Information Sciences 358–359 (2016) 129–150

Contents lists available at ScienceDirect

Information Sciences

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

Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach

Inés Couso[a,*], Luciano Sánchez[b]

[a] Department of Statistics and Operational Research, University of Oviedo, Spain
[b] Department of Computer Sciences, University of Oviedo, Spain

ARTICLE INFO

Article history:
Received 15 September 2015
Revised 7 March 2016
Accepted 6 April 2016
Available online 12 April 2016

Keywords:
Regression
Classification
Loss function
Generalized stochastic ordering
Set-valued data
Low-quality data

ABSTRACT

We study those problems where the goal is to find "optimal" models with respect to some specific criterion, in regression and supervised classification problems. Alternatives to the usual expected loss minimization criterion are proposed, and a general framework where this criterion can be seen as a particular instance of a general family of criteria is provided.

In the new setting, each model is formally identified with a random variable that associates a loss value to each individual in the population. Based on this identification, different stochastic orderings between random variables lead to different criteria to compare pairs of models. Our general setting encompasses the classical criterion based on the minimization of the expected loss, but also other criteria where a numerical loss function is not available, and therefore the computation of its expectation does not make sense.

The presentation of the new framework is divided into two stages. First, we consider the new framework under standard situations about the sample information, where both the collection of attributes and the response variables are observed with precision. Then, we assume that just incomplete information about them (expressed in terms of set-valued data sets) is provided. We cast some comparison criteria from the recent literature on learning methods from low-quality data as particular instances of our general approach.

1. Introduction

We deal with those machine learning problems where the goal is to find an optimal model $f : \mathcal{X} \to \mathcal{Y}$ relating some response variable $Y : \Omega \to \mathcal{Y}$ to a collection of attributes $\mathbf{X} : \Omega \to \mathcal{X}$, both of them defined on the same population Ω. These optimization problems usually aim at minimizing the expected loss, according to some loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that assigns a specific value to every pair $(Y(\omega), f(\mathbf{X}(\omega))$, composed by the outcome of the response variable and its estimate based on the collection of attributes, for every individual $\omega \in \Omega$. Very typical examples of that are the square and the absolute value of the difference ($\Delta(y, \hat{y}) = (y - \hat{y})^2$ and $\Delta'(y, \hat{y}) = |y - \hat{y}|$, respectively), both of them commonly used in regression problems, as well as the $0-1$-valued loss function $\Delta(y, \hat{y}) = 1_{\hat{y} \neq y}$, frequently used in classification problems. But sometimes, a numerical valued loss function is impossible to assess. For instance, an expert can tell us that classifying a girl with severe dyslexia as non-dyslexic is worse than classifying her as having a moderate dyslexia, but he may be unable to provide us with specific loss values on a numerical scale.

* Corresponding author. Tel.: +34 985181906.
E-mail addresses: couso@uniovi.es (I. Couso), luciano@uniovi.es (L. Sánchez).

http://dx.doi.org/10.1016/j.ins.2016.04.016
0020-0255/

On the other hand during the last years, there has been a growing interest in the development of learning models from set-valued datasets, extending existing learning algorithms to the case where our data points are not elements in the Cartesian product $\mathcal{X} \times \mathcal{Y}$ anymore, but (crisp or fuzzy) subsets of it (see [8,9,11,17,25,27], among others). In order to do so, one should first distinguish between the "ontic" and the "epistemic" interpretations of set-valued data (see [4,12]). Under the ontic approach (also called the "conjunctive" approach), sets are understood as complex entities observed with precision. As pointed out by Hüllermeier in [17], this interpretation suggests learning models that produce sets as predictions, i.e., models that *reproduce* the observed data. Thus, methods based on this interpretation of (fuzzy) sets usually produce parametric models where the parameters are indeed subsets of the parametric space, instead of elements of it. On the contrary, under the epistemic approach (also called the "disjunctive" approach), sets are used to describe our (in) complete knowledge about the true outcomes of the vector of attributes and/or the response variable: we do not observe their exact values, but we just can provide sets that contain them with total certainty. In that case, we aim to find a crisp model that relates the (possibly ill-observed) response variable to the (also possibly ill-observed) random vector of attributes. Models are therefore usual functions of the form $f : \mathcal{X} \to \mathcal{Y}$, but our information about their respective performances over a particular individual ω is incomplete, and it can be naturally expressed in terms of a subset of the form:

$$\{\Delta(y, f(\mathbf{x})) \ : \ (\mathbf{x} \in \mathbb{X}(\omega), y \in \mathbb{Y}(\omega)\},$$

where $\mathbb{X}(\omega)$ and $\mathbb{Y}(\omega)$ denote the most precise sets that respectively contain $\mathbf{X}(\omega)$ and $Y(\omega)$ with certainty, according to our incomplete information. Our information about the expected loss is therefore also incomplete, and a partial or a total (pre)ordering over the class of subsets of the real line, extending the usual order, needs to be considered, in order to compare two different models. This technique gives birth to different extended methods, depending on the nature of the algorithms to be extended and the partial/total ordering selected [19–22,27,30,31].

From a practical point of view, "ontic" and "epistemic" interpretations are appropriate for different categories of problems. In short, ontic interpretations are used when the set-valued data does not model imprecision and "epistemic" interpretations are related to vagueness. For instance, using the set {English, Spanish, French} for describing the languages spoken by a person is an ontic interpretation, but a set-valued medical differential diagnosis {hyperparathyroidism, cancer}, meaning that other diagnosis than hyperparathyroidism and cancer are discarded and a finer diagnostic is not yet possible, is an epistemic interpretation. The uncertainty in the data is propagated to the loss function, as mentioned: if a patient has cancer and is diagnosed as {hyperparathyroidism, cancer}, this cannot be regarded as a successful classification, neither it is a failure. The methods mentioned in the preceding paragraph address this issue by means of a generalization of the expected error, i.e. the expected number of errors will be set-valued. Consequently, certain order must defined over these set-valued expected errors, and in case that this order is partial, the "best" model is in turn generalized to the set of minimal elements of this partial order.

Notice also that this methodology is not too different than that used in the so-called "multi-criteria" modeling problems. For example, consider a regression problem with two output variables. In this kind of problems, one could aggregate the expected squared error of all variables into an scalar value and solve the problem with the same optimization algorithms used in the univariate case. Alternatively, one can assign a vector of losses to each model and define a "dominance" operator among these multi-valued losses. In this case, the "best" model is not sought but a set of nondominated elements, the so-called "Pareto front" [1].

This paper makes use of these techniques in a more general context, where a numerical loss function is not necessarily defined, and therefore an expected loss minimization does not necessarily make sense. In our framework, every model is identified with a random variable representing its reward (the opposite to its loss). According to this view, any stochastic ordering will lead to a specific pairwise comparison criterion between models. In particular, the expected loss minimization criterion is based on a well known stochastic ordering called "dominance in the sense of expected utility". Here, we replace such a particular stochastic ordering criterion by a general family of criteria involving a wider family of stochastic orderings, including it just as a particular case. Other stochastic orderings do not require the compared random elements to be numerical valued, and therefore non numeric loss functions can be considered in our general environment.

A specific comparison criterion leads us either to an optimal model (if the criterion produces a total ordering between the different alternatives) or to a collection of non-dominated models (when the criterion produces a partial ordering). Different criteria may therefore produce different optimal solutions. The goal in the paper not to argue that any particular optimality criterion is better than another, but simply to present a general framework that facilitates future studies about different possible optimality criteria.

2. Basics and notation

Let Ω denote the population under study. Let $\mathcal{X}$ be the input space (the set of possible outcomes of the vector of attributes) and let $\mathcal{Y}$ denote the output space, i.e., the set of possible outcomes for the response variable. For instance, in a regression problem, the output space may coincide with the real line, while in a classification problem it will represent the collection of classes. $\mathbf{X} = (X_1, \ldots, X_d) : \Omega \to \mathcal{X}$ will denote the random vector of attributes and $Y : \Omega \to \mathcal{Y}$ will stand for the response variable (in particular, in classification problems, $Y(\omega)$ will represent the class of object ω.).

We will consider a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, penalizing wrong predictions. A model will incur a penalty $\Delta(y, \hat{y})$ if the true output value is y and the model predicts $\hat{y}$. Examples of loss functions commonly used in regression problems are