Information Sciences 255 (2014) 16–29

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty

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

[a] Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str., 35032 Marburg, Germany
[b] Department of Family Medicine, University of Marburg, Karl-von-Frisch-Str. 4, 35043 Marburg, Germany
[c] Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznan, Poland

ARTICLE INFO

Article history:
Received 16 October 2012
Received in revised form 18 May 2013
Accepted 29 July 2013
Available online 6 August 2013

Keywords:
Machine learning
Classification
Uncertainty
Medical decision making
Medical diagnosis

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.

* Corresponding author. Tel.: +49 6421 28 21 569; fax: +49 6421 28 21 573.
E-mail address: eyke@mathematik.uni-marburg.de (E. Hüllermeier).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.07.030

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 Fig. 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 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?