Reliable Classification: Learning Classifiers that Distinguish Aleatoric and Epistemic Uncertainty

Robin Senge [a], Stefan Bösner [b], Krzysztof Dembczynski [c], Jörg Haasenritter [b], Oliver Hirsch [b], Norbert Donner-Banzhoff [b], Eyke Hüllermeier [a,*]

[a] Department of Mathematics and Computer Science
[b] Department of Family Medicine
University of Marburg, Germany

[c] Institute of Computing Science
Poznań University of Technology, Poland

* Corrsponding author: eyke@mathematik.uni-marburg.de

Abstract

A proper representation of the uncertainty involved in a prediction is an important prerequisite for the acceptance of machine learning and decision support technology in safety-critical application domains such as medical diagnosis. Despite the existence of various probabilistic approaches in these fields, there is arguably no method that is able to distinguish between two very different sources of uncertainty: aleatoric uncertainty, which is due to statistical variability and effects that are inherently random, and epistemic uncertainty which is caused by a lack of knowledge. In this paper, we propose a method for binary classification that does not only produce a prediction of the class of a query instance but also a quantification of the two aforementioned sources of uncertainty. Despite being grounded in probability and statistics, the method is formalized within the framework of fuzzy preference relations. The usefulness and reasonableness of our approach is confirmed on a suitable data set with information about patients suffering from chest pain.

1 Introduction

Intelligent systems play an increasingly important role in the medical domain, where they are typically used for the purpose of decision support. This includes the application of machine learning methods for predictive modeling, that is, the data-driven construction of models that can be used for predictive purposes [22]. As a simple example, imagine a classifier system that predicts a diagnosis based on symptoms and different types of

patient data. Apart from making predictions, the construction of such models may serve other goals, too. In particular, a model is often useful to gain further insight into the dependencies between predictors and the target variable, and thus may hint at hitherto unknown or incompletely known causal relationships.

Learning from data is inseparably connected with uncertainty. This is largely due to the fact that learning, understood as generalizing beyond a finite set of observed data, is necessarily based on a process of *induction*. Inductive inference replaces specific observations by general models of the data-generating process, but these models are always hypothetical and, therefore, afflicted with uncertainty. Indeed, observed data can generally be explained by more than one candidate theory, which means that one can never be sure of the truth of a particular model (and the predictions it implies). Apart from the uncertainty inherent in inductive inference, additional sources of uncertainty may exist, including erroneous data, incorrect model assumptions, or simply random effects.

Needless to say, a trustworthy representation of uncertainty is desirable and should be considered as a key feature of a machine learning method, all the more in safety-critical application domains such as medicine [3, 20, 25, 14]. Traditionally, all sorts of uncertainty in classification, like in data analysis in general, have been modeled in a probabilistic way, and indeed, probability theory has always been perceived as the ultimate tool for uncertainty handling in fields like statistics and machine learning.

Without questioning the probabilistic approach in general, we argue that conventional methods fail to distinguish two inherently different sources of uncertainty, which are often referred to as *aleatoric* and *epistemic* uncertainty [12]. 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. As opposed to this, epistemic (*aka* systematic) uncertainty refers to uncertainty caused by a lack of knowledge, i.e., it refers to the epistemic state of the decision maker.

The prototypical example of aleatoric uncertainty is coin flipping: The data-generating process in this type of experiment has a stochastic component that cannot be reduced by whatsoever additional information. Consequently, even the best model of this process will only be able to provide probabilities for the two possible outcomes, heads and tails, but no definite answer. Epistemic uncertainty, on the other hand, can in principle be reduced on the basis of additional information. For example, as long as nothing relevant is known about a patient, a medical doctor will be completely *ignorant* about the true diagnosis. Gathering more and more information in the form of medical tests etc., this ignorance will disappear step by step.

In other words, epistemic uncertainty refers to the *reducible* part of the (total) uncertainty, whereas aleatoric uncertainty refers to the *non-reducible* part. From a knowledge representation and decision making point of view, a distinction between these two sources of

uncertainty is arguably important, especially in cases where the ultimate decision can be delayed. A medical doctor, for example, who knows that his uncertainty about the illness of a patient is caused by a lack of knowledge about the disease in question, may decide to consult the literature or ask a colleague before making a decision.

In this paper, we introduce a new approach to reliable classification, in which the aforementioned sources of uncertainty are carefully distinguished. Moreover, we illustrate the usefulness of this approach in the context of medical decision making. Before presenting details of our method in Section 4, we elaborate on the important role of model assumptions and background knowledge in learning from data (Section 2) and propose a formalization of the classification problem within the framework of fuzzy preference relations (Section 3). Section 5 is devoted to a case study, in which our approach is applied to a medical data set with information about patients suffering from chest pain. Additional experiments with benchmark data are presented in Section 6, before concluding the paper in Section 7.

2 Knowledge and Data

The problem we are tackling is to quantify aleatoric and epistemic uncertainty in the context of learning from data. In this context, it is natural to assume that epistemic uncertainty will strongly depend on the amount of data seen so far: the larger the number of observations, the less ignorant we will be when having to make a new prediction. Although this is true in general, it is important to realize that the data is only one source of information. Another important source of information is the background knowledge about the dependency to be learned. In statistics and machine learning, this background knowledge is represented in terms of *model assumptions*, that is, through the specification of the underlying hypothesis (model) space. This specification always comes with an *inductive bias*, which is indeed essential for learning from data. In fact, without any bias, learning would be impossible [18].

Both aleatoric and epistemic uncertainty (ignorance) depend on the way in which background knowledge and data interact with each other. Roughly speaking, the stronger the background knowledge, the less data is needed to resolve ignorance. In the extreme case, the true model is already known, and data is completely superfluous. Normally, however, background knowledge is specified by assuming a certain type of model, for example a linear relationship. Then, all else (namely the data) being equal, the degree of ignorance (epistemic uncertainty) depends on how flexible the corresponding model class is. Informally speaking, the more restrictive the model assumptions are, the smaller the level of ignorance will be.

This is illustrated in Figure 1, where a class prediction is requested for the point marked by a cross. Assuming that the two classes, positive (black) and negative (white), can be

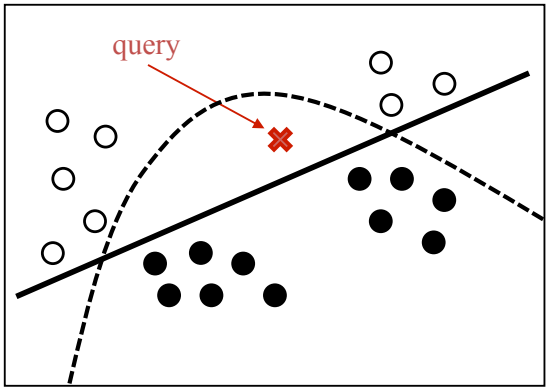


Figure 1: All linear models consistent with the data, like the one shown as a solid line, will predict the query instance as negative (white). A richer model space, however, may include both models voting for positive (like the one shown as dashed line) as well as for negative.

separated by a linear decision boundary, the case is quite clear: All consistent models, i.e., models correctly classifying the training data (like the one shown as a solid line), will predict the negative class. However, being less sure about the shape of the decision boundary and, therefore, expanding the model space by allowing also non-linear (e.g. quadratic) discriminant functions, the level of ignorance increases. In fact, under this assumption, the class of consistent models will not vote unanimously: there are models predicting the positive (like the one shown as a dashed line) as well as models predicting the negative class (like the linear model).

At this point, two further remarks are indicated. First, the background knowledge is normally not limited to assumptions about the shape of the decision boundary, as might be suggested by the previous example. Instead, it will also comprise other assumptions about the data-generating process, for example assumptions about the statistical distribution of error terms. Second, our approach takes the correctness of the background knowledge for granted. In other words, our predictions are conditioned on the underlying model assumptions. Informally, the question we seek to answer can thus be summarized as follows: Looking at the data from a point of view that is biased by our model assumptions, what can we reliably say about the class of the query instance under consideration?

3 Classification as Fuzzy Preference Modeling

Classification can be seen as a decision making problem: Given a new query instance to be classified, a decision in favor of one class has to made against the background of the information at hand (which, as discussed above, essentially consists of data and background

knowledge). Indeed, the classification problem is often formalized within the framework of Bayesian decision theory [4].

For the reasons explained in the following, we seek to connect the classification problem to another type of decision-theoretic framework, namely *fuzzy preference modeling* [10]. From the discussion so far, it is clear that aleatoric and epistemic uncertainty should no be considered as *bivalent* notions: Normally, it does not make sense to categorize a situation as either uncertain or not uncertain, or a decision maker as being either ignorant or not ignorant. Instead, it makes sense to say that a situation is uncertain to some degree, or that a decision maker is ignorant to some extent. For example, given perfect knowledge about the probability of a binary event, the aleatoric uncertainty of the decision (prediction) should be highest for the probability $1/2$ and become smaller for probabilities close to 0 or close to 1. Likewise, the degree of ignorance will depend on the abundance of data, but will not suddenly disappear by adding a single observation.

Our goal, therefore, is to model aleatoric and epistemic uncertainty as *gradual* notions. As mentioned above, we therefore establish a connection to the field of fuzzy logic or, more specifically, fuzzy preference modeling. In fuzzy preference modeling, the point of departure is a *preference structure* $(\mathcal{P}, \mathcal{I}, \mathcal{J})$, consisting of the following binary relations on the set $\mathcal{A}$ of decision alternatives:

- a strict preference relation $\mathcal{P} : \mathcal{A} \times \mathcal{A} \to [0, 1]$,
- an indifference relation $\mathcal{I} : \mathcal{A} \times \mathcal{A} \to [0, 1]$,
- an incomparability relation $\mathcal{J} : \mathcal{A} \times \mathcal{A} \to [0, 1]$.

These relations are "valued" or "fuzzy" in the sense of assuming values in $[0, 1]$ instead of $\{0, 1\}$. That is, for two decision alternatives $A, A' \in \mathcal{A}$, the former can be preferred to the latter to a certain degree $\mathcal{P}(A, A') \in [0, 1]$. Likewise, indifference $\mathcal{I}(A, A')$ and incomparability $\mathcal{J}(A, A')$ are both a matter of degree. Nevertheless, the relations are supposed to satisfy certain properties; for example, $\mathcal{I}$ and $\mathcal{J}$ should be symmetric. Besides, for all $A, A' \in \mathcal{A}$,

$$\mathcal{P}(A, A') + \mathcal{P}(A', A) + \mathcal{I}(A, A') + \mathcal{J}(A, A') = 1 \tag{1}$$

The indifference and incomparability relation are especially interesting for us, since a direct correspondence can be established between the concepts of indifference and aleatoric uncertainty on the one side and incomparability and epistemic uncertainty on the other side. In fact, the the sum $\mathcal{I}(A, A') + \mathcal{J}(A, A')$, which corresponds to the symmetric part of (1), can be seen as the degree to which A and A' cannot be distinguished, albeit for different reasons: $\mathcal{I}(A, A')$ is the degree to which A and A' are inherently equal, whereas

$\mathcal{J}(A, A')$ is the degree to which they (yet) cannot be compared. This fits quite nicely with our view of aleatoric uncertainty as the irreducible and epistemic uncertainty as the reducible part of the uncertainty.

In Section 4, we shall propose a formal approach to classification in which predictions are represented in terms of fuzzy preferences, with indifference re-interpreted as aleatoric uncertainty and incomparability re-interpreted as epistemic uncertainty. Thus, given two classes, positive ($\oplus$) and negative ($\ominus$), a prediction will be given in the form of a 4-tuple

$$\big(p_\oplus,\, p_\ominus,\, u_a,\, u_e\big) = \Big(\mathcal{P}(\oplus,\ominus),\, \mathcal{P}(\ominus,\oplus),\, \mathcal{I}(\oplus,\ominus),\, \mathcal{J}(\oplus,\ominus)\Big), \tag{2}$$

where $p_\oplus$ is interpreted as the strict preference in favor of predicting the positive class, $p_\ominus$ the strict preference in favor of the negative class, u_a as the degree of aleatoric uncertainty, and u_e as the degree of epistemic uncertainty.

Thus, we restrict ourselves to the case of two classes (alternatives) in this paper. Roughly speaking, our approach allows for producing a single entry in the relations defining a fuzzy preference structure. An obvious extension to the multi-class case is to apply the method for each pair of classes. However, this extension as well as the question of how to exploit a preference structure $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ for different decision making purposes are beyond the scope of this paper.

4 A Formal Approach

Our point of departure is a binary classification problem with classes $\mathcal{Y} = \{\oplus, \ominus\}$ and instance space $\mathcal{X}$. Besides, we assume to be given a class $\mathfrak{M}$ of candidate models, where each $M \in \mathfrak{M}$ is a probabilistic classifier. Thus, a model M is an $\mathcal{X} \to [0, 1]$ mapping such that, for each instance $\boldsymbol{x} \in \mathcal{X}$, the value $M(\boldsymbol{x})$ is the probability that the class of $\boldsymbol{x}$ is positive.

Suppose to be given a set of training data (observations) $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$ (normally assumed to be independent and identically distributed) and a new query instance $\boldsymbol{x}_0$ for which a prediction is sought. The method we propose is inductive in the sense that we generally learn global models from the data, but transductive (and "lazy") in the sense that the main inference step is specifically tailored for the query instance $\boldsymbol{x}_0$.

4.1 Bayesian Inference

Adopting a Bayesian perspective, we can consider the posterior distribution on the model space:

$$\mathbf{P}(M \,|\, \mathcal{D}) \;\propto\; \mathbf{P}(M) \cdot \mathbf{P}(D \,|\, M),$$

where $\mathbf{P}(M)$ is the prior probability of the model $M \in \mathfrak{M}$, $\mathbf{P}(D \,|\, M)$ is the probability of the data given M, and $\mathbf{P}(M \,|\, \mathcal{D})$ is the posterior probability of M. Often, the specification of a suitable prior distribution is difficult, and hence the likelihood function

$$L(M) \;=\; \mathbf{P}(D \,|\, M)$$

is considered instead. This is essentially equivalent to using a uniform prior and leads to the posterior $\mathbf{P}(M \,|\, \mathcal{D}) \propto L(M)$.

In proper Bayesian inference, a prediction for $\boldsymbol{x}_0$ is obtained through *model averaging*:

$$\mathbf{P}(\oplus \,|\, \boldsymbol{x}_0) = \int_{\mathfrak{M}} M(\boldsymbol{x}_0)\, d\, \mathbf{P}(M \,|\, \mathcal{D}) \tag{3}$$

Thus, the predicted probability of the positive class is the *expected* prediction $M(\boldsymbol{x}_0)$, where the expectation over the model is taken with respect to the posterior $\mathbf{P}(M \,|\, \mathcal{D})$. Correspondingly, the probability of the negative class is given by

$$\mathbf{P}(\ominus \,|\, \boldsymbol{x}_0) = 1 - \mathbf{P}(\oplus \,|\, \boldsymbol{x}_0). \tag{4}$$

In this type of inference, aleatoric and epistemic uncertainty is not distinguished or, more specifically, epistemic uncertainty is "averaged out". Consider again coin flipping as an example, and let the model class be given by $\mathfrak{M} = \{M_\alpha \,|\, 0 \leq \alpha \leq 1\}$, where M_α is modeling a biased coin landing heads ($\oplus$) with a probability of α and tails ($\ominus$) with a probability of $1 - \alpha$. According to (3), we derive a probability of $1/2$ for $\oplus$ and $\ominus$, regardless of whether the (posterior) distribution on $\mathfrak{M}$ is given by the uniform distribution (all coins are equally probable, i.e., the case of complete ignorance) or the one-point measure assigning probability 1 to $M_{1/2}$ (the coin is known to be fair with complete certainty).

4.2 Plausibility and Necessity of Events

The above problem is related the fact that *ignorance* cannot be represented in terms of a (standard) probability measure. This problem, in turn, is rooted in the normalization constraint (4) or, more generally, the self-duality of a probability measure:

$$\mathbf{P}(A) = 1 - \mathbf{P}(\overline{A}), \tag{5}$$

where A is an event and $\overline{A}$ the complement of A. Thus, assigning a certain probability mass to A (i.e., supporting A) can only be done by removing this mass from $\overline{A}$. Consequently, it is not possible to say that A is plausible without saying that $\overline{A}$ is implausible. Roughly speaking, the probabilistic approach assumes a "constant" amount of knowledge, namely a single unit mass. Although this mass can be distributed to different events, the total

mass remains the same.

To circumvent this limitation, we refer to the more flexible framework of *possibility theory*, in which uncertainty is modeled in terms of *two* measures instead of a single one [9]. These two measures, which are called, respectively, *possibility* (*plausibility*) measure Π and *necessity* measure N, are dual in the sense that

$$\Pi(A) = 1 - N(\overline{A}). \tag{6}$$

Thus, an event A is plausible insofar as the complement of A is not necessary. Or, stated differently, an event A necessarily occurs if the complement of A is not possible. These two measures allow for expressing ignorance in a proper way, mainly because A can be declared plausible without declaring $\overline{A}$ implausible. In particular, $\Pi(A) \equiv 1$ models complete ignorance: Everything is completely plausible, and hence nothing is necessary ($N(A) = 1 - \Pi(\overline{A}) = 0$ for all A).

Comparing (5) and (6), it becomes obvious that a probability measure is playing both roles simultaneously, namely the role of the possibility and the role of the necessity measure. This explains why it is more constrained and hence less expressive.

4.3 From Plausibility to Aleatoric and Epistemic Uncertainty

In our case, the underlying space $\mathcal{Y}$ has only two elements, which are hence complementary to each other. Thus,

$$\begin{aligned}\pi(\oplus) &= \Pi(\{\oplus\}) = 1 - N(\{\ominus\}),\\ \pi(\ominus) &= \Pi(\{\ominus\}) = 1 - N(\{\oplus\}),\end{aligned}$$

where $\pi(\oplus)$ denotes the plausibility of $\oplus$ (given the query $\boldsymbol{x}_0$) and, correspondingly, $\pi(\ominus)$ the plausibility of $\ominus$. Now, suppose the two plausibility degrees to be given, and recall that $0 \leq \pi(\oplus), \pi(\oplus) \leq 1$ without any further constraints. We then define the *degree of epistemic uncertainty* as

$$u_e = \min\big(\pi(\oplus), \pi(\ominus)\big),$$

that is, the degree to which both $\oplus$ and $\ominus$ are plausible (the minimum plays the role of a generalized conjunction [13]). Likewise, we define the *degree of aleatoric uncertainty* as

$$\begin{aligned}u_a &= \min\big(1 - \pi(\oplus), 1 - \pi(\ominus)\big)\\ &= \min\big(N(\{\ominus\}), N(\{\oplus\})\big)\\ &= 1 - \max\big(\pi(\oplus), \pi(\ominus)\big),\end{aligned}$$

that is, the degree to which neither $\oplus$ nor $\ominus$ is plausible. These definitions are completely in agreement with the standard way of deriving the degrees of indifference and ignorance from the degrees of weak preference in fuzzy preference modeling [10].

Note that

$$u_a + u_e = 1 - \max\big(\pi(\oplus), \pi(\ominus)\big) + \min\big(\pi(\oplus), \pi(\ominus)\big) \leq 1.$$

Thus, the total uncertainty (aleatoric + epistemic) is upper-bounded by 1. We assign the difference $1-(u_a+u_e)$ as strict preference to the class with the higher plausibility, thereby satisfying the normalization condition (1).

4.4 Modeling Class Plausibility

An important question has been left open so far, namely how to define the plausibilities $\pi(\oplus)$ and $\pi(\ominus)$ of the positive and negative class, respectively. To this end, we exploit the posterior model distribution (likelihood function), just like in Bayesian inference. However, for the reasons explained above, we avoid averaging.

Instead, we define the plausibility of the positive class as follows:

$$\pi(\oplus) = \sup_{M \in \mathfrak{M}} \min\big(\pi_{\mathfrak{M}}(M), f(M(\boldsymbol{x}_0))\big), \tag{7}$$

where $f(a) = \max(2a-1, 0)$ is the membership function of the fuzzy set of *high probabilities*, and $\pi_{\mathfrak{M}}$ is a possibility distribution on the model space. Recalling that the supremum operator plays the role of a generalized existential quantifier, the expression (7) can be read as follows: The class $\oplus$ is plausible insofar there exists a model M that is plausible and that assigns a high probability to $\oplus$. Analogously, we define the plausibility for $\ominus$:

$$\pi(\ominus) = \sup_{M \in \mathfrak{M}} \min\big(\pi_{\mathfrak{M}}(M), f(1 - M(\boldsymbol{x}_0))\big). \tag{8}$$

The possibility distribution $\pi_{\mathfrak{M}}$ on the model space is essentially a normalized probability or likelihood [23, 24]. In the case of likelihood inference, we define it as

$$\pi_{\mathfrak{M}}(M) = \frac{L(M)}{L(M^{ml})} \ ,$$

where $M^{ml} \in \mathfrak{M}$ is the maximum likelihood (ML) estimation on the data $\mathcal{D}$. Thus, the plausibility of a model is in direct proportion to its likelihood, with the ML estimation having the highest plausibility of 1; for a deeper analysis of the idea of possibility as normalized likelihood, see [8, 7]. Likewise, in the Bayesian case, we can normalize the posterior probabilities:

$$\pi_{\mathfrak{M}}(M) = \frac{\mathbf{P}(M \,|\, \mathcal{D})}{\sup_{M^* \in \mathfrak{M}} \mathbf{P}(M^* \,|\, \mathcal{D})}.$$

Although algorithmic aspects are not in the focus of this paper, it is worth to mention that the computation of (7) (and likewise of (8)) may become rather complex. In fact, the computation of the supremum comes down to solving an optimization problem, the complexity of which strongly depends on the model space $\mathfrak{M}$.

Often, a local search in the model space $\mathfrak{M}$ will be appropriate, at least of $\pi_{\mathfrak{M}}$ is unimodal: Starting the search at $M = M^{ml}$ (or any other model with highest plausibility), we have $\pi_{\mathfrak{M}}(M) = 1$ but probably $f(M(\boldsymbol{x}_0)) < 1$; thus, the minimum will be determined by the second argument, $f(M(\boldsymbol{x}_0))$. In order to increase the minimum, one should move into the direction of models M that assign a higher probability to $\oplus$. This way, $f(M(\boldsymbol{x}_0))$ will become larger while $\pi_{\mathfrak{M}}(M)$ will become smaller. The optimal solution achieves a compromise between model and class plausibility and (in the continuous case) is characterized by the equality $\pi_{\mathfrak{M}}(M) = f(M(\boldsymbol{x}_0))$.

4.5 Summary of the Approach

Summarizing the steps described above, our approach (based on likelihood inference) can be summarized as follows. For a given data set $\mathcal{D}$ and a query instance $\boldsymbol{x}_0$, it computes a quadruple (2) with

$$
\begin{aligned}
\pi(\oplus) &= \sup_{M \in \mathfrak{M}} \min \left(\frac{L(M)}{L(M^{ml})} \,,\, \max\big(2M(\boldsymbol{x}_0) - 1, 0\big) \right), \\
\pi(\ominus) &= \sup_{M \in \mathfrak{M}} \min \left(\frac{L(M)}{L(M^{ml})} \,,\, \max\big(1 - 2M(\boldsymbol{x}_0)), 0\big) \right), \\
u_a &= \min\Big(1 - \pi(\oplus), 1 - \pi(\ominus)\Big), \\
u_e &= \min\Big(\pi(\oplus), \pi(\ominus)\Big), \\
p_\oplus &= \begin{cases} 1 - (u_a + u_e) & \pi(\oplus) > \pi(\ominus) \\ \frac{1-(u_a+u_e)}{2} & \pi(\oplus) = \pi(\ominus) \\ 0 & \pi(\oplus) < \pi(\ominus) \end{cases}, \\
p_\ominus &= 1 - (p_\oplus + u_a + u_e).
\end{aligned}
$$

4.6 Special Cases

The above considerations already reveal that the class plausibilities will strongly depend on the "peakedness" of $\pi_{\mathfrak{M}}$. If this function has a strong peak, suggesting a high certainty about the true model (e.g., because $\mathcal{D}$ consists of many observations), $\pi_{\mathfrak{M}}$ will quickly drop when moving away from this model. Thus, it will not be possible to achieve a value

significantly better than $f(M(\boldsymbol{x}_0))$ in the case of $\oplus$ and of $f(1 - M(\boldsymbol{x}_0))$ in the case of $\ominus$. Consequently, not both plausibility degrees $\pi(\oplus)$ and $\pi(\ominus)$ can be large at the same time, which in turn means that there is a low level of epistemic uncertainty.

As a concrete example, consider the case where a probability distribution

$$(\alpha, 1-\alpha) = (\mathbf{P}(\oplus \,|\, \boldsymbol{x}_0), \mathbf{P}(\ominus \,|\, \boldsymbol{x}_0))$$

is known with certainty; the case of fair coin flipping, for instance, corresponds to $\alpha = 1/2$. Since $\pi_{\mathfrak{M}}(M)$ is 1 for the model prescribing these probabilities and 0 otherwise, (7) and (8) are given, respectively, by $f(\alpha)$ and $f(1-\alpha)$. This implies that either $\pi(\oplus)$ or $\pi(\ominus)$ (or both) are 0. Consequently, there is no epistemic uncertainty ($u_e = 0$), while the aleatoric uncertainty is given by $u_a = \min(2\alpha, 2(1-\alpha))$. Thus, the aleatoric uncertainty is highest (namely 1) for $\alpha = 1/2$ and lowest (namely 0) for $\alpha \in \{0, 1\}$. This is in perfect agreement with our expectations.

The other extreme case corresponds to $\pi_{\mathfrak{M}} \equiv 1$, where all models are considered as equally plausible; this case is obtained, for example, from a flat likelihood and a uniform posterior on $\mathfrak{M}$. Then, the first argument of the minimum in (7) and (8) is always 1, which essentially means that the model can be chosen freely. Consequently, $\pi(\oplus) = \pi(\ominus) = 1$. This corresponds to the case of complete ignorance or, say, full epistemic uncertainty ($u_e = 1$).

The last example also makes clear that our approach can be seen as a generalization of standard *version space learning* [17], in which the class $\mathcal{M}_{cons} \subseteq \mathfrak{M}$ of all consistent models is maintained. Standard version space learning considers a noise-free setting, in which models are binary classifiers ($\mathcal{X} \to \{0, 1\}$ mappings), and a model is consistent if it does not make any mistake on the examples seen so far. This setting can be modeled in terms of $\{0, 1\}$-valued possibility distributions, in which case the supremum in (7) and (8) becomes a real existential quantifier: $\pi(\oplus) = 1$ if there exists a model $M \in \mathcal{M}_{cons}$ that predicts $\oplus$ for $\boldsymbol{x}_0$ and $\pi(\oplus) = 0$ otherwise.

4.7 An Illustration

As a more concrete example, consider a sequence of Bernoulli experiments: A coin with a fixed but unknown probability $p \in [0, 1]$ for landing heads is thrown repeatedly, and after each trial, the problem is to predict the outcome of the next trial. Formally, the model space is given by $\mathfrak{M} = \{M_p \,|\, 0 \leq p \leq 1\}$, where $M_p = M_p(x_0)$ is the model that produces outcome $\oplus$ (heads) with probability p and $\ominus$ (tails) with probability $1 - p$.[1]

Needless to say, the prediction of the next outcome is afflicted with uncertainty. In the beginning, this uncertainty is mainly of epistemic nature, since nothing is known about

[1] Although we do not need an instance space in this example, one may formally define a space $\mathcal{X} = \{x_0\}$ consisting of a single instance which is repeatedly chosen with probability 1 (and can hence be ignored).

the parameter p. In the course of time, however, more and more is learned about this parameter, so the epistemic uncertainty becomes smaller and smaller; in the limit of an infinite sample size, it will vanish completely, since p can be estimated (based on relative frequencies) to an arbitrary degree of precision. The remaining uncertainty, then, is purely aleatoric.

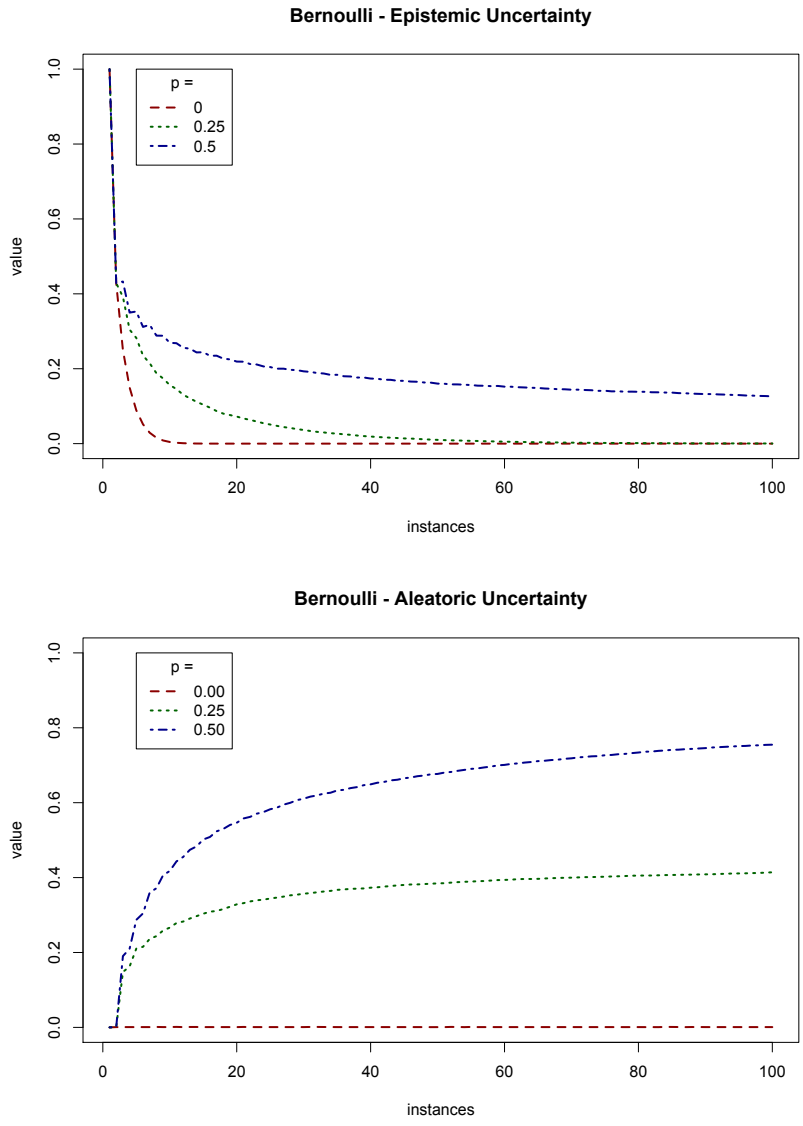


Figure 2: Sequential coin flipping example: Epistemic and aleatoric uncertainty for different values of p as a function of the number of coin flips.

More specifically, after each trial, we applied our approach with the likelihood function

$$L(M_p) = \begin{pmatrix} N \\ K \end{pmatrix} \cdot p^K \cdot (1-p)^{N-K} \ ,$$

where N is the number of trials and K the number of heads so far. Figure 2 plots the expected degree of the two types of uncertainty (approximated by the average over a large number of repetitions) as a function of the number of trials. These results obviously confirm our expectations. Qualitatively, the curves are similar for different values of p. Yet, the closer p is to $1/2$, the slower the epistemic uncertainty disappears, and the higher the level of aleatoric uncertainty that eventually remains. The case $p = 1/2$ is especially noticeable, as it corresponds to "complete uncertainty". Being fully epistemic in the beginning, the uncertainty becomes entirely aleatoric in the limit $N \to \infty$. More importantly, however, the total amount of uncertainty is not reduced (the curve for aleatoric uncertainty slowly converges to 1): Even precise knowledge about p does not help in predicting the outcome of the next trial. Obviously, the situation is different for parameters $p \neq 1/2$, since in this case, even approximate knowledge of p will help to do better than random guessing.

5 Case Study

In this section, we present a series of experimental studies that mainly serve two purposes: First, we seek to show that our approach is able to capture uncertainty in classification in a meaningful way; as will be seen shortly, the data set to be used in the experiments is especially suitable in this regard. Second, we like to highlight again the relevance of our method for the practical problem of medical decision making.

5.1 Data Set

The data set has originally been used to evaluate the diagnostic accuracy of symptoms and signs for coronary heart disease (CHD) in patients presenting with chest pain in primary care (PC). Chest pain is a common complaint in primary care, with CHD being the most concerning of many potential causes. Based on the medical history and physical examination, GPs have to classify patients into two classes: patients in whom an underlying CHD can be safely ruled out and patients in whom chest pain is probable caused by CHD. Needless to say, a wrong decision may have serious consequences. All the more it is important to incorporate measures of reliability into computer systems assisting the GP. Design and conduct of the study were described elsewhere [5].

Briefly, 74 general practitioners (GP) recruited consecutively patients aged $\geq$ 35 who

presented with chest pain as primary or secondary complaint. GPs took a standardized history and performed a physical examination. Patients and GPs were contacted six weeks and six months after the consultation. All relevant information about course of chest pain, diagnostic procedures and treatments had been gathered during six months. An independent expert panel of one cardiologist, one GP and one research staff member reviewed each patient's data and established the reference diagnosis by deciding whether or not CHD was the underlying reason of chest pain. This reference diagnosis hence can be taken as a kind of ground-truth.[2]

The data set is comprised of 1199 (135 CHD and 1064 non-CHD) patients described by six binary attributes:

- patient assumes pain is of cardiac origin
- pain not reproducible by palpation
- known clinical vascular disease
- age gender compound
- pain depends on exercise
- reference diagnosis

The first five attributes have been found to be the most predictive ones among all those collected in the full survey [6]. For our purpose, the data set is especially interesting because, in addition to the patient information and the classification, it contains a subjective value of the GP's uncertainty about the classification: The GP was asked to assign a probability for CHD being the underlying reason. Thus, values close to zero or close to one indicate a high confidence of the GP, whereas values close to 0.5 indicate a high level of uncertainty.

5.2 Classifiers

Having in mind our discussion about the influence of background knowledge (Section 2), we decided to make use of two classifiers. The first one is standard logistic regression (LR), which is a quite common approach in medical data analysis in general and was also used for analyzing our data set in [5]. Since this classifier fits a linear decision boundary in the instance space, it is rather restrictive and makes strong model assumptions. In particular, it makes a strong independence assumption: The influence of each attribute (on the probability of CHD as underlying cause $\oplus$, i.e., a CHD case) is independent of the values of all other attributes.

[2] Needless to say, one cannot entirely exclude the possibility of mistakes even for this diagnosis.

The second classifier is more flexible and simply estimates the probability of $\oplus$ for each possible patient (based on relative frequencies). This is possible due to the quite limited size of the instance space $\mathcal{X} = \{0,1\}^5$: Five binary attributes give rise to only 32 possible combinations $\boldsymbol{x} = (x_1, \ldots, x_5)$. Since this classifier simply tabulates the estimated probabilities $\mathbf{P}(\oplus \,|\, \boldsymbol{x})$ for each $\boldsymbol{x} \in \mathcal{X}$, we call is "table classifier" (TC).

Prior to elaborating on the representation of uncertainty, which is our main target, we compared the predictive accuracy of the original classifiers with our "reliable variants". The reason is to make sure that an improved representation of uncertainty in classification does not come at the price of a drop in predictive accuracy.

Given a set of training data and a new query instance, our approach solves the optimization problems (7-8) defining the plausibility degrees for the positive and the negative class using the well-known CMAES method [11]. Based on these degrees, both types of uncertainty (aleatoric and epistemic) are calculated, as well as the degrees of strict preference. A prediction is then made on the basis of the latter, that is, in favor of the class with the higher degree of strict preference. In terms of computational complexity, this approach is quite efficient: the time needed to make a single prediction is 1.52 ± 0.14 seconds.

LR	Reliable LR	TC	Reliable TC
0.921 ± 0.025	0.921 ± 0.025	0.916 ± 0.034	0.915 ± 0.031

Table 1: Mean classification rate $\pm$ standard deviation for different classifiers.

The results of a 10-fold cross validation (repeated 5 times and averaged) are summarized in Table 1 and Figure 3. As can be seen, there are essentially no differences between the original classifiers and their respective reliable variants. In other words, extracting different types of uncertainty from a classifier and basing a prediction on what remains, namely the strict preferences, does apparently not harm the predictive accuracy.

5.3 Evolution of Aleatoric and Epistemic Uncertainty

In our model, epistemic uncertainty is expected to decrease with an increasing sample size, which can be seen as a measure of the GP's "experience" (the number of patients seen so far). As opposed to this, aleatoric uncertainty will normally not vanish completely, unless a perfect classifier with zero prediction error can be learned. This, however, is only possible in a noise-free setting (and provided the model assumptions are correct).

In order to verify these expectations, we generated a series of training data sets of increasing size. For each sample size, the classifiers were trained and used to predict the two types of uncertainty (aleatoric and epistemic) for each instance in the remaining test data (i.e., the data not used for training so far).

Figures 4 and 5 illustrate the results in terms of the average level of epistemic and aleatoric

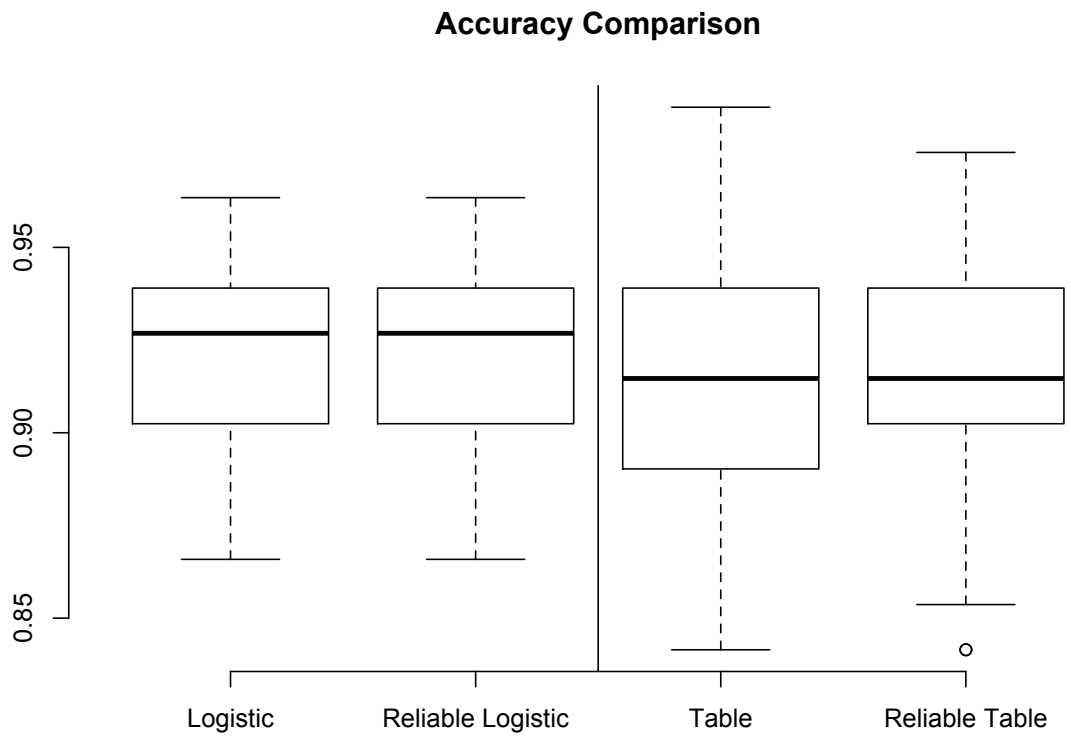


Figure 3: Accuracy (classification rate) of the original classifiers compared to the respective reliable variant.

uncertainty as a function of sample size. As expected, epistemic uncertainty decreases with an increasing number of patients, while aleatoric uncertainty slightly increases in the beginning but then remains at a level of about 0.1. Interestingly, but in complete agreement with our expectation, the level of epistemic uncertainty is much higher for the table classifier than for logistic regression, reflecting the fact that the latter starts with more background knowledge than the former.

5.4 Comparing Predicted and Reported Uncertainty

In another experiment, we analyzed to what extent the uncertainty "predicted" by our model is in agreement with the GP's level of uncertainty in a classification. In fact, a positive dependency between the model's uncertainty and the GP's uncertainty would be a strong indicator of the fact that the former captures uncertainty in a reasonable and realistic way. It would also be a prerequisite for using our model in order to select uncertain cases in an automatic way.

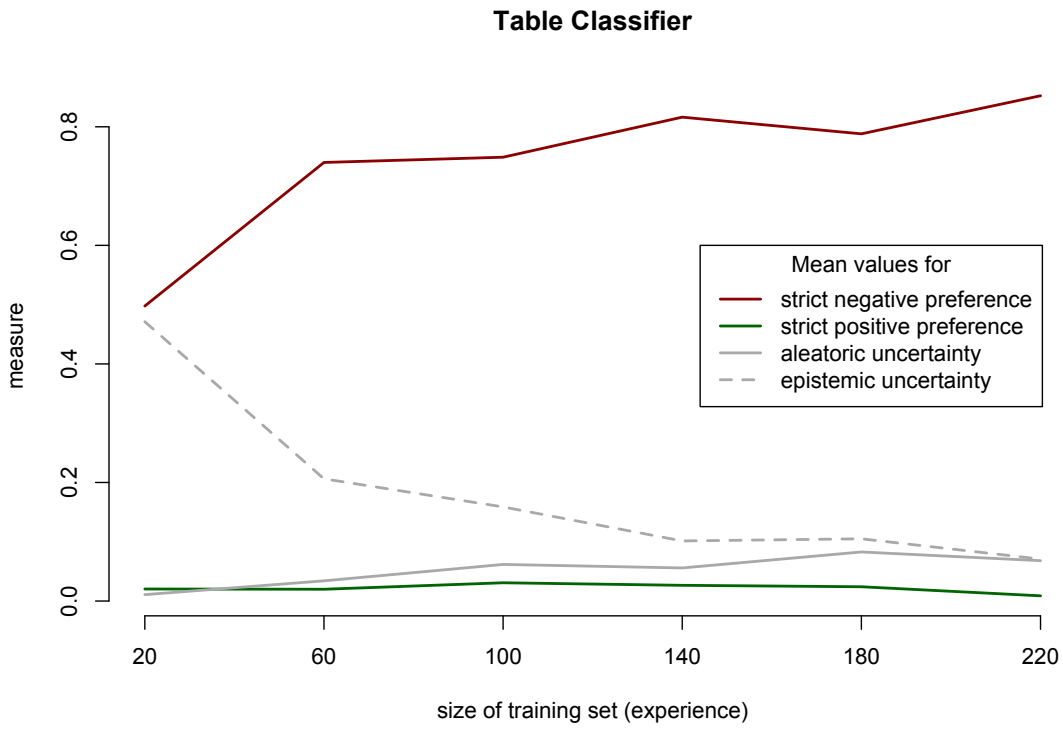


Figure 4: Aleatoric and epistemic uncertainty for the table classifier as a function of sample size.

One should note, however, that a strong dependency of that kind should not necessarily be expected, even if our model is quantifying uncertainty in a proper way. First, the model assumptions underlying our classifiers (LR and TC) will probably not fully fit the background knowledge and decision procedure of the GP. Our classifier assumes a model with only five attributes of a patient and is trained on a subsample of the patients in this study. The physicians, on the other hand, are "trained" by their own experience and education. Moreover, they will probably use more information about a patient than our five attributes.

Second, the types of uncertainty are not exactly the same. As already explained earlier, the GPs were not directly asked for their uncertainty about a diagnosis, let alone to differentiate between aleatoric and epistemic uncertainty. Instead, they were asked to provide the probability p_{CHD} that a patient $\boldsymbol{x}_0$ is a CHD case. Therefore, we approximated the GPs subjective uncertainty using the following transformation:

$$U_{GP}(\boldsymbol{x}_0) \;=\; 1 - \frac{|p_{CHD}(\boldsymbol{x}_0) - 0.5|}{0.5}$$

In order to quantify the dependency between the GP's uncertainty and our model's un-

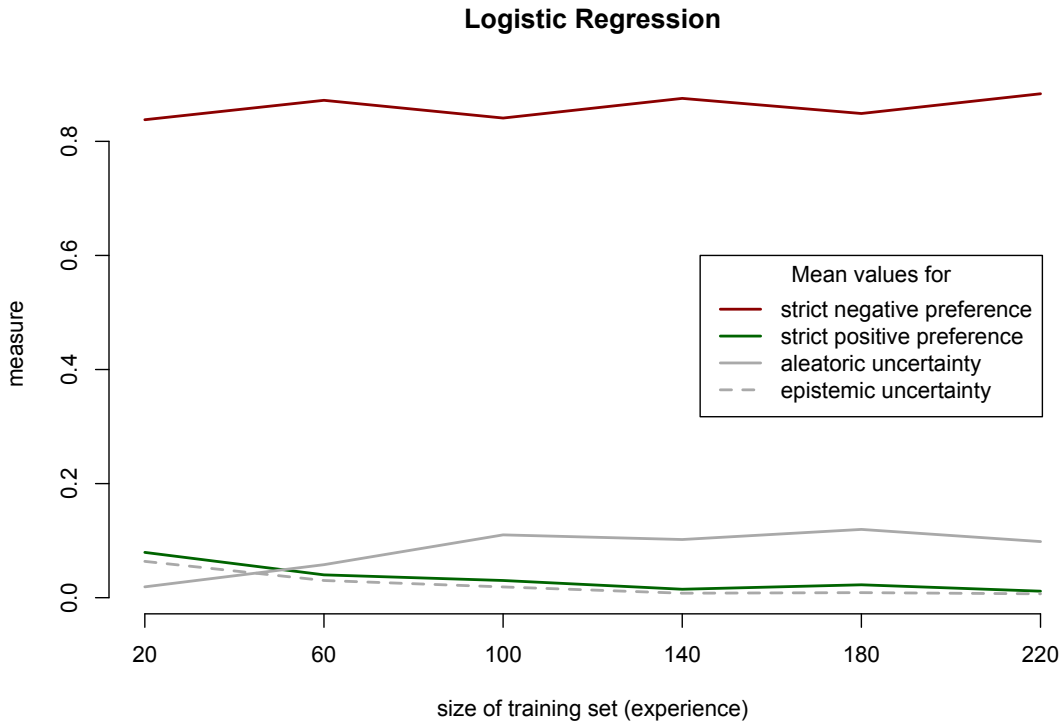


Figure 5: Aleatoric and epistemic uncertainty for logistic regression as a function of sample size.

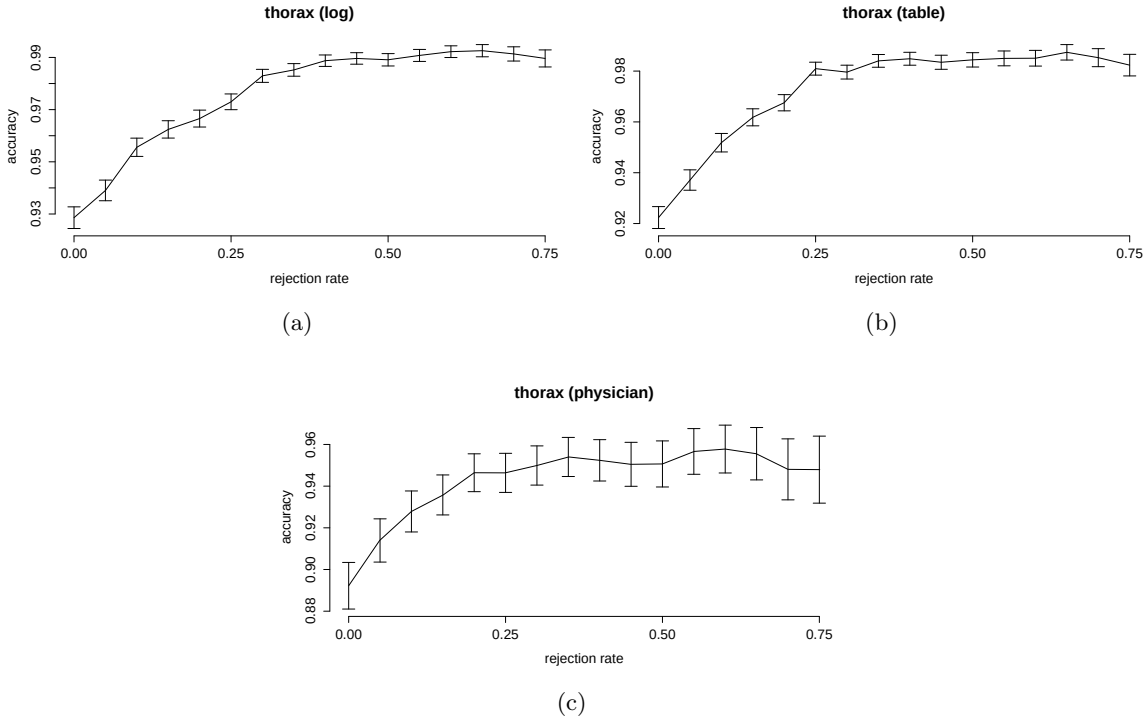
certainty, we calculated the (Spearman) rank correlation between $U_{GP}(\boldsymbol{x}_0)$ and the total uncertainty $u_a(\boldsymbol{x}_0) + u_e(\boldsymbol{x}_0)$, i.e., the sum of the aleatoric and epistemic uncertainty predicted by our model. Although the correlation is not perfect, it is significantly positive, with a value of more than 0.5 for logistic regression and 0.43 for the table classifier. This clearly shows that the uncertainty quantified by our approach is in well agreement with the subjective uncertainty as perceived by the physicians. Since the correlation is a bit higher for logistic regression than for the table classifier, one might be tempted to believe that the physicians' way of reasoning is closer to the simple model assuming independence of the predictor variables than to the more complex model that looks at all attributes simultaneously. Yet, given the small difference in terms of correlations and keeping in mind the factors that compromise full comparability, this conjecture is admittedly very speculative.

5.5 Accuracy-Rejection Curves

In the previous experiment, we compared the uncertainty as quantified by our model with the subjective uncertainty of the medical expert. The goal of the experiment presented

in this section is to validate the reasonableness of our approach in a somewhat more "objective" way. To this end, we compute so-called *accuracy-rejection* curves [19]. Roughly speaking, the idea is that, if $u(\boldsymbol{x}_0)$ is a reliable measure of the uncertainty involved in the classification of an instance $\boldsymbol{x}_0$, then this value should correlate with the probability to make a correct decision. Or, stated differently, when being allowed to abstain from the classification of all instances whose uncertainty exceeds a certain threshold ($u(\boldsymbol{x}_0) > t$), the classification accuracy should improve on the remaining instances. An accuracy-rejection curve is obtained by varying the rejection-threshold; it plots the classification accuracy (on the non-rejected instances) as a function of the reject rate (percentage of rejected instances).

In this experiment, we based the reject decision on the overall uncertainty $u(\boldsymbol{x}_0) = u_a(\boldsymbol{x}_0) + u_e(\boldsymbol{x}_0)$ of the classifier. In the case of the GPs, we again used U_{GP} as their measure of uncertainty. In order to produce the accuracy-rejection curves, we run a 5-times 10-fold cross-validation experiment to obtain predictions and their corresponding uncertainties.



(a) (b)

(c)

Figure 6: Accuracy-Rejection curves for (a) logistic regression, (b) table classifier and (c) the GPs.

The expected monotone dependency between reject rate and classification accuracy is confirmed by the results shown in Figure 6, not only for the classifiers but also for the GPs. Thus, although the overall accuracy of the GPs is lower, they seem to be well aware

of their uncertainty, which is a quite interesting finding. More importantly, however, the same can be observed for both classifiers, showing that the uncertainty produced by our model is indeed a reliable indicator of the uncertainty involved in a classification decision.

5.6 Statistics for Different Patient Groups

As mentioned earlier, the instance space consists of only 32 different patients or, more specifically, patient representations (patient groups) in this study. Table 2 summarizes some properties of these patient groups. From this table, several interesting observations can be made, including the following:

- Epistemic uncertainty is supposed to be higher for smaller patient groups. And indeed, the (Spearman) correlation between epistemic uncertainty and the relative group size is -0.91 in the case of the table classifier. Despite being still negative, the absolute correlation is much smaller for logistic regression (-0.15), a result which is again expected in light of our discussion about model assumptions and the incorporation of background knowledge.
- Misclassifications are expected to happen more often for patients with a high overall uncertainty. The corresponding correlation for the physicians, calculated for the mean uncertainty of the physicians and their error rate, is 0.67. For the classifiers, this correlation is slightly weaker (0.51 for logistic regression and 0.61 for the table classifier).

6 Experiments on Benchmark Data

In order to broaden our analysis, we conducted additional experiments with standard (binary classification) benchmark data sets from the UCI [1] and the StatLib [16] repositories. Of course, in contrast to the data we used in the medical case study, these benchmark data sets do not provide any extra information about the reliability of a classification. Therefore, most of the analyses of the previous section cannot be repeated here.

What can still be done with this kind of data is the computation of accuracy-rejection curves, which, as explained in Section 5.5, provide at least an indication of the learner's ability to estimate and quantify reliability in a proper way. The curves that we produced for the data sets included in our study are presented in Figure 7. As can be seen, these curves are coherent with our previous results and fully confirm the conclusions that we have already drawn in the medical case study.

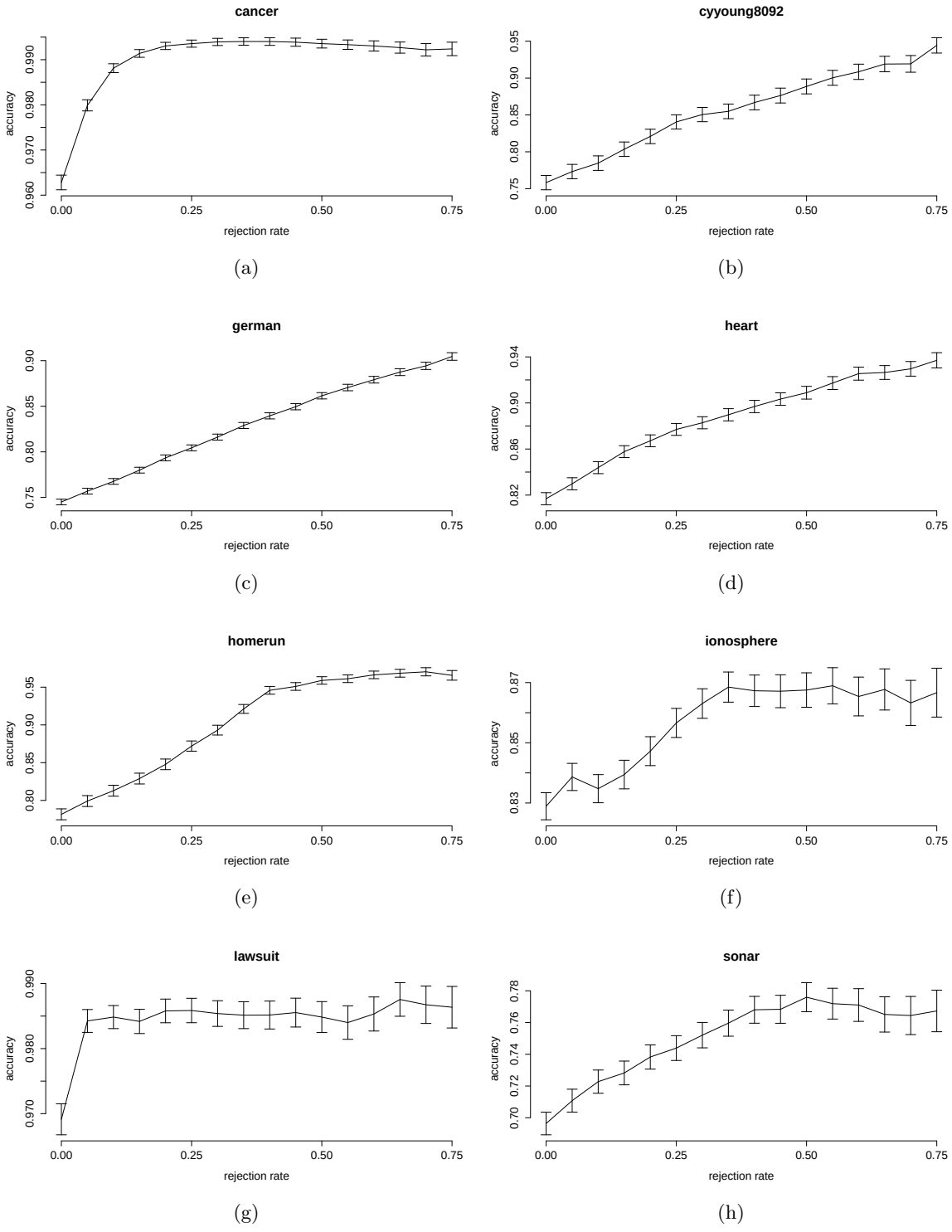


Figure 7: Accuracy-rejection curves for logistic regression on benchmark data sets.

7 Summary and Conclusions

In this paper, we made a distinction between two types of uncertainty in prediction, namely aleatoric and epistemic uncertainty, and argued that this distinction should be of interest in the context of learning from data, especially in safety-critical domains such as medical diagnosis. Our understanding of a "reliable classifier" is a system that is "self-aware" in the sense of knowing what it knows—and what not. Correspondingly, in addition to the predictions themselves, the system should provide information about the reliability of these predictions.

Here, a first step toward a system of that kind has been made. More concretely, we developed a method for binary classification which, given a new query instance and two different decision alternatives, does not only decide in favor of one of them, but instead produces a prediction in the form of a quadruple: a degree of (strict) preference for the first alternative, the same for the second alternative, the level of aleatoric uncertainty, and the level of epistemic uncertainty.

Empirically, we evaluated our approach on a practically relevant medical data set containing information about more than 1000 chest pain patients and their diagnoses. Our results are very promising insofar as they show that our measures of uncertainty harmonize quite well both with the uncertainty expressed by the medical experts and the difficulty of the diagnoses (as reflected by the probability of a wrong decision).

For future work, we plan to further expand on empirical studies of that kind, not restricted to the medical domain but also in other fields. Besides, there is of course scope for improvements and further developments on the methodological side. One important aspect, for example, is the computational complexity of our approach to reliable classification. In fact, since each prediction involves the solution of an optimization problem, the method is clearly critical from this point of view. Needless to say, the question of how to solve this optimization problem in the most efficient way strongly depends on the underlying model class which is used for learning. Developing efficient implementations for the most common approaches to (binary) classification is therefore another important topic of future work.

Apart from methodological advancements and applications in the context of medical decision making, we are also interested in corroborating our formal model with empirical evidence and in testing its "cognitive plausibility". For example, several neurophysiological studies using functional magnetic resonance imaging (fMRT) have supported the assertion that different forms of uncertainty exist [2, 15, 21], although definitions of uncertainty seem to vary. Levy et al. [15] were able to show that the neural representation of subjective value in situations where the estimation of different outcome probabilities is possible and in situations in which this is not possible, relies on the same structures,

namely the striatum and the medial prefrontal cortex. Pushkarskaya et al. [21] argue to differentiate between individuals who are averse or tolerant regarding missing information because uncertainty might have different neural representations depending on the attitude towards vague probabilities. On the one side, these studies can be regarded as an external validation of our mathematical model of different types of uncertainties. On the other side, our model could perhaps also be used to complement such studies in the sense of providing a tool for cognitive modeling, i.e., for the construction of formal models of cognitive processes complementing or explaining neurophysiological data.

References

[1] A. Asuncion and D. Newman. Uci machine learning repository, 2009. Accessed 13 Nov 2009.

[2] D.R. Bach, B. Seymour, and R.J. Dolan. Neural activity associated with the passive prediction of ambiguity and risk for aversive events. *Journal of Neuroscience*, 29(6):1648–56, 2009.

[3] V.N. Balasubramanian, R Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, and R.M. Siegel. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In *Proceedings of the IEEE Conference on Computers in Cardiology*, 2009.

[4] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] S. Bösner, A. Becker, M.A.A. Hani, H. Keller, A.C. Sonnichsen, J. Haasenritter, K. Karatolios, J.R. Schäfer, E. Baum, and N. Donner-Banzhoff. Accuracy of symptoms and signs for coronary heart disease assessed in primary care. *British Journal of General Practice*, 60(575):e246–e257, 2010.

[6] S. Bösner, J. Haasenritter, A. Becker, K. Karatolios, P. Vaucher, B. Gencer, L. Herzig, M. Heinzel-Gutenbrunner, J.R. Schäfer, M. Abu Hani H. Keller, A.C. Sönnichsen, E. Baum, and N. Donner-Banzhoff. Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule. *CMAJ*, 182(12):1295–1300, 2010.

[7] M. Cattaneo. Likelihood-based statistical decisions. In *Proc. 4th Int. Symposium on Imprecise Probabilities and their Applications*, pages 107–116, 2005.

[8] D. Dubois, S. Moral, and H. Prade. A semantics for possibility theory based on likelihoods. *Journal of Mathematical Analysis and Applications*, 205(2):359–380, 1997.

[9] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, 1988.

[10] J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[11] Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*, pages 312–317. IEEE, 1996.

[12] S.C. Hora. 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.

[13] EP. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic Publishers, 2002.

[14] A. Lambrou, H. Papadopoulos, and A. Gammerman. Reliable confidence measures for medical diagnosis with evolutionary algorithms. *IEEE Trans. on Information Technology in Biomedicine*, 15(1):93–99, 2011.

[15] I. Levy, J. Snell, A.J. Nelson, A. Rustichini, and P.W. Glimcher. Neural representation of subjective value under risk and ambiguity. *Journal of Neurophysiology*, 103(2):1036–47, 2010.

[16] M. Meyer and P. Vlachos. Statlib data, software and news from the statistics community, 2009. Accessed 13 Nov 2009.

[17] T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings* IJCAI-77, pages 305–310, 1977.

[18] T.M. Mitchell. The need for biases in learning generalizations. Technical Report TR CBM–TR–117, Rutgers University, 1980.

[19] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *Journal of Machine Learning Research - Proceedings Track*, 8:65–81, 2010.

[20] H. Papadopoulos, A. Gammerman, and V. Vovk. Confidence predictions for the diagnosis of acute abdominal pain. In *Artificial Intelligence Applications and Innovations III*, pages 175–184. 2009.

[21] H. Pushkarskaya, X. Liu, M. Smithson, and J.E. Joseph. Beyond risk and ambiguity: deciding under ignorance. *Cognitive, Affective, and Behavioral Neuroscience*, 10(3):382–91, 2010.

[22] N. Savage. Better medicine through machine learning. *Communications of the ACM*, 55(1):17–19, 2012.

[23] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[24] L.A. Wasserman. Belief functions and statistical evidence. *The Canadian Journal of Statistics*, 18(3):183–196, 1990.

[25] F. Yang, H. Zhen Wanga, H. Mi, C. de Lin, and W. Wen Cai. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, 10, 2009.

Patient					Group Statistics			Physician		Logistic Regression			Table Classifier		
patient believes its heart pain	pain reproducible	risk factor vascular disease	age gender compound	pain depends on exercise	relative group size	group size	relative frequency of CVD	mean uncertainty of physician	error rate of physician	mean aleatoric uncertainty	mean epistemic uncertainty	error rate classifier	mean aleatoric uncertainty	mean epistemic uncertainty	error rate classifier
0	0	0	0	0	.078	60	0.0	.177	.017	.024	0.0	0.0	.021	0.0	0.0
1	0	0	0	0	.107	82	0.0	.397	.146	.048	0.0	0.0	.031	0.0	0.0
0	1	0	0	0	.119	91	.011	.097	.011	.006	0.0	.011	.021	0.0	.011
1	1	0	0	0	.095	73	.014	.181	.014	.011	0.0	.014	.024	0.0	.014
0	0	1	0	0	.003	2	0.0	.250	.500	.072	0.0	0.0	.001	.600	0.0
1	0	1	0	0	.008	6	.333	.253	.333	.133	0.0	.333	.380	.315	.333
0	1	1	0	0	.005	4	0.0	.115	0.0	.017	0.0	0.0	.001	.124	0.0
1	1	1	0	0	.007	5	0.0	.348	0.0	.037	0.0	0.0	.002	.057	0.0
0	0	0	1	0	.055	42	.024	.250	.071	.108	0.0	.024	.035	0.0	.023
1	0	0	1	0	.112	86	.140	.475	.221	.213	0.0	.140	.261	0.0	.143
0	1	0	1	0	.061	47	0.0	.104	0.0	.026	0.0	0.0	.016	0.0	0.0
1	1	0	1	0	.072	55	.036	.321	.091	.052	0.0	.036	.054	0.0	.036
0	0	1	1	0	.016	12	.333	.250	.167	.296	0.0	.333	.347	.146	.308
1	0	1	1	0	.038	29	.241	.472	.241	.524	0.0	.241	.336	.013	.241
0	1	1	1	0	.022	17	.118	.221	.059	.082	0.0	.118	.147	.010	.118
1	1	1	1	0	.021	16	.063	.234	0.0	.159	0.0	.063	.084	.003	.063
0	0	0	0	1	.016	12	0.0	.312	0.0	.093	0.0	0.0	.099	.010	0.0
1	0	0	0	1	.023	18	.111	.469	.278	.174	0.0	.111	.146	.008	.111
0	1	0	0	1	.026	20	0.0	.094	0.0	.024	0.0	0.0	.007	0.0	0.0
1	1	0	0	1	.026	20	0.0	.217	.050	.046	0.0	0.0	.067	0.0	0.0
0	0	1	0	1	0.0	0	-	-	-	-	-	-	-	-	-
1	0	1	0	1	.001	1	0.0	.600	1.0	.465	0.0	0.0	.001	.999	1.0
0	1	1	0	1	.001	1	0.0	.600	0.0	.085	0.0	0.0	.001	.999	1.0
1	1	1	0	1	.003	2	0.0	.500	.500	.134	0.0	0.0	.001	.333	0.0
0	0	0	1	1	.008	6	.333	.533	.333	.341	0.0	.333	.355	.369	.333
1	0	0	1	1	.022	17	.412	.466	.353	.598	.067	.412	.549	.251	.412
0	1	0	1	1	.004	3	0.0	.053	0.0	.109	0.0	0.0	.001	.171	0.0
1	1	0	1	1	.016	12	.083	.498	.333	.184	0.0	.083	.087	.008	.083
0	0	1	1	1	.003	2	1.0	.600	.500	.641	.113	0.0	.259	.387	0.0
1	0	1	1	1	.026	20	.600	.366	.300	.472	0.0	.400	.448	.116	.400
0	1	1	1	1	.005	4	0.0	.140	0.0	.303	0.0	0.0	.001	.117	0.0
1	1	1	1	1	.003	2	.500	.400	.500	.434	0.0	.500	.001	.333	1.0

Table 2: Characteristics of all patient groups including uncertainty values and error rates (which are computed for the physician on the complete data and as out-of-sample statistics for the classifiers).