International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions

Luciano Sánchez [a,*], Inés Couso [b]

[a] Universidad de Oviedo, Departamento de Informática, Campus de Viesques, 33071 Gijón, Asturias, Spain
[b] Universidad de Oviedo, Departamento de Estadística e I.O. y D.M., Campus de Viesques, 33071 Gijón, Asturias, Spain

ARTICLE INFO

Article history:
Received 3 May 2017
Received in revised form 29 September 2017
Accepted 5 October 2017
Available online xxxx

Keywords:
Fuzzy rule-based models
Soft computing
Imprecise data

ABSTRACT

A framework is proposed for learning fuzzy rule-based systems from low quality data where the differences between observed and true values may introduce systematic bias in the model. It is argued that there are problems where aggregating imprecise losses into numerical or fuzzy-valued risk functions discards useful information, thus generalizing the risk of a model to a vector of fuzzy losses is preferred. The principles governing a learner that is capable of optimizing these fuzzy multivariate risk functions are discussed. Illustrative use cases are worked to exemplify those situations where new framework could become the alternative of choice.

1. Introduction

The term *uncomfortable science* was coined by Tukey [1] to describe cases where an inference must be drawn from a limited sample of data because the collection of further data is not feasible. Inducing knowledge from limited datasets poses a completely different set of challenges than big data analysis because, apart from computational efficiency considerations, small sets of data force the researcher to use the same information for both exploratory data analysis (unveiling cause/effect relationships) and confirmatory data analysis (testing whether these relationships are supported by the data) [2]. Systematic bias is potentially introduced [3] which may invalidate the confirmatory analysis. Post-hoc theorizing from small datasets is still possible [4], but a great effort must be made to ensure that all the available data is exploited to the full. Discarding "less than perfect" data is not an option for small datasets.

In this paper, some techniques for exploiting "less than perfect" data are studied. Datasets comprising incomplete observations, intermediate between precise perceptions and missing data, will be considered. These will be referred to as "low quality data". An encompassing treatment of these data as epistemic fuzzy sets is adopted.

1.1. Low quality data

Generally speaking, "low quality data" comprises those cases where the inaccurate perception of the data can introduce systematic bias in the learning, as shown in the following prototypical examples:

* Corresponding author.
E-mail address: luciano@uniovi.es (L. Sánchez).

https://doi.org/10.1016/j.ijar.2017.10.008
0888-613X/

- Discretized data ($a \leq x \leq b$, with x being the actual value and a, b perceived bounds) is very common because digital processing introduces this kind of uncertainty. For instance, if the true weight of an object is 7.3 and a digital scale displays "7" then our information about the weight is $6.5 \leq \text{weight} < 7.5$.
- Censored ($x \leq a$). For instance, in studies in mortality rate, the most that can be said about the death date of live individuals is that it is higher than the current date.
- Restrictions in multiple variables ($f(x_1, x_2, \ldots) = 0$). For instance, the marks of one student, given the class average score, or a high resolution image, given low resolution images [5].
- Tolerance intervals ($P(a \leq x \leq b) \geq 1-\alpha$), that are slightly more informative than coarsely discretized data. For instance, a GPS sensor may indicate us that our position is in a circle of radius 5 m, centered at a point in the map, and this information is guaranteed to be valid at least 95% of times. GPS sensors operate this way because the size of a radius that is true 100% of times would be too large for any practical purposes. Moreover, stacked tolerance intervals may be given for different degrees of confidence, i.e. the radius is lower than 5 m more than 95% of times, lower than 10 m. 99% of times, etc. Under the level-cut representation of a fuzzy set, stacked tolerance intervals are fuzzy subsets of the observable space [6].

In all these cases, the key for avoiding the systematic bias is to distinguish the fact that an event is observed ("the value displayed at the scale is 7") from the fact that an event has happened ("the weight of the object is 7.3"). This problem has been realized early [7][8][9][10] but has not received a complete treatment until recently [11][12].

Observe that incomplete observations can be either crisp or fuzzy subsets of complete observations. However, not all of the interpretations of a fuzzy set are pertinent to this problem. This paper regards the "epistemic" or "disjunctive" approach, where sets are used to describe an incomplete knowledge about the vector of attributes and/or the response variable [13].

1.2. Context and aim of the study

In this paper, learning a model is understood as determining the model with the best empirical risk, given a sample or "training dataset" and a parametric family of models. We also consider that uncertain measurements are not different than partially missing values: the degree of knowledge about a value can be complete (precise data), null (missing data) or partial (imprecise data).

It is well known that missing data can be either removed or imputed while the data is preprocessed, and the same can be said about partially missing data, that can be either removed via instance or feature selection, or imputed (although the term "defuzzification" is preferred for fuzzy data). In turn, imputation can be single (the uncertain value is replaced by a precise measurement) or multiple (the uncertain value is replaced by a set of precise measurements). Lastly, multiple imputation can be combined with also multiple models (a different model is learnt for each of the surrogate measurements) or a single model (whose risk is different for each of the surrogate measurements). The drawback of learning multiple models is that predictions become set-valued (one prediction for each model) but learning a single model is also troublesome because in this last case it is the empirical risk that becomes set-valued ("fuzzy fitness") and the learning algorithm must be designed in accordance.

This paper is about this last research line. We are interested in epistemic fuzzy models that operate with generalized definitions of loss and risk, thus a single model can be learnt without the need of preprocessing the data for removing the uncertainty. In this context, this paper aims to illustrate some cases where current techniques for learning fuzzy models from imprecise data still have room for improvement, in connection with recent advances in the use of generalized loss functions in machine learning [14].

The paper is organized as follows: in Section 2, a brief study about the state of the art in epistemic models for imprecise data is given, along with a discussion about the weakness of the current models. It is shown that none of the existing methods is better than the others, and also that not all the available information is being used in the learning, because models with different merits can be assigned the same risk. Because of this, new criteria for performing compared evaluations of fuzzy models are proposed in Section 3. These new criteria do not aggregate imprecise losses into fuzzy-valued risks, but the risk of a model is replaced by a vector of fuzzy losses. The principles governing a learner that would be capable of optimizing these fuzzy multivariate risk functions are discussed. In Section 4 three case studies are worked to illustrate those cases where the new framework is expected to have a competitive advantage. Section 5 concludes the paper. Lastly, the metaheuristic that was developed in order to preview the results of the new research line in the cases of study is described in the Appendix.

2. State of the art about epistemic models for imprecise data

When data is partially incomplete, machine learning is not straightforward. There are studies about the conditions for which the difference between true and observed data can be ignored. Others support learning a different model for each possible completion, in a process that shares certain points in common with multiple imputation. Finally, epistemic interval-valued or fuzzy representations of the incomplete data can be used, and this last representation can also be understood as an (implicit) set-valued imputation. The most relevant lines of research about these four subjects are reviewed in the present section, along with a critical view and future research paths.

Please cite this article in press as: L. Sánchez, I. Couso, A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions, Int. J. Approx. Reason. (2017), https://doi.org/10.1016/j.ijar.2017.10.008