbol{x}) = \sup_{\theta \in \Theta} \min \left[\pi_\Theta(\theta), p_\theta(1 \mid \boldsymbol{x}) - p_\theta(0 \mid \boldsymbol{x})\right],$$
$$\pi(0 \mid \boldsymbol{x}) = \sup_{\theta \in \Theta} \min \left[\pi_\Theta(\theta), p_\theta(0 \mid \boldsymbol{x}) - p_\theta(1 \mid \boldsymbol{x})\right].$$

So, $\pi(1 \mid \boldsymbol{x})$ is high if and only if a highly plausible model supports the positive class much stronger (in terms of the assigned probability mass) than the negative class (and $\pi(0 \mid \boldsymbol{x})$ can be interpreted analogously)[4]. Note that, with $f(a) = 2a - 1$, we can also write

$$\pi(1 \mid \boldsymbol{x}) = \sup_{\theta \in \Theta} \min \left[\pi_\Theta(\theta), f(h_\theta(\boldsymbol{x}))\right], \tag{19}$$

$$\pi(0 \mid \boldsymbol{x}) = \sup_{\theta \in \Theta} \min \left[\pi_\Theta(\theta), f(1 - h_\theta(\boldsymbol{x}))\right]. \tag{20}$$

Given the above degrees of support, the degrees of epistemic uncertainty u_e and aleatoric uncertainty u_a are defined as follows:

$$u_e(\boldsymbol{x}) = \min \left[\pi(1 \mid \boldsymbol{x}), \pi(0 \mid \boldsymbol{x})\right], \tag{21}$$

$$u_a(\boldsymbol{x}) = 1 - \max \left[\pi(1 \mid \boldsymbol{x}), \pi(0 \mid \boldsymbol{x})\right]. \tag{22}$$

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 (22) is the degree to which none of the classes is supported. Roughly speaking, aleatoric uncertainty is due to influences on the data-generating process that are inherently random, whereas epistemic uncertainty is caused by a lack of knowledge. Or, stated differently, u_e and u_a measure the *reducible* and the *irreducible* part of the total uncertainty, respectively.

It is thus tempting to assume that epistemic uncertainty is more relevant for active learning: While it makes sense to query additional class labels in regions where uncertainty can be reduced, doing so in regions of high aleatoric uncertainty appears to be less reasonable. This leads us to suggest the principle of *epistemic uncertainty sampling*, which prescribes the selection

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \mathbf{U}} u_e(\boldsymbol{x}). \tag{23}$$

[4] Technically, we assume that, for each $\boldsymbol{x} \in \mathcal{X}$, there are hypotheses $h, h' \in \mathcal{H}$ such that $h(\boldsymbol{x}) \geq 0.5$ and $h'(\boldsymbol{x}) \leq 0.5$, which implies $\pi(1 \mid \boldsymbol{x}) \geq 0$ and $\pi(0 \mid \boldsymbol{x}) \geq 0$.

For comparison, we will also consider an analogous selection rule based on the aleatoric uncertainty, i.e.,

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathbf{U}} u_a(\boldsymbol{x})\,. \tag{24}$$

As already said, this approach is completely generic and can in principle be instantiated with any hypothesis space $\mathcal{H}$. The uncertainty measures (21–22) can be derived very easily from the support degrees (19–20). The computation of the latter may become difficult, however, as it requires the solution of an optimization problem, the properties of which depend on the choice of $\mathcal{H}$. We are going to present practical methods to determine (19–20) for the cases of a simple Parzen window classifier and logistic regression in Sects. A.1 and A.2, respectively.

4 Discussion and comparison of the approaches

4.1 EAU versus EBU

Although the concepts of "conflicting evidence" and "insufficient evidence" of Sharma and Bilgic (2017) appear to be quite related, respectively, to aleatoric and epistemic uncertainty, the correspondence becomes much less obvious (and in fact largely disappears) upon a closer inspection. Besides, a direct comparison is complicated due to various technical issues with the evidence-based approach to uncertainty sampling. In particular, due to the preselection of the top-t uncertain instances (the set $\mathbf{S}$), evidence-based uncertainty sampling is actually a variant of standard (entropy-based) uncertainty sampling, and completely degenerates to the latter for $t = 1$. As we are more interested in alternative measures of uncertainty, we will subsequently ignore the preselection step, and instead focus our discussion on the nature of the evidence measures themselves. In other words, we consider a version of evidence-based uncertainty sampling with a very large t. Before proceeding, let us emphasize that this is not the version proposed by the authors. Therefore, our discussion should be taken with a grain of salt.

As a first important observation, note that the evidences $E_0(\boldsymbol{x})$ and $E_1(\boldsymbol{x})$ solely depend on the *relation* of the class-conditional probabilities $p_\theta(x^m \mid 1)$ and $p_\theta(x^m \mid 0)$, which hides the number of training examples they have been estimated from, and hence their confidence. The latter, however, has an important influence on whether something is qualified as aleatorically or epistemically uncertain. As an illustration, consider a simple example with two binary attributes, the first with domain $\{a_1, a_2\}$ and the second with domain $\{b_1, b_2\}$. Denote by $n_{i,j} = (n^+_{i,j}, n^-_{i,j})$ the number of positive and negative examples observed for $(x_1, x_2) = (a_i, b_j)$. Here are two scenarios:

	b_1	b_2		b_1	b_2
a_1	(1, 1)	(1, 1)	a_1	(100, 100)	(100, 100)
a_2	(1, 1)	(1, 1)	a_2	(100, 100)	(100, 100)

In the both scenarios, the insufficient evidence would be high, because all class-conditional probabilities are equal. In EAU, however, the first scenario would largely be a case of epistemic uncertainty, due to the few number of training examples, whereas the second

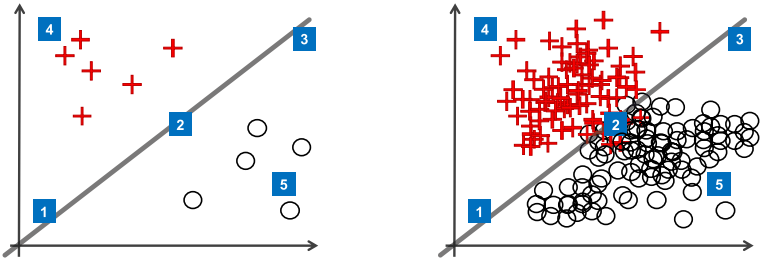


Fig. 1 Two scenarios for logistic regression: training data with positive (red crosses) and negative examples (black circles) and five query instances

would be aleatoric, because the equal posteriors[5] are sufficiently "confirmed". Similar remarks apply to conflicting evidence. In the scenario

	b_1	b_2
a_1	(1, 1)	(10, 1)
a_2	(1, 10)	(1, 1)

the latter would be high for (a_1, b_1), because $p_\theta(a_1 \mid 1) \gg p_\theta(a_1 \mid 0)$ and $p_\theta(b_1 \mid 0) \gg p_\theta(b_1 \mid 1)$. The same holds for (a_2, b_2), whereas the uncertainties for (a_1, b_2) and (a_2, b_1) would be low. Note, however, that in all these cases, exactly the same conditional probability estimates $p_\theta(x^m \mid 1)$ and $p_\theta(x^m \mid 0)$ are involved.

We would argue that epistemic uncertainty should directly refer to these probabilities, because they constitute the parameter θ of the model. Thus, to reduce epistemic uncertainty (about the right model θ), one should look for those examples that will mostly improve the estimation of these probabilities. Aleatoric uncertainty may occur in cases of posteriors close to 1/2, in which the conflicting evidence may indeed be high (although, as already mentioned, the latter ignores the class priors). Yet, we would not necessarily call such cases a "conflict", because the predictions are completely in agreement with the underlying model (Naïve Bayes), which assumes class-conditional independence of attributes, i.e., an independent combination of evidences on different attributes.

Another illustration is provided in Fig. 1, now for the case of logistic regression. A first important observation is that the uncertainties due to conflicting and insufficient evidence are exactly the same in both scenarios in Fig. 1, the left and the right one. This is because these uncertainties are merely derived from the single model h_θ learned from the training data. Thus, like in the example for Naïve Bayes, the evidence-based approach does not capture *model uncertainty*, i.e., uncertainty about the truly optimal model (which is clearly larger on the left and smaller on the right), which EAU essentially measures in terms of epistemic uncertainty.

For the first three queries, the evidence-based uncertainties are very different: The first query has a high insufficient-evidence uncertainty, the third has a high

[5] The class priors are ignored here.

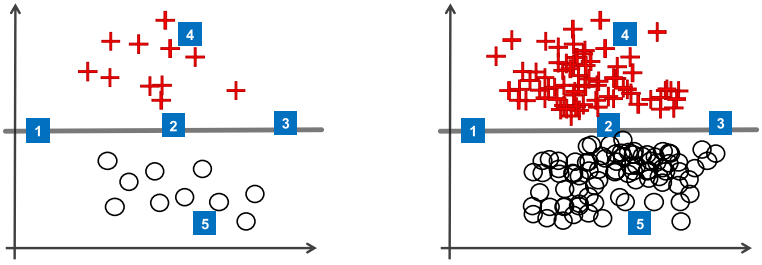


Fig. 2 Two scenarios for logistic regression: training data with positive (red crosses) and negative examples (black circles) and five query instances

conflicting-evidence uncertainty, and the second none of the two. According to EAU, the uncertainties for these three cases are all high (because they are all located close to the decision boundary) and, more importantly, of the same nature: mostly aleatoric in the right and a mix of aleatoric and epistemic in the left scenario. For the second, fourth and fifth query, the evidence-based uncertainties are roughly the same. Again, this is very different from EAU, which assigns a high uncertainty to the second but very low uncertainties to the fourth and fifth query.

Figure 2 shows two very similar scenarios, but now with a bias term. As already said, the evidence-approach does not account for such a bias. Moreover, since $w_1 \approx 0$ (the first feature does not seem to have an influence), there is essentially no negative evidence, i.e., $E_0(\boldsymbol{x})$ is always close to 0. Consequently, the product $E_0(\boldsymbol{x}) \times E_1(\boldsymbol{x})$ will be small, too, suggesting that conflicting-evidence uncertainty is always low and insufficient-evidence uncertainty always high.

As shown by these examples, the additional uncertainty captured by EBU is very different from aleatoric and epistemic uncertainty in EAU. In particular, the evidence-based approach can be criticized for ignoring model uncertainty as well as properties of the model class. Although the measures of evidence for the positive and negative class, such as (10), are derived from the model, the evidences are "feature-based" in the sense of considering the evidence provided by each feature in isolation. What is not taken into account, however, is the way in which the model combines the features into an overall prediction. In logistic regression, for example, the features are linearly combined into a single score, and the class probabilities are expressed as a function of this score. For instance, a model like the one we considered in our example,

$$p(y = 1 \,|\, \boldsymbol{x}) = \frac{1}{1 + \exp(-\gamma(x_2 - x_1))} \,,$$

assumes that the probability of the positive class is a function of the *difference* between x_2 and x_1. One may wonder, therefore, why one should consider a case where both x_1 and x_2 are large as a *conflict* (and, likewise, a case with both values being small as not providing sufficient evidence for a prediction). From this point of view, the very idea of conflicting (and, likewise, insufficient) evidence may appear somewhat questionable.

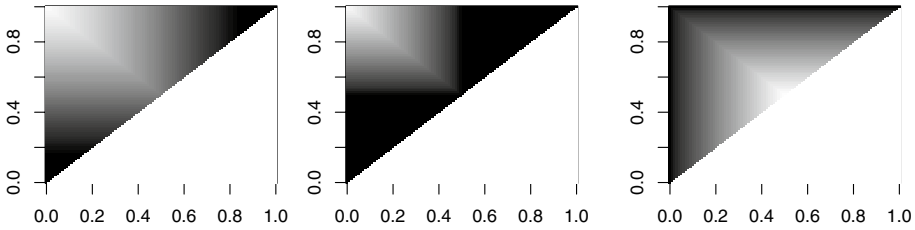


Fig. 3 From left to right: exponential rescaling of the credal uncertainty measure (17), epistemic uncertainty u_e and aleatoric uncertainty u_a for intervals $[\underline{p}, \overline{p}]$ with lower probability $\underline{p}$ (x-axis) and upper probability $\overline{p}$ (y-axis). Lighter colors indicate higher values

4.2 EAU versus CU

Credal uncertainty (sampling) seems to be closer to EAU, at least in terms of the underlying principle. In both approaches, model uncertainty is captured in terms of a set of plausible candidate models from the underlying hypothesis space, and this (epistemic) uncertainty about the right model is translated into uncertainty about the prediction for a given $\boldsymbol{x}$. In credal uncertainty sampling, the candidate set is given by the credal set C, which corresponds to the distribution π_Θ in EAU—as a difference, we thus note that the latter is a "graded set", to which a candidate θ belongs with a certain degree of membership (the relative likelihood), whereas a credal set is a standard set in which a model is either included or not. Using machine learning terminology, C plays the role of a *version space* (Mitchell 1977), whereas π_Θ represents a kind of generalized (graded) version space (Hüllermeier 2003).

More specifically, the wider the interval $[\underline{p}(1 \,|\, \boldsymbol{x}), \overline{p}(1 \,|\, \boldsymbol{x})]$ in (17), the larger the score $s(\boldsymbol{x})$, with the maximum being obtained for the case [0, 1] of complete ignorance. This is well in agreement with the degree of epistemic uncertainty in EAU. In the limit, when $[\underline{p}(1 \,|\, \boldsymbol{x}), \overline{p}(1 \,|\, \boldsymbol{x})]$ reduces to a precise probability $p(1 \,|\, \boldsymbol{x})$, i.e., the epistemic uncertainty disappears, (17) is maximal for $p(1 \,|\, \boldsymbol{x}) = 1/2$ and minimal for $p(1 \,|\, \boldsymbol{x})$ close to 0 or 1. Again, this behavior is in agreement with the conception of aleatoric uncertainty in EAU. More generally, comparing two intervals of the same length, (17) will be larger for the one that is closer to the middle point 1/2. Thus, it seems that the credal uncertainty score (17) combines both epistemic and aleatoric uncertainty in a single measure.

Yet, upon closer examination, its similarity to epistemic uncertainty is much higher than the similarity to aleatoric uncertainty. Note that, for EAU, the special case of a credal set C can be imitated with the measure $\pi_\Theta(\theta) = 1$ if $\theta \in C$ and $\pi_\Theta(\theta) = 0$ if $\theta \notin C$. Then, (19) and (20) become

$$\pi(1 \,|\, \boldsymbol{x}) = \sup_{\theta \in C} \max[\, 2\, p_\theta(1 \,|\, \boldsymbol{x}) - 1, 0\,] = \max[\, 2\, \overline{p}(1 \,|\, \boldsymbol{x}) - 1, 0\,],$$
$$\pi(0 \,|\, \boldsymbol{x}) = \sup_{\theta \in C} \max[\, 2\, p_\theta(0 \,|\, \boldsymbol{x}) - 1, 0\,] = \max[\, 1 - 2\, \underline{p}(1 \,|\, \boldsymbol{x}), 0\,],$$

and u_e and u_a can be derived from these values as before. Figure 3 shows a graphical illustration of the credal uncertainty score[6] (17) as a function of the probability bounds $\underline{p}$ and

[6] The score s is not well scaled, and may assume very large negative values. For better visibility, we therefore plotted the monotone transformation $\exp(s)$.

Table 1 Data sets used in the experiments

#	name	# instances	# features	attributes	% of major class
1	parkinsons	197	22	real	75.0
2	vertebral	310	6	real	67.7
3	ionosphere	351	34	real	64.1
4	climate	540	18	real	91.4
5	breast	569	30	real	62.7
6	blood	748	5	real	76.2
7	QSAR	1055	41	integer, real	66.2
8	banknote	1372	4	real	55.0
9	madelon	4400	500	real	50.0
10	spambase	4601	57	integer, real	60.5

$\overline{p}$, and the same illustration is given for epistemic uncertainty u_e and aleatoric uncertainty u_a. From the visual impression, it is clear that the credibility score closely resembles u_e, while behaving quite differently from u_a. This impression is corroborated by a simple correlation analysis, in which we ranked the intervals

$$[\underline{p}, \overline{p}] \in \left\{ I_{a,b} = \left[\frac{a}{100}, \frac{b}{100} \right] \middle| a, b \in \{0, 1, \ldots, 100\}, a \leq b \right\},$$

i.e., a quantization of the class of all probability intervals, according to the different measures, and then computed the Kendall rank correlation. While the ranking according to (17) is strongly correlated with the ranking for u_e (Kendall is around 0.86), it is almost uncorrelated with u_a.

In summary, the credal uncertainty score appears to be quite similar to the measure of epistemic uncertainty in EAU. As potential advantages of the latter, let us mention the following points. First, the degree of epistemic uncertainty is normalized and bounded, and thus easier to interpret. Second, it is complemented by a degree of aleatoric uncertainty—the two degrees are carefully distinguished and have a clear semantics. Third, handling candidate models in a graded manner, and modulating their influence according to their plausibility, appears to be more reasonable than creating an artificial separation into plausible and non-plausible models (i.e., the credal set and its complement).

5 Experiments

This section starts with a description of the experimental setting and the data sets used in the experiments. Some technical details, e.g., regarding the choice of the model parameters and instantiations of aleatoric and epistemic uncertainty, are deferred to Sects. 1.1 and 1.2 in the appendix. Finally, the results of the experiments are presented and analyzed.

5.1 Data sets and experimental setting

We perform experiments on binary classification data sets from the UCI repository[7], the properties of which are summarized in Table 1. To make sure that the data is amenable to all methods without the need for further preprocessing, we only selected data with numerical features. Each data set is randomly split into 10% training, 80% pool, and 10% test data. The training data is used to obtain an initial model. Then, in each iteration, the learner is allowed to evaluate the instances from the pool and query a (mini-)batch of these instances — according to the strategy of uncertainty sampling, the learner selects those instances with the highest degrees of uncertainty. The chosen instances are labelled (by an oracle or expert) and added to the training data $\mathbf{D}$, on which the model is then re-trained. The budget of the active learner is fixed to the size of the pool, and the performance of the classifiers is monitored over the entire active learning process. The whole procedure is repeated 1000 times and test accuracies are averaged. The following variants of uncertainty sampling are included in the experimental studies:

- Rand: Random sampling
- ENT: Standard uncertainty sampling based on the entropy measure[8] (1)
- CEU: Conflicting-evidence uncertainty sampling (8)
- IEU: Insufficient-evidence uncertainty sampling (9)
- CU: Credal uncertainty sampling[9] (17)
- EU: Epistemic uncertainty sampling(21)
- AU: Aleatoric uncertainty sampling (22)

5.1.1 Local learning

By local learning, we refer to a class of non-parametric models that derive predictions from the training information in a local region of the instance space, for example the local neighborhood of a query instance (Bottou and Vapnik 1992; Cover and Hart 1967). As a simple example, we consider the Parzen window classifier (Chapelle 2005), to which most of the mentioned approaches can be applied in a quite straightforward way. For a given instance $\boldsymbol{x}$, we define the set of its neighbours as follows:

$$R(\boldsymbol{x}, \epsilon) = \left\{(\boldsymbol{x}_i, y_i) \in \mathbf{D} \mid \|\boldsymbol{x}_i - \boldsymbol{x}\| \leq \epsilon\right\},$$

where ϵ is the width of the Parzen window. In binary classification, a local region $R(\boldsymbol{x}, \epsilon)$ can be associated with a constant hypothesis $h_\theta, \theta \in \Theta = [0, 1]$, where $p_\theta(1|\boldsymbol{x}) = h_\theta(\boldsymbol{x}) \equiv \theta$. With p and n the number of positive and negative instances, respectively, within a Parzen window $R(\boldsymbol{x}, \epsilon)$, the likelihood function and the maximum likelihood estimate are, respectively, given by

[7] http://archive.ics.uci.edu/ml/index.php

[8] In all the experiments, we use entropy as a measure of uncertainty to select unlabelled instances.

[9] The interval-valued probabilities associated to each region can be determined using the numbers of positive and negative instances as described in (Antonucci et al. 2012).

$$L(\theta) = \binom{p+n}{p} \theta^p (1-\theta)^n \,, \text{ and } \hat{\theta} = \frac{p}{p+n} \,. \tag{25}$$

Since the likelihood function is well-defined, we can determine the degrees of epistemic and aleatoric uncertainty as described in Sect. 3.3; we refer to Sect. A.1 for the technical details.

How to determine the width ϵ of the Parzen window? This value is difficult to assess, and an appropriate choice strongly depends on properties of the data and the dimensionality of the instance space. Intuitively, it is even difficult to say in which range this value should lie. Therefore, instead of fixing ϵ, we fixed an absolute number K of neighbors in the training data, which is intuitively more meaningful and easier to interpret. A corresponding value of ϵ is then determined in such a way that the average number of nearest neighbours of instances $\boldsymbol{x}_i$ in the training data $\mathbf{D}$ is just K. In other words, ϵ is determined indirectly via K. Furthermore, since we are not, in the first place, interested in maximizing performance, but in analyzing the effectiveness of active learning approaches, we simply fix the neighborhood size K as the square root of the size of the data set (number of instances in the initial training and pool set) as suggested by Lall and Sharma (1996). A practical algorithm for determining ϵ given K, and the way in which we handle empty Parzen windows, are also given in Sect. A.1.

In a similar way, the approach can be applied to decision tree learning (Quinlan 1986; Safavian and Landgrebe 1991). In fact, recall that a decision tree partitions the instance space $\mathcal{X}$ into (rectangular) regions $R_1, \ldots, R_L$ (i.e., $\bigcup_{i=1}^{L} R_i = \mathcal{X}$ and $R_i \cap R_j = \emptyset$ for $i \neq j$) associated with corresponding leafs of the tree (each leaf node defines a region R). Again, in the case of binary classification, we can assume each region R to be associated with a constant hypothesis h_θ, $\theta \in \Theta = [0,1]$, where $h_\theta(\boldsymbol{x}) \equiv \theta$ is the probability of the positive class. Therefore, degrees of epistemic and aleatoric uncertainty can be derived in the same way as described for Parzen window.

For the Parzen window classifier and decision trees[10], we fixed the batch size to 1% of the initial pool dataset. For the approach based on credal uncertainty (CU), we determine the lower and upper probabilities based on the number of positive and negative examples in a region, following the procedure described in "Appendix 1 and 2" of (Antonucci and Cuzzolin 2010). Note that the evidence-based approach (Sect. 3.1) is not immediately applicable to these learners, and therefore omitted from the experiments.

5.1.2 Logistic regression

In contrast to nonparametric, local learning methods such as the Parzen window classifier, logistic regression is a parametric class of linear models, and hence coming with comparatively restrictive assumptions. Recall that logistic regression assumes posterior probabilities to depend on feature vectors $\boldsymbol{x} = (x^1, \ldots, x^d) \in \mathbb{R}^d$ in the following way:

$$h(\boldsymbol{x}) = p(1 \mid \boldsymbol{x}) = \frac{\exp\left(\theta_0 + \sum_{i=1}^{d} \theta_i x^i\right)}{1 + \exp\left(\theta_0 + \sum_{i=1}^{d} \theta_i x^i\right)} \,.$$

[10] For an implementation in Python, see https://scikit-learn.org/stable/modules/tree.html

This means that learning the model comes down to estimating a parameter vector $\theta = (\theta_0, \ldots, \theta_d)$, which is commonly done through likelihood maximization (Menard 2002). For numerical stability, we employ L_2-regularization, which comes down to maximizing the following strictly concave function (Rennie 2005):

$$l(\theta) = \log L(\theta) = \sum_{n=1}^{N} y_n \left(\theta_0 + \sum_{i=1}^{d} \theta_i x_n^i \right) \tag{26}$$

$$- \sum_{n=1}^{N} \ln \left(1 + \exp \left(\theta_0 + \sum_{i=1}^{d} \theta_i x_n^i \right) \right) - \frac{\gamma}{2} \sum_{i=1}^{d} \theta_i^2 , \tag{27}$$

where the regularization term γ is fixed to 1. On the basis of this likelihood function, the degrees of epistemic and aleatoric uncertainty can again be determined as described in Sect. 3.3; as before, technical details and a practical algorithm are deferred to Sect. 1.2 in the appendix.

For the case of logistic regression, the evidence-based approach can be applied as well (cf. Sect. 3.1.3). Following Sharma and Bilgic (2017), we set the number of the top uncertain instances to be evaluated to 5 times of the batch size.

5.2 Results

As can be seen in Fig. 4, in the case of the Parzen window classifier, EU performs the best and AU the worst. Moreover, standard uncertainty sampling (ENT) and random sampling are in-between the two. This is in agreement with our expectations and supports our conjecture that, from an active learning point of view, epistemic uncertainty is the more useful information. Even if the improvements compared to ENT are not huge, they are still visible and quite consistent. The performance provided by CU is competitive to the one of EU, and again in agreement with our expectations — as discussed in Sect. 4.2, both CU and EU have the ability of capturing model uncertainty. The results for decision tree learning (cf. Fig. 5) are quite similar. Now, however, standard uncertainty sampling based on entropy performs worse, and the advantage of epistemic uncertainty sampling is even more pronounced.

In the case of logistic regression (cf. Fig. 6), the picture looks a bit different. Here, epistemic, aleatoric, and standard uncertainty sampling perform more or less the same (and all significantly better than random sampling), whereas no general pattern can be drawn for the evidence-based uncertainty measures. As a plausible explanation, note that, in contrast to the local learning methods in the first experiment, logistic regression comes with a very strong learning bias in the form of a linearity assumption. Therefore, the epistemic (or model) uncertainty disappears quite quickly: The linear decision boundary stabilizes relatively early in the learning process, and then, the learner is rather "certain" about its predictions (regardless of whether this certainty is warranted or not). According to the logistic model, the uncertain cases are those closest to the current decision boundary, some of them with a slightly higher epistemic and others with a higher aleatoric uncertainty. In any case, all three methods, EU, AU, and ENT, are sampling near the decision boundary. Thus, it is hardly surprising that they show similar performance.

Overall, the experiments nevertheless confirm that, in the context of uncertainty sampling for active learning, epistemic uncertainty is a viable alternative to standard

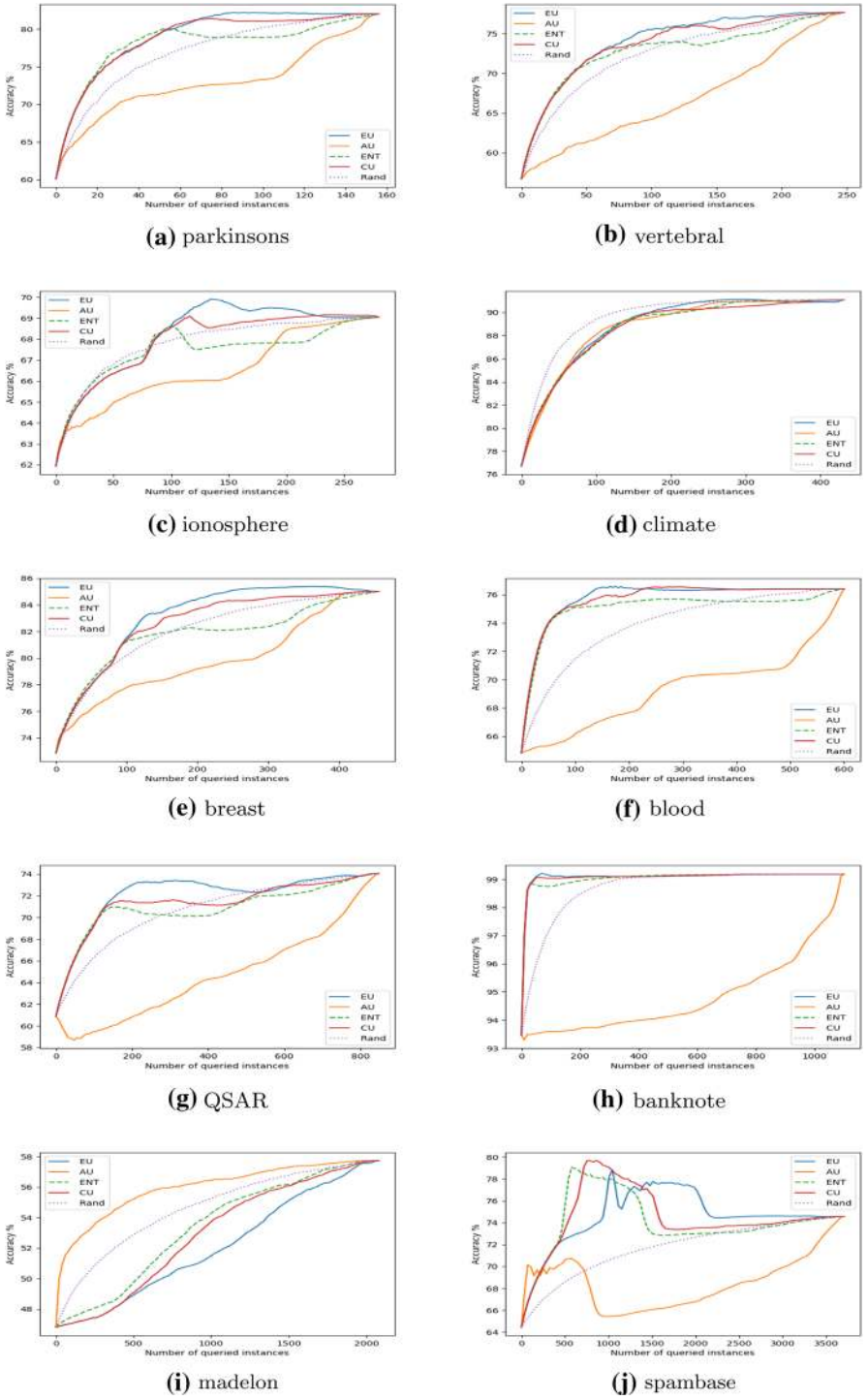
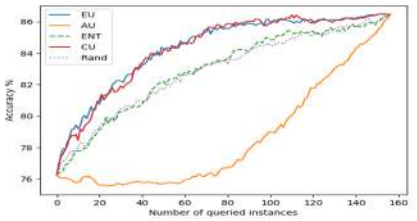
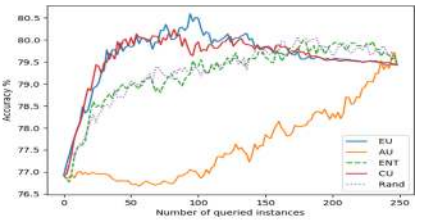


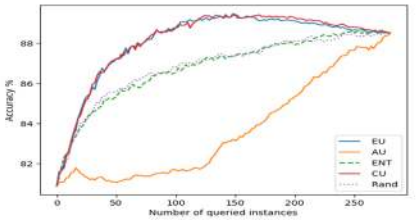
Fig. 4 Average accuracies (y-axis) for the Parzen window classifiers as a function of the number of instances queried from the pool (x-axis)



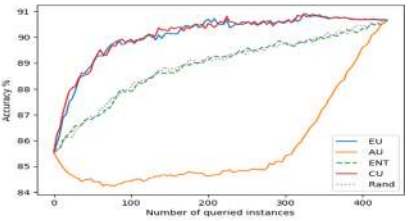
(a) parkinsons



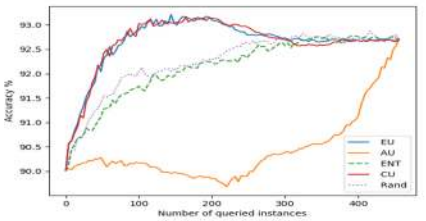
(b) vertebral



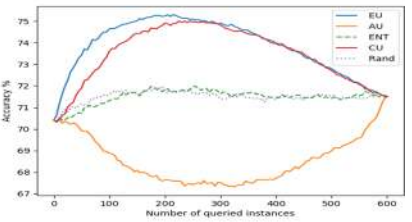
(c) ionosphere



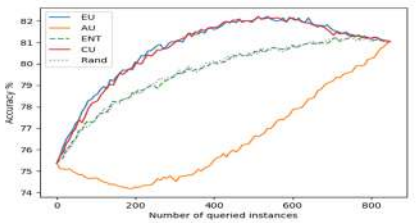
(d) climate



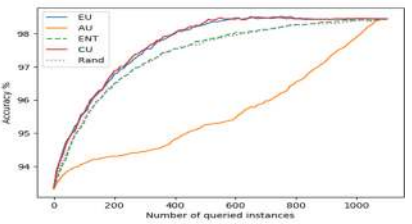
(e) breast



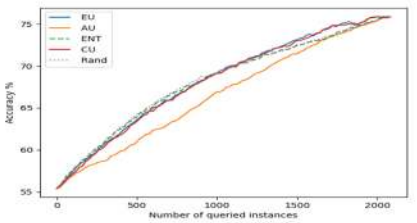
(f) blood



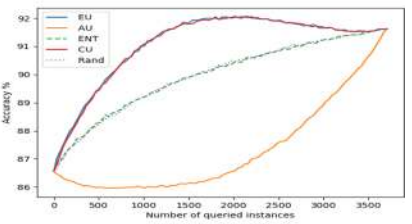
(g) QSAR



(h) banknote



(i) madelon



(j) spambase

Fig. 5 Average accuracies (y-axis) for the decision trees as a function of the number of instances queried from the pool (x-axis)

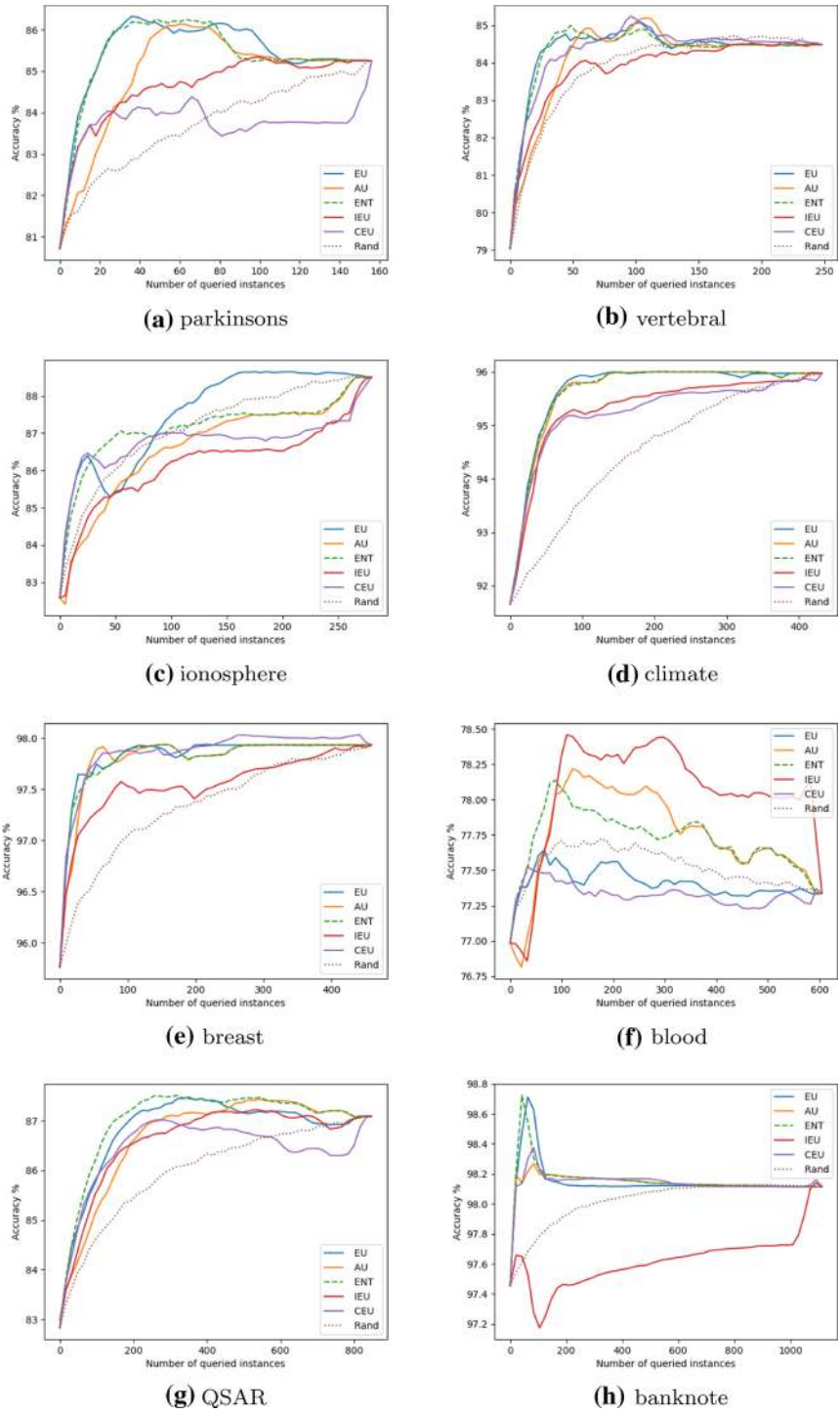


Fig. 6 Average accuracies (y-axis) for logistic regression as a function of the number of instances queried from the pool (x-axis)

uncertainty measures like entropy: For local learning methods, in which epistemic uncertainty tends to be higher, epistemic uncertainty sampling improves upon standard uncertainty sampling, and for global methods with a strong learning bias, it performs at least on a par. Credal uncertainty, which behaves similarly to epistemic uncertainty (cf. Sect. 4.2), shows strong performance as well.

As an aside, note that the learning curves are not all monotone increasing, which might be surprising at first sight. Actually, however, this kind of behavior is not uncommon and may occur if a data set, in addition to useful examples, also comprises low-quality (e.g., noisy or otherwise misleading) instances. In this case, a strong active learning strategy may succeed in selecting the informative, high-quality examples first, leading to a good model with strong predictive performance. In the end, if the pool needs to be exhausted, the active learner is "forced" to pick the low-quality examples, too, thereby causing a drop in performance.

5.3 Influence of model bias

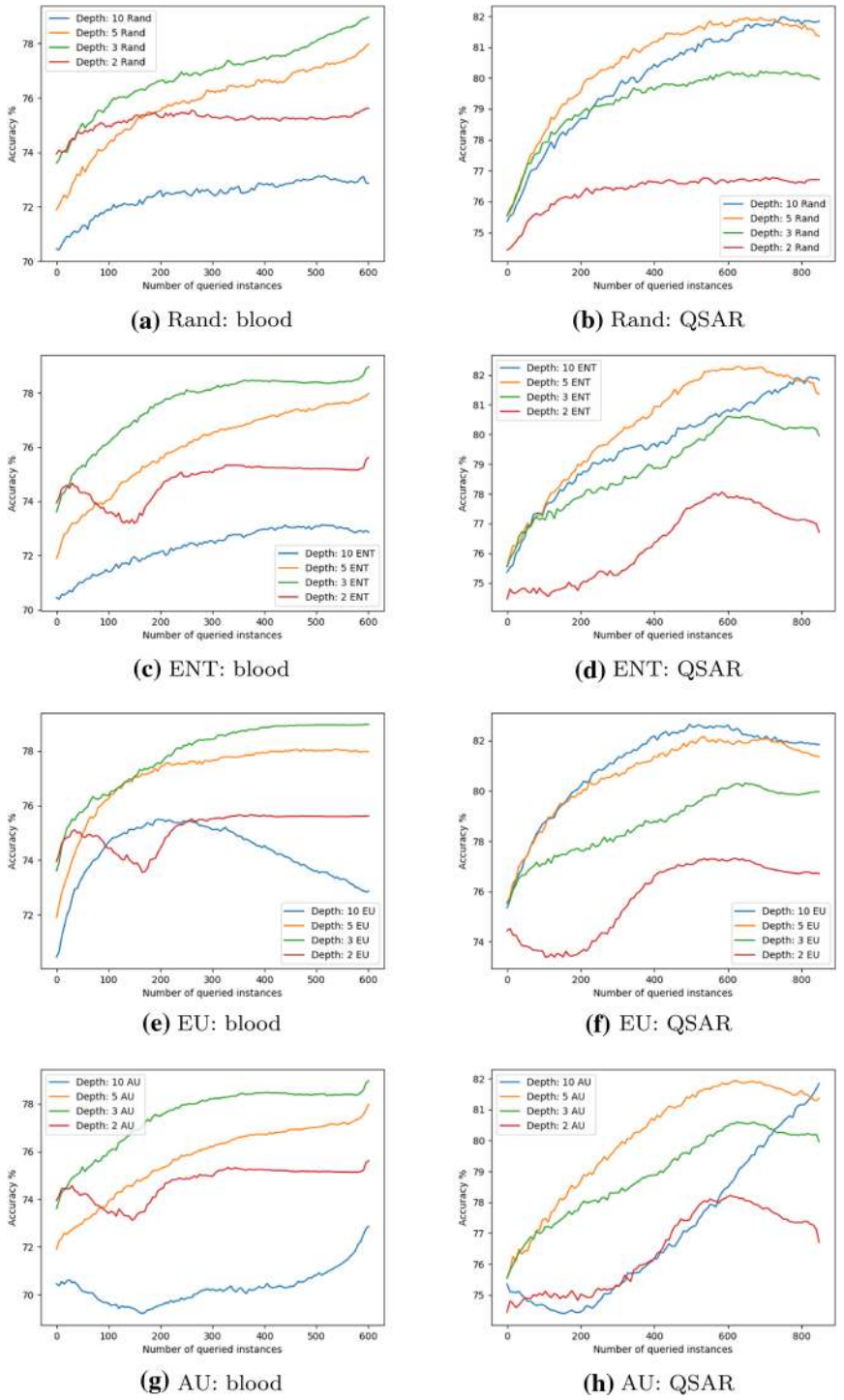
The results presented above suggest that epistemic uncertainty might be more advantageous for (active) learners with a low bias and less so for learners with a strong bias. To corroborate this conjecture, we conducted an additional experiment, using decision trees with a maximum depth limit as model classes. This allows for contolling the bias in a seamless manner: The higher the depth limit, the less restricted the model class, the lower the bias.

Figure 7 shows the learning curves for the depth limits $\{2, 3, 5, 10\}$ on the data sets blood and QSAR. As expected, different depth limits appear to be optimal for different problems (the best limit for blood is 3, for QSAR 10). However, more interesting for our purpose is the slope of the learning curves, which indeed seem to support our conjecture: For epistemic uncertainty, the learning curves increase faster for larger and slower for lower depth limits — for aleatoric uncertainty, it is just the other way around. To make this even clearer, Fig. 8 plots the *relative* performance in comparison to standard (entropy-based) uncertainty sampling, i.e., the performance ratio, both for epistemic and aleatoric uncertainty. As can be seen, EU tends to be superior, because the ratio is mostly larger than 1, while the opposite holds for AU. Again more importantly, the depth limit (bias) is in perfect agreement with the "order" of the curves: The higher the limit, the better EU (worse AU) in terms of relative performance.

5.4 Uncertainty as stopping criterion

In a last experiment, we analyze the potential of epistemic uncertainty to serve as a stopping criterion for an active learning process. Indeed, this appears to be a rather natural idea, because epistemic uncertainty — as opposed to aleatoric or total uncertainty — reflects the state of knowledge of the learner, and the potential to improve this knowledge through additional data. If the epistemic uncertainty is low for all instances remaining in the pool, this suggests that almost nothing can be gained anymore through additional sampling.

The criterion just outlined is an instance of the third type of stopping criteria commonly used in active learning (Li and Sethi 2006; Zhu et al. 2010): The active learning process ends if



(a) Rand: blood **(b)** Rand: QSAR

(c) ENT: blood **(d)** ENT: QSAR

(e) EU: blood **(f)** EU: QSAR

(g) AU: blood **(h)** AU: QSAR

Fig. 7 Average accuracies (y-axis) for decision tree as a function of the number of instances queried from the pool (x-axis)

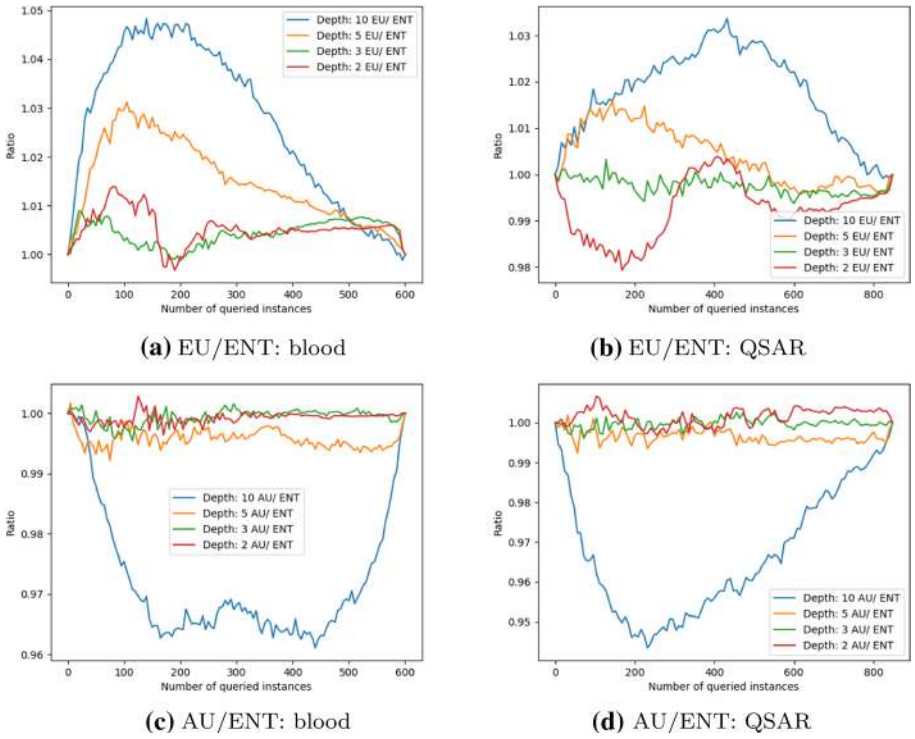


Fig. 8 Average accuracies (y-axis) for decision tree as a function of the number of instances queried from the pool (x-axis)

- the training data set reaches a desired size;
- a targeted performance level is achieved;
- no informative examples are available anymore.

While it is difficult to pre-define either a desirable size of the training data set or a targeted performance level, the last criterion can be easily implemented by setting some predefined uncertainty threshold and stopping the active learning process if the degree of uncertainty falls below the threshold (Zhu et al. 2010).

The potential usefulness of epistemic uncertainty is confirmed by our results (shown in Fig. 9 for two data sets–results for the other data sets are similar and can be found in Sect. 2 in the appendix).

6 Conclusion

This paper reconsiders the principle of uncertainty sampling in active learning from the perspective of uncertainty modeling and quantification. More specifically, it starts from the supposition that, when it comes to the question which instances to select from a pool of

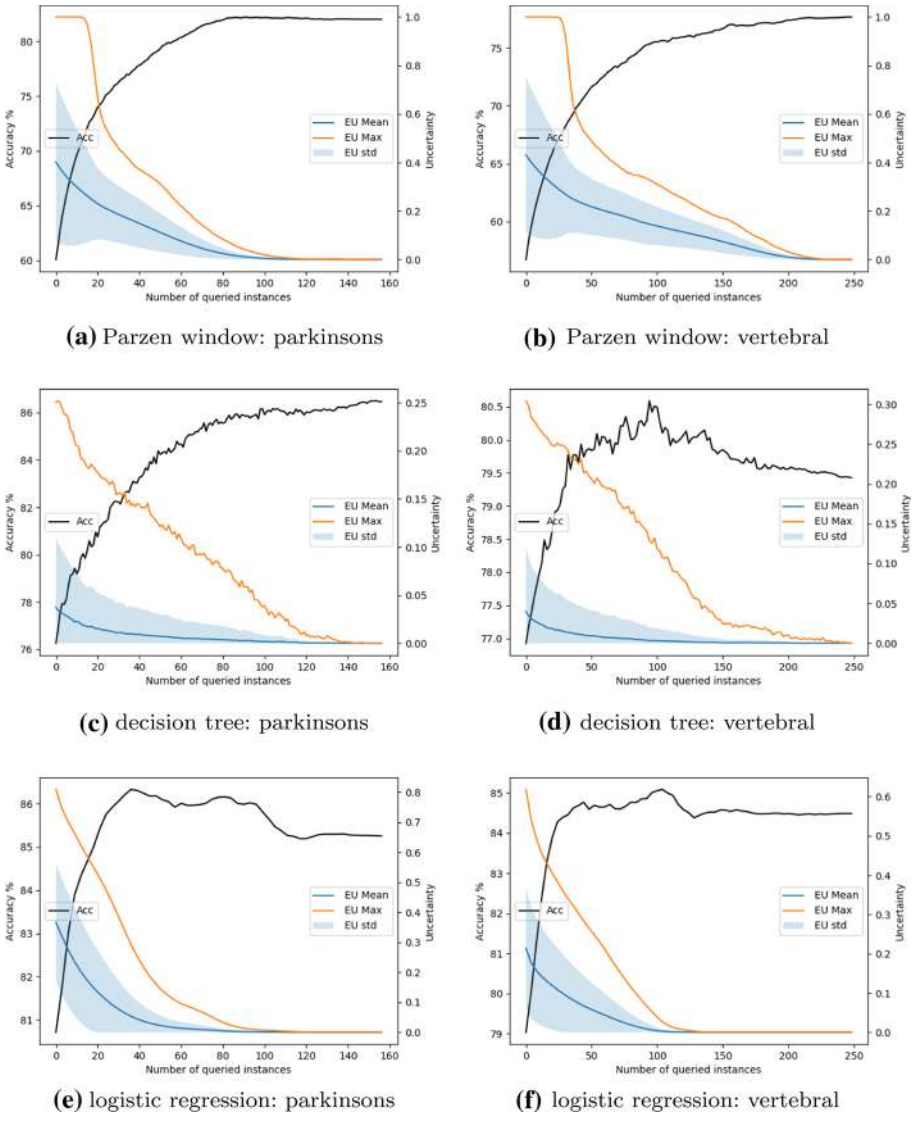


Fig. 9 Average accuracies and degrees of epistemic uncertainty (mean and maximum over instances in the pool, y-axis) as functions of the number of instances queried from the pool (x-axis)

candidates, a learner's predictive uncertainty due to "not knowing" should be more relevant than its uncertainty due to confirmed randomness.

To corroborate this conjecture, we revisited recent approaches to uncertainty quantification in machine learning, with a specific emphasis on methods that allow for separating different types of uncertainty, and incorporated them in the general uncertainty sampling procedure. Following a comparison and critical discussion of these approaches, a series of experiments with different learning algorithms was conducted. In these experiments, a distinction between so-called epistemic and aleatoric uncertainty proved to be especially

useful. More specifically, *epistemic uncertainty sampling* in the sense of uncertainty sampling based on measures of epistemic uncertainty in a prediction shows strong performance and consistently improves on standard uncertainty sampling. These results, which we interpret as clear evidence in favor of our conjecture, and indeed quite plausible: Epistemic and aleatoric uncertainty can be thought of, respectively, as the reducible and irreducible part of the total uncertainty. Consequently, querying an instance with a high epistemic uncertainty may provide useful information for the learner, whereas an aleatorically uncertain instance is unlikely to do so.

Given this affirmation, we are now encouraged to elaborate on epistemic uncertainty sampling in more depth, and to develop it in more sophistication. In this regard, there are various directions to be followed:

- Depending on the underlying model class, the quantification of epistemic uncertainty based on the generic approach by Senge et al. (2014) can be computationally expensive. Therefore, efficient instantiations for important learning methods would be desirable.
- Similar approaches for measuring epistemic uncertainty, which have been proposed in the literature more recently, should be investigated as possible alternatives (Depeweg et al. 2018).
- Our experimental results suggest that a distinction between epistemic and aleatoric uncertainty is more useful for learners with a weak inductive bias and less useful for learners with a strong bias. This observation ought to be analyzed in more detail and corroborated by further experiments. In fact, the learning algorithms included in our study constitute extremes on this spectrum (Parzen classifier and decision trees have a very low bias, logistic regression a very strong one), and additional experiments with learners having a "mediocre" bias would certainly be useful.
- Quite interestingly, the very notion of epistemic uncertainty seems to share many commonalities with other principles that have been suggested for active learning — probably not by chance. One example is so-called *expected model change* or the related principle of *expected model output change* (EMOC), where the idea is to query instances that, if added to the training data, are likely to cause large changes of the hypothesis or the predictions produced by the hypothesis (Freytag et al. 2014). According to our quantification of epistemic uncertainty (but also other formalizations), such instances should also have a high epistemic uncertainty. Therefore, epistemic uncertainty sampling seems to have much in common with EMOC, perhaps with the notable difference that the former looks at the uncertainty for a single instance, whereas the latter considers the expected change over all instances. Nevertheless, elaborating on this connection more closely seems to be worthwhile. The same holds true for another well-established active learning strategy, namely query-by-committee (QBC) approach (Seung et al. 1992). In fact, the diversity of the predictions of an ensemble of hypotheses, which is used as a selection criterion in QBC, has recently also been advocated as a suitable means for quantifying epistemic uncertainty (Shaker and Hüllermeier 2020).
- It might be interesting to combine the quantification of uncertainty with the notion of *relevance* or *representativeness* in active learning (McCallum and Nigam 1998; Lindenbaum et al. 2004): How relevant is a certain improvement of the current model for the overall performance of the learner? In local learning algorithms, for example, epistemic uncertainty tends to be high in sparse regions of the instance space, so that an active learner is tempted to sample here. At the same time, however, such regions appear to be less important for the generalization performance, simply because future queries will more likely

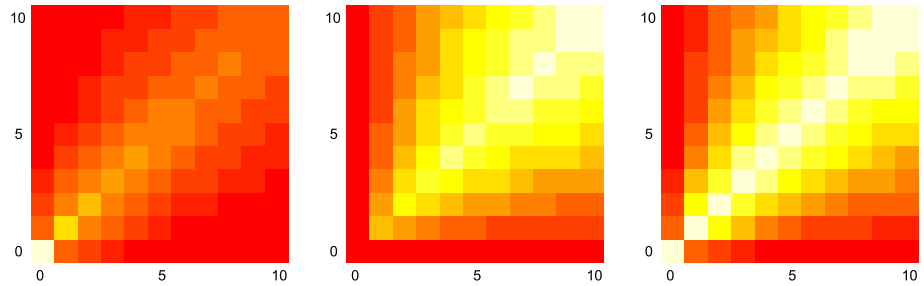


Fig. 10 From left to right: epistemic, aleatoric, and total uncertainty (epistemic + aleatoric) as a function of the numbers $p, n \in \{0, 1, \ldots, 10\}$ of positive and negative examples in a region (Parzen window) of the instance space (lighter colors indicate higher values)

occur in dense regions. Overall, a small improvement in a dense region may thus be more beneficial than a big improvement in a sparse region. This observation motivates a kind of *density weighting*, i.e., the combination (multiplication) of an uncertainty degree with the (estimated) density of a data point (Krempl et al. 2015).

- Last but not least, going beyond uncertainty sampling for binary classification as considered in this paper, the idea of epistemic uncertainty sampling should also be extended toward other learning problems, such as multi-class classification and regression.

Appendix 1: Instantiations of aleatoric and epistemic uncertainty

This section presents an instantiation of the approach outlined in Sect. 3.3 for the case of local learning using a Parzen window classifier (Chapelle 2005), as well as logistic regression. As already said, instantiating the approach essentially means to address the question of how to compute the degrees of support (19–20), from which everything else can easily be derived.

Appendix 1.1: Local learning

With the likelihood function and the the maximum likelihood estimate defined in (25), the degrees of support for the positive and negative classes are

$$\pi(1 \,|\, \boldsymbol{x}) = \sup_{\theta \in [0,1]} \min \left(\frac{\theta^p (1-\theta)^n}{\left(\frac{p}{n+p}\right)^p \left(\frac{n}{n+p}\right)^n}, \, 2\theta - 1 \right), \tag{28}$$

$$\pi(0 \,|\, \boldsymbol{x}) = \sup_{\theta \in [0,1]} \min \left(\frac{\theta^p (1-\theta)^n}{\left(\frac{p}{n+p}\right)^p \left(\frac{n}{n+p}\right)^n}, \, 1 - 2\theta \right). \tag{29}$$

Solving (28) and (29) comes down to maximizing a scalar function over a bounded domain, for which standard solvers can be used. We applied Brent's method[11] (which is a variant of the golden section method) to find a local minimum in the interval $\theta \in [0, 1]$. From (28–29), the epistemic and aleatoric uncertainty associated with the region R can be derived according to (23) and (24), respectively. For different combinations of n and p, these uncertainty degrees can be pre-computed (cf. Fig. 10).

Algorithm 2: Determining the width ϵ.

Input: **D**-normalized data, K-number
Output: the local width ϵ_K

```
foreach x_n ∈ D do
    foreach x_m ≠ x_n do
        compute d(x_n, x_m);
    form 1 × (n − 1) vector d_n = (d(x_n, x_m) | n ≠ m);
    sort d_n by increasing order and determine the K-th element d_n^K;
return ε_K = (Σ_{n=1}^{|D|} d_n^K) / |D|;
```

Given a number K, the corresponding ϵ is determined using Algorithm 2. Furthermore, as K is an average, individual instances may have more or less neighbors in their Parzen windows. In particular, a Parzen window may also be empty. In this case, we set $u_e(\boldsymbol{x}) = 1$ by definition, i.e., we consider this as a case of full epistemic uncertainty. Likewise, the uncertainty is considered to be maximal for all other sampling techniques. If the accuracy of the Parzen classifier needs to be determined, we assume that it yields a wrong prediction.

For the approach based on credal uncertainty (CU), we determine the lower and upper probabilities based on the number of positive p and negative n examples in a region. The probability $p(y \,|\, \boldsymbol{x})$, $y \in \{0, 1\}$, obtained from the counts a region is

$$p(y \,|\, \boldsymbol{x}) = \frac{p + q \cdot t(y)}{p + n + q} \,,$$

which corresponds to a Bayesian learning approach with Dirichlet priors with parameters $\{t(y) \,|\, y \in \{0, 1\}\}$. The problem of learning the lower and upper probabilities (Antonucci and Cuzzolin 2010) can be rewritten as

$$\underline{p}(1 \,|\, \boldsymbol{x}) = \inf_{t(1)} \frac{p + q \cdot t(1)}{p + n + q} \text{ and } \overline{p}(1 \,|\, \boldsymbol{x}) = \sup_{t(1)} \frac{p + q \cdot t(1)}{p + n + q} \,,$$

where

$$\frac{\eta}{2} \leq t(y) \leq 1 - \frac{\eta}{2} \,, y \in \{0, 1\} \,, \sum_{y \in \{0,1\}} t(y) = 1 \text{ and } \eta \in (0, 1) \,.$$

Thus,

[11] For an implementation in Python, see https://docs.scipy.org/doc/scipy-0.19.1/reference/generated/scipy.optimize.minimize_scalar.html

$$\underline{p}(1 \mid \boldsymbol{x}) = \frac{p + q \cdot \frac{\eta}{2}}{p + n + q} \text{ and } \overline{p}(1 \mid \boldsymbol{x}) = \frac{p + q \cdot (1 - \frac{\eta}{2})}{p + n + q}.$$

To this end, we set $q = 1$ and $\eta = 0.001$ and refer to (Bernard 2005) for an intensive discussion on choosing these parameters.

Appendix 1.2: Logistic regression

In the following, we present a practical procedure for determining the degree of support (19) for the positive class, and then summarize the results for the negative class (which can be determined in a similar manner). Associating each hypothesis $h \in \mathcal{H}$ with a vector $\theta \in \mathbb{R}^{d+1}$, the degree of support (19) can be rewritten as follows:

$$\pi(1 \mid \boldsymbol{x}) = \sup_{\theta \in \mathbb{R}^{d+1}} \min \left[\pi(\theta), 2h(\boldsymbol{x}) - 1 \right]. \tag{30}$$

It is easy to see that the target function to be maximized in (30) is not necessarily concave. Therefore, we propose the following approach.

Let us first note that whenever $h(\boldsymbol{x}) < 0.5$, we have $2h(\boldsymbol{x}) - 1 \leq 0$ and $\min \left[\pi_{\mathcal{H}}(h), 2h(\boldsymbol{x}) - 1 \right] \leq 0$. Thus the optimal value of the target function (19) can only be achieved for some hypotheses h such that $h(\boldsymbol{x}) \in [0.5, 1]$. For a given value $\alpha \in [0.5, 1]$, the set of hypotheses h such that $h(\boldsymbol{x}) = \alpha$ corresponds to the convex set

$$\theta^{\alpha} = \left\{ \theta \mid \theta_0 + \sum_{i=1}^{d} \theta_i x^i = \ln \left(\frac{\alpha}{1 - \alpha} \right) \right\}. \tag{31}$$

The optimal value $\pi^*_{\alpha}(1 \mid \boldsymbol{x})$ that can be achieved within the region (31) can be determined as follows:

$$\pi^*_{\alpha}(1 \mid \boldsymbol{x}) = \sup_{\theta \in \theta^{\alpha}} \min \left[\pi(\theta), 2\alpha - 1 \right] = \min \left[\sup_{\theta \in \theta^{\alpha}} \pi(\theta), 2\alpha - 1 \right]. \tag{32}$$

Thus, to find this value, we maximize the concave log-likelihood over a convex set:

$$\theta^*_{\alpha} = \arg \sup_{\theta \in \theta^{\alpha}} l(\theta). \tag{33}$$

As the log-likelihood function (26) is concave and has second-order derivatives, we tackle the problem with a Newton-CG algorithm (Nocedal and Wright 2006). Furthermore, the optimization problem (33) can be solved using sequential least squares programming[12] (Philip and Elizabeth 2010). Since regions defined in (31) are parallel hyperplanes, the solution of the optimization problem (19) can then be obatined by solving the following problem:

$$\sup_{\alpha \in [0.5,1)} \pi^*_{\alpha}(1 | \boldsymbol{x}) = \sup_{\alpha \in [0.5,1)} \min \left[\pi(\theta^*_{\alpha}), 2\alpha - 1 \right]. \tag{34}$$

[12] For an implementation in Python, see https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

Following a similar procedure, we can estimate the degree of support for the negative class (20) as follows:

$$\sup_{\alpha\in(0,0.5]} \pi^*_\alpha(0|\boldsymbol{x}) = \sup_{\alpha\in(0,0.5]} \min\left[\pi(\theta^*_\alpha), 1-2\alpha\right]. \tag{35}$$

Note that limit cases $\alpha = 1$ and $\alpha = 0$ cannot be solved, since the region (31) is then not well-defined (as $\ln(\infty)$ and $\ln(0)$ do not exist). For the purpose of practical implementation, we handle (34) by discretizing the interval over α. That is, we optimize the target function for a given number of values $\alpha \in [0.5, 1)$ and consider the solution corresponding to the α with the highest optimal value of the target function $\pi^*_\alpha(1\,|\,\boldsymbol{x})$ as the maximum estimator. Similarly, (35) can be handled over the domain (0, 0.5]. In practice, we evaluate (34) and (35) on uniform discretizations of cardinality 50 of [0.5, 1) and (0, 0.5], respectively. We can further increase efficiency by avoiding computations for values of α for which we know that $2\alpha - 1$ and $1 - 2\alpha$ are lower than the current highest support value given to class 1 and 0, respectively. See Algorithm 3 for a pseudo-code description of the whole procedure.

Algorithm 3: Degrees of support for logistic regression

Input: Q, $\mathbf{D}$, θ^{ml}, $\boldsymbol{x}$- initial pool, training data, classifier, unlabelled instance
Output: $\pi(1\,|\,\boldsymbol{x})$, $\pi(0\,|\,\boldsymbol{x})$ - degrees of support

initialize subsets Q_p, Q_n of cardinality Q;
$\pi(1\,|\,\boldsymbol{x}) = \max(2h^{ml}(\boldsymbol{x}) - 1, 0)$, $\pi(0\,|\,\boldsymbol{x}) = \max(1 - 2h^{ml}(\boldsymbol{x}), 0)$;
for $q = 1, \ldots, Q$ **do**
    $\alpha_p = \max(Q_p)$; $\alpha_n = \min(Q_n)$;
    if $2\alpha_p - 1 > \pi(1\,|\,\boldsymbol{x})$ **then**
        solve (33) for $\boldsymbol{x}$, α_p and return θ;
        $\pi(1\,|\,\boldsymbol{x}) = \max(\pi(1\,|\,\boldsymbol{x}), \min(\pi_{\mathcal{H}}(\theta), 2\alpha_p - 1))$;
    if $1 - 2\alpha_n > \pi(0\,|\,\boldsymbol{x})$ **then**
        solve (33) for $\boldsymbol{x}$, α_n and return θ;
        $\pi(0\,|\,\boldsymbol{x}) = \max(\pi(0\,|\,\boldsymbol{x}), \min(\pi_{\mathcal{H}}(\theta), 1 - 2\alpha_p))$;
    $Q_p = Q_p \setminus \{\alpha_p\}$, $Q_n = Q_n \setminus \{\alpha_n\}$;
Return $\pi(1\,|\,\boldsymbol{x})$, $\pi(0\,|\,\boldsymbol{x})$;

Appendix 2: Additional experiments

In addition to the experiments which presented in Sect. 5.4, similar results for the other data sets are presented in Figs. 11, 12 and 13.

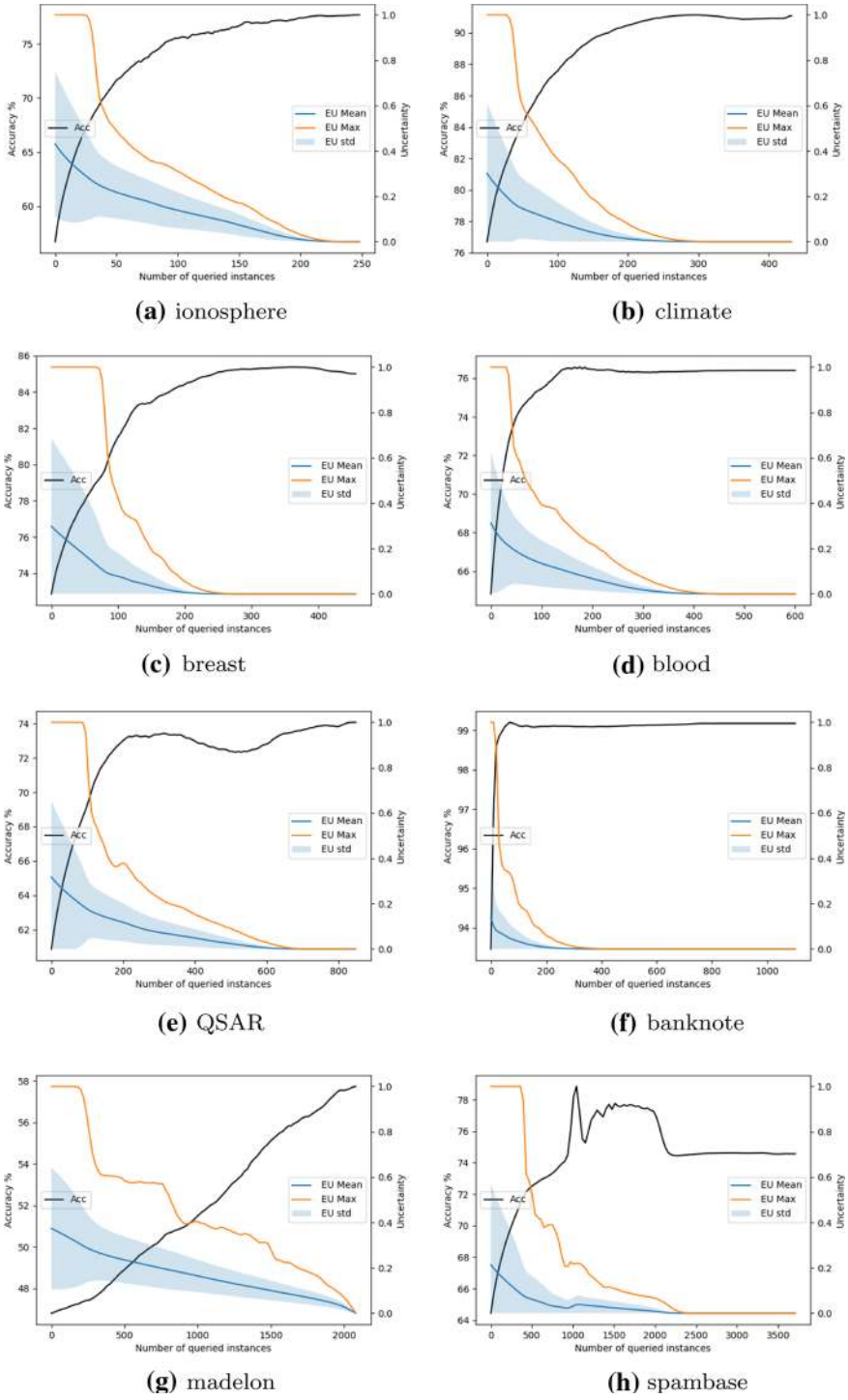


Fig. 11 Average accuracies and degrees of epistemic uncertainty (y-axis) for Parzen window classifiers as functions of the number of instances queried from the pool (x-axis)

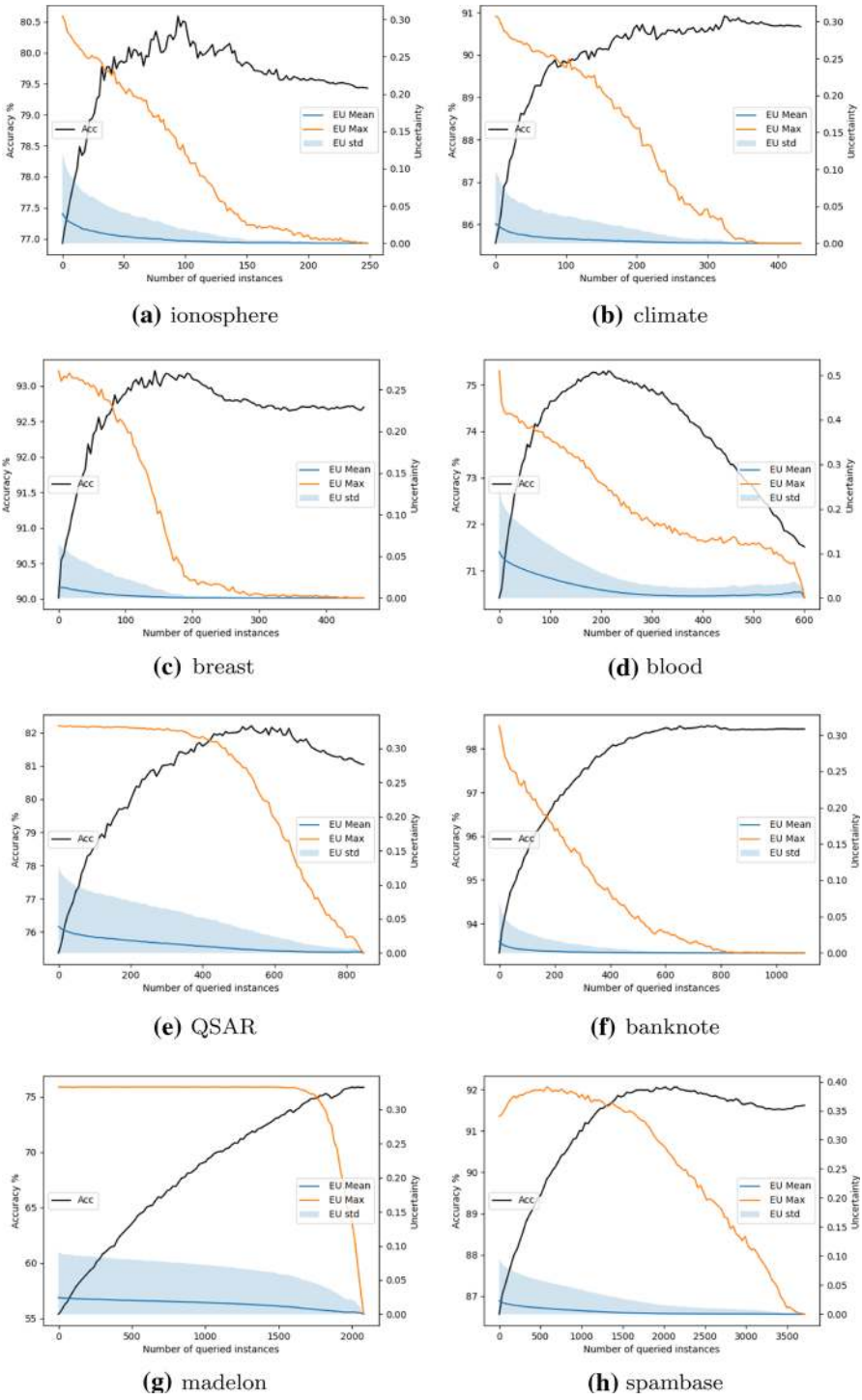


Fig. 12 Average accuracies and degrees of epistemic uncertainty (y-axis) for decision tree as functions of the number of instances queried from the pool (x-axis)

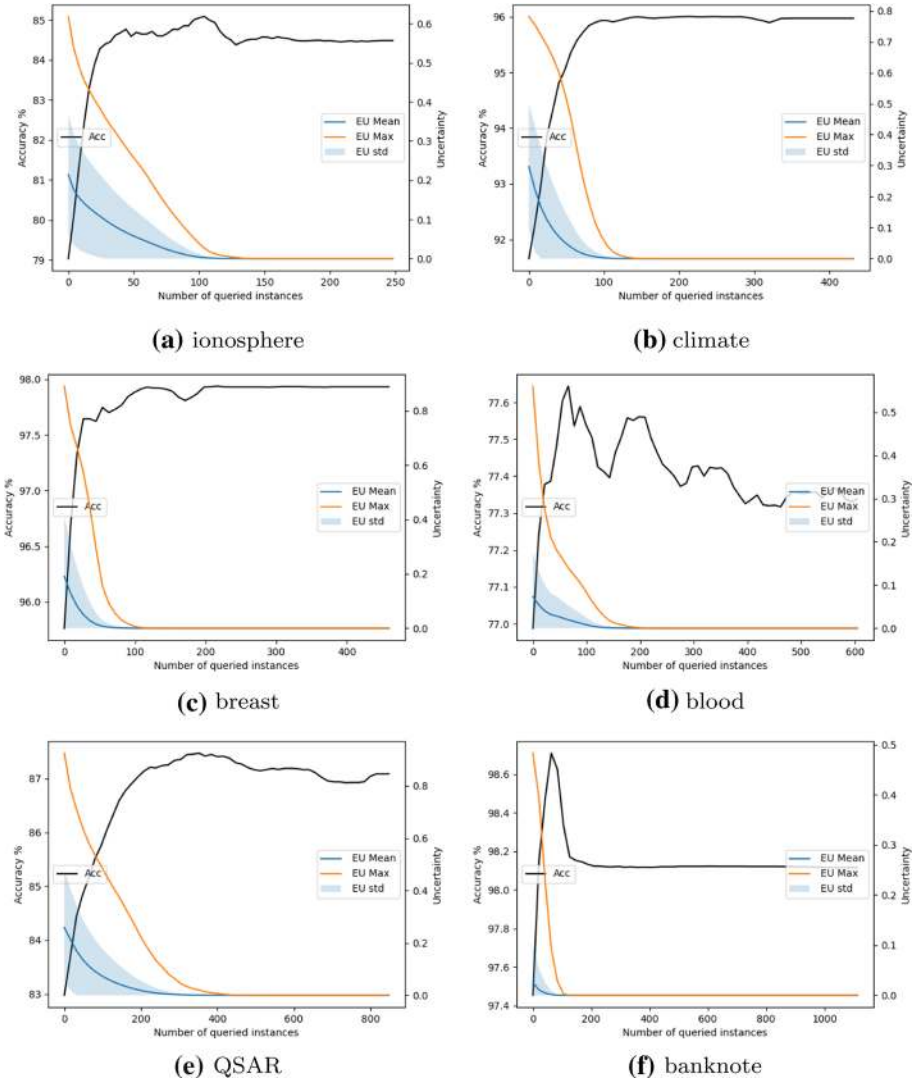


Fig. 13 Average accuracies and degrees of epistemic uncertainty (y-axis) for logistic regression as functions of the number of instances queried from the pool (x-axis)

Acknowledgements The authors gratefully acknowledge financial supported by the German Research Foundation (DFG) under grant number 400845550 as well as the Federal Ministry for Education and Research (BMBF). Valuable technical support was provided by the Paderborn Center for Parallel Computing (PC^2).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Antonucci, A., & Cuzzolin, F . (2010) . Credal sets approximation by lower probabilities: Application to credal networks. In: Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Springer, pp. 716–725.

Antonucci, A., Corani, G., & Gabaglio, S .(2012) . Active learning by the naive credal classifier. In: Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM), pp. 3–10.

Bernard, J. M. (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning, 39*(2–3), 123–150.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association, 57*(298), 269–306.

Bottou, L., & Vapnik, V. (1992). Local learning algorithms. *Neural Computation, 4*(6), 888–900.

Chapelle, O. (2005). Active learning for Parzen window classifier. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS), 5,* 49–56.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

De Campos, L. M., Huete, J. F., & Moral, S. (1994). Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2*(02), 167–196.

Depeweg, S., Hernandez-Lobato, J., Doshi-Velez, F., & Udluft, S .(2018) . Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML), vol 3, pp. 1920–1934.

Freytag, A., Rodner, E., & Denzler, J. (2014). Selecting influential examples: Active learning with expected model output changes. In: Proceedings of the 13th European Conference on Computer Vision (ECCV), Springer, pp. 562–577.

Fu, Y., Zhu, X., & Li, B .(2013) . A survey on instance selection for active learning. Knowledge and Information Systems, pp. 1–35.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer, 27*(2), 83–85.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering and System Safety, 54*(2–3), 217–223.

Hüllermeier, E .(2003) .Inducing fuzzy concepts through extended version space learning. In: Bilgic T, De Baets B, Kaynak O (eds) Proc. IFSA–03, 10th International Fuzzy Systems Association World Congress, Springer, Istambul, no. 2715 in LNAI, pp. 677–648.

Hüllermeier, E., & Waegeman, W. (2021) .Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. Machine Learning (arXiv preprint 191009457)

Kendall, A., & Gal, Y . (2017) .What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS).

Krempl, G., Kottke, D., & Lemaire, V. (2015). Optimised probabilistic active learning (OPAL). *Machine Learning, 100*(2), 449–476.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research, 32*(3), 679–693.

Lewis, D. D., & Gale, W. A .(1994) . A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International SIGIR Conference on Research and Development in Information Retrieval, Springer, pp. 3–12.

Li, M., & Sethi, I. K. (2006). Confidence-based active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(8), 1251–1261.

Lindenbaum, M., Markovitch, S., & Rusakov, D. (2004). Selective sampling for nearest neighbor classifiers. *Machine learning, 54*(2), 125–152.

McCallum, A., & Nigam, K .(1998) . Employing EM and pool-based active learning for text classification. In: Proceedings of the Fifteenth International Conference on Machine Learning (ICML), pp. 350–358.

Menard, S .(2002) .Applied logistic regression analysis. Sage Publishing.

Mitchell, TM .(1977) . Version spaces: A candidate elimination approach to rule learning. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI), pp. 305–310.

Nguyen, VL., Destercke, S., H & üllermeier, E .(2019) . Epistemic uncertainty sampling. In: Proceedings of the 22nd International Conference on Discovery Science (DS), Springer, pp. 72–86.
Nocedal, J., & Wright, S. (2006). *Numerical optimization*. New York: Springer.
Philip, E., & Elizabeth, W .(2010) Sequential . Quadratic programming methods. UCSD Department of Mathematics Technical Report NA-10-03.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.
Rennie, J. D. (2005). Regularized Logistic Regression is Strictly Convex. Technical report, MIT.
Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660–674.
Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences, 255,* 16–29.
Settles, B .(2009) . Active learning literature survey. Technical Report, University of Wisconsin, Madison (TR1648).
Settles, B., & Craven, M .(2008) . An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 1070–1079.
Seung, HS., Opper, M., & Sompolinsky, H .(1992) .Query by committee. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT), pp. 287–294.
Shaker, MH., & Hüllermeier, E .(2020) . Aleatoric and epistemic uncertainty with random forests. In: Proceedings of the Eighteenth International Symposium on Intelligent Data Analysis (IDA), Springer, pp. 444–456.
Sharma, M., & Bilgic, M .(2013) . Most-surely vs. least-surely uncertain. In: Proceedings of the IEEE 13th International Conference on Data Mining (ICDM), IEEE, pp. 667–676.
Sharma, M., & Bilgic, M. (2017). Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery, 31*(1), 164–202.
Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.
Walley, P., & Moral, S. (1999). Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(4), 831–847.
Zaffalon, M. (2002). The naive credal classifier. *Journal of Statistical Planning and Inference, 105*(1), 5–21.
Zhu, J., Wang, H., Hovy, E., & Ma, M. (2010). Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing, 6*(3), 1–24.