International Journal of Approximate Reasoning 55 (2014) 1519–1534

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization

Eyke Hüllermeier

Department of Mathematics and Computer Science, University of Marburg, Germany

ARTICLE INFO

Article history:
Available online 18 October 2013

Keywords:
Imprecise data
Fuzzy sets
Machine learning
Extension principle
Data disambiguation
Loss function

ABSTRACT

Methods for analyzing or learning from "fuzzy data" have attracted increasing attention in recent years. In many cases, however, existing methods (for precise, non-fuzzy data) are extended to the fuzzy case in an ad-hoc manner, and without carefully considering the interpretation of a fuzzy set when being used for modeling data. Distinguishing between an *ontic* and an *epistemic* interpretation of fuzzy set-valued data, and focusing on the latter, we argue that a "fuzzification" of learning algorithms based on an application of the generic extension principle is not appropriate. In fact, the extension principle fails to properly exploit the inductive bias underlying statistical and machine learning methods, although this bias, at least in principle, offers a means for "disambiguating" the fuzzy data. Alternatively, we therefore propose a method which is based on the generalization of loss functions in empirical risk minimization, and which performs model identification and data disambiguation simultaneously. Elaborating on the fuzzification of specific types of losses, we establish connections to well-known loss functions in regression and classification. We compare our approach with related methods and illustrate its use in logistic regression for binary classification.

1. Introduction

The learning of models from imprecise data, such as interval data or, more generally, data modeled in terms of fuzzy subsets of an underlying reference space, has gained increasing interest in recent years [4,6,7,23,30]. Indeed, while problems such as fuzzy regression analysis [2,8,9,12,13,27] have already been studied for a long time, the scope is currently broadening, both in terms of the problems tackled (e.g., classification, clustering, ranking) and the uncertainty formalisms used (e.g., probability distributions, histograms, intervals, fuzzy sets, belief functions).

Needless to say, learning from imprecise and uncertain data also requires the extension of corresponding learning algorithms. Unfortunately, this is often done without clarifying the actual meaning of an uncertain observation, although representations such as intervals or fuzzy sets can obviously be interpreted in different ways. In particular, an *ontic* interpretation of (fuzzy) set-valued data should be carefully distinguished from an *epistemic* one [10]. This difference is reflected, for example, in different approaches to fuzzy statistics, where fuzzy random variables can be formalized in an epistemic [17–19] as well as an ontic way [21]; see [5] for a comparison of these views in this context. Surprisingly, however, the fact that these two interpretations also call for very different types of extensions of existing learning algorithms and methods for data analysis seems to be largely ignored in the literature.

Under the ontic view, a variable can assume a fuzzy set as its "true value"; for example, one may argue that assigning a precise value to the variable "daily sunshine duration" is not very meaningful, and that a specification of sunshine durations

E-mail address: eyke@mathematik.uni-marburg.de.

0888-613X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2013.09.003

Table 1
Summary of the main notation used throughout the paper.

Notation	Meaning
$\boldsymbol{z}_i$, $(\boldsymbol{x}_i, y_i)$	(precise) data point, input/output sample
$\hat{\boldsymbol{z}}_i$, $\hat{y}_i$	(precise) prediction/estimator
Z_i, X_i, Y_i	sets or fuzzy sets (imprecise/fuzzy data)
$\mathcal{Z}, \mathcal{X}, \mathcal{Y}$	data space, input space, output space
$\mathbb{F}(\mathcal{Z})$	class of fuzzy subsets of $\mathcal{Z}$
$\mathcal{D}$, $\mathbf{D}$	sample of (precise) data points, class of potential samples
$\mathbb{D}$	sample of imprecise/fuzzy data
$\mathrm{INS}(\mathbb{D})$	set of instantiations of $\mathbb{D}$
$L(y, \hat{y})$	loss function, loss caused by $\hat{y}$ when compared to y
$\mathcal{L}(Y, \hat{y})$	loss caused by $\hat{y}$ when compared to set Y
$\mathbb{L}(Y, \hat{y})$	loss caused by $\hat{y}$ when compared to fuzzy set Y
M, $\mathbf{M}$	model, model space
$\mathcal{R}$, $\mathcal{R}_{emp}$	risk, empirical risk
$\overline{\mathcal{R}}_{emp}$	aggregated (empirical) risk
r_M	risk function mapping levels $\alpha \in (0, 1]$ to risk values

in terms of intervals or fuzzy sets is more appropriate. This interpretation suggests the learning of models that produce fuzzy sets as predictions, that is to say, models that *reproduce* the observed data. As opposed to this, a reproduction of the data appears less reasonable under the epistemic view, where fuzzy sets are used to describe, not the data itself, but the uncertain or imprecise *knowledge* about the data: A fuzzy set defines a possibility distribution that specifies a degree of plausibility for each potential precise value. As we shall explain in more detail later on, one should then rather try to "disambiguate" the data instead of reproducing it.

The possibilistic interpretation of fuzzy sets in the epistemic case, that we focus on in this paper, naturally suggests a "fuzzification" of learning algorithms based on an application of the generic extension principle [1,31]. As we shall argue, however, this approach is not appropriate and prone to fail in the context of data analysis. The main reason, to be detailed in Section 3, is a lack of differentiation between the possible data instantiations (i.e., the instantiation of each imprecise observation by a precise value). Such a differentiation, however, is typically suggested by the model assumptions through which the learning algorithm justifies its generalization beyond the data observed.

This idea of differentiating between instantiations of the data leads us to the notion of "data disambiguation" that we already mentioned above: *When learning from imprecise data under the epistemic view, model identification and data disambiguation should go hand in hand*. To this end, we propose an approach based on the generalization of loss functions in empirical risk minimization.

The rest of the paper is organized as follows. In the next section, we introduce the basic setting that we consider and the main notation that we shall use throughout the paper (see Table 1 for a summary). In Section 3, we explain the aforementioned problems caused by the use of the extension principle and elaborate on our idea of data disambiguation. Our new approach to learning from fuzzy data based on generalized loss functions is then introduced in Section 4. Section 5 is devoted to a comparison with an alternative and closely related method that was recently introduced by Denoeux [6,7]. In Section 6, we illustrate our approach on a concrete learning problem. Finally, we conclude with a summary and some additional remarks in Section 7.

2. Notation and basic setting

We consider the problem of *model induction*, which, roughly speaking, consists of passing from a specific data sample to a general (though hypothetical) model describing the data-generating process or at least certain properties of this process. In this setting, a learning (data analysis) algorithm **ALG** is given as input a set

$$\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^N \in \mathcal{Z}^N \tag{1}$$

of data points $\boldsymbol{z}_i \in \mathcal{Z}$. As output, the algorithm produces a model $M \in \mathbf{M}$, where $\mathbf{M}$ is a predefined model class. Formally, the algorithm can hence be seen as a mapping

$$\mathbf{ALG} : \mathbf{D} \longrightarrow \mathbf{M}, \tag{2}$$

where $\mathbf{D}$ is the space of potentially observable data samples. For instance, the data points might be vectors in $\mathcal{Z} = \mathbb{R}^d$, and the model could be a partitioning of the data into a finite set of disjoint groups (clusters). Or, the model could be a probability density function characterizing the underlying data-generating process. In fact, the data points $\boldsymbol{z}_i$ are typically assumed to be independent and identically distributed (i.i.d.) according to an underlying (though unknown) probability distribution. Moreover, the model class $\mathbf{M}$ is often parameterized, which means that each model $M \in \mathbf{M}$ is uniquely identified by a parameter $\theta \in \Theta$ (in other words, there is a bijection between the model space $\mathbf{M}$ and the parameter space Θ).