International Journal of Approximate Reasoning 96 (2018) 36–55

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Partial data querying through racing algorithms ☆

Vu-Linh Nguyen [a], Sébastien Destercke [a,*], Marie-Hélène Masson [a,b]

[a] UMR CNRS 7253 Heudiasyc, Sorbonne Universités, Université de Technologie de Compiègne, CS 60319 – 60203 Compiègne cedex, France
[b] Université de Picardie Jules Verne, France

ARTICLE INFO

Article history:
Received 5 April 2017
Received in revised form 10 March 2018
Accepted 11 March 2018
Available online 13 March 2018

Keywords:
Partial data
Interval-valued data
Set-valued labels
Data querying
Active learning
Racing algorithms

ABSTRACT

The paper studies the problem of actively learning from instances characterized by imprecise features or imprecise class labels, where by actively learning we understand the possibility to query the precise value of imprecisely specified data. We differ from classical active learning by the fact that in the later, data are either fully precise or completely missing, while in our case they can be partially specified. Such situations can appear when sensor errors are important to encode, or when experts have only specified a subset of possible labels when tagging data. We provide a general active learning technique that can be applied in principle to any model. It is inspired from racing algorithms, in which several models are competing against each others. The main idea of our method is to identify the query that will be the most helpful in identifying the winning model in the competition. After discussing and formalizing the general ideas of our approach, we illustrate it by studying the particular case of binary SVM in the case of interval valued features and set-valued labels. The experimental results indicate that, in comparison to other baselines, racing algorithms provide a faster reduction of the uncertainty in the learning process, especially in the case of imprecise features.

1. Introduction

Although classical learning schemes assume that every instance is fully specified, there are situations where such an assumption is unlikely to hold, and where the data can be qualified of *partial* or *imprecise*. By "partial data", we refer to the situation where either some features or the labels are imperfectly known, that is are specified by sets of possible values rather than a precise one. For example, when the label of some training instances is only known to belong to a set of labels, or when some features are imprecisely given in the form of intervals.

Classical statistical solutions to solve this problem include the use of different imputation techniques [5] or the use of likelihood-based techniques such as the EM algorithm [4] and its extensions. The use of such techniques however implies to satisfy specific statistical assumptions about the missingness process (e.g., missing-at-random assumption), that can be very hard or impossible to check in practice, especially since we do not have access to the original precise data. More recently, the problem of learning from partial data has gained an increasing interest within the machine learning community, and many methods [2,3,10] that have shown their efficiency for different problems have been developed. Yet, even if these

☆ This paper is part of the Virtual special issue on Uncertainty Management in Machine Learning Applications, Edited by Van-Nam Huynh.

* Corresponding author.
E-mail addresses: linh.nguyen@hds.utc.fr (V.-L. Nguyen), sebastien.destercke@hds.utc.fr (S. Destercke), mmasson@hds.utc.fr (M.-H. Masson).

https://doi.org/10.1016/j.ijar.2018.03.005
0888-613X/

methods can handle partial data, their performances usually degrade as data become more and more partial or imprecise, as more and more uncertainty is present in the learning process.

This work explores the following question about learning from partial data: if we have the possibility to gain more information on some of the partial instances, which instance and what feature of this instance should we query? In the case of a completely missing label (and to a lesser extent of missing features), this problem known as active learning has already been largely treated [16] and applied in different fields like natural language processing, text or image classification, recommender systems [6,14,23,19]. However, we are not aware of similar works concerning the case of partial data. Note that for the case of features, there is even very few active learning methods addressing the problem of missing features. In this work, we provide a new general active learning technique that can be applied in principle to any model and partially missing input/features, and illustrate it on the case of SVM. It is inspired from the concept of racing algorithms [12], in which several models are competing against each others. They were initially introduced to select an optimal configuration of a given lazy learning model (e.g., K-nn methods), and since then have been applied to other settings such as multi-armed bandits [9]. The idea of such racing algorithms is to oppose a (finite) set of alternatives in a race, and to progressively discard losing ones as the race goes along. In our case, the set of alternatives will be different possible models, and the race will consist in iteratively querying the precise value of some partial features or labels. Indeed, as data are partial, the performance of each model is uncertain and several candidate models can be optimal. By iteratively making queries, i.e. asking to an oracle the precise value of a partial data, these performances will become less and less uncertain, and more models will be discarded from the race. The key question is then to identify those data that will be the most helpful in reducing the set of possible winners in the race, in order to converge as quickly as possible to the optimal model.

The rest of this paper is organized as follows: we present in Section 2 the basic notations used in this paper. Section 3 introduces the general principles of racing algorithms and formalizes the problem of quantifying the influence of a query on the race. We then study the application of our approach using the particular case of a binary SVM. Section 4 is focused on interval-valued features, while Section 5 explores the case of set-valued labels. Some experiments are then performed in Section 6 to demonstrate the effectiveness of our proposals. Before concluding the paper, Section 7 discusses some computational issues of the presented approaches, generalizing some of the results concerning SVM method. Note that this paper is an extension of [13], with full proofs, larger experiments as well as the addition of the set-valued label case for binary SVM and a discussion about the complexity of the approach.

2. Preliminaries

In classical supervised setting, the goal of the learning approach is to find a model $m : \mathcal{X} \to \mathcal{Y}$ within a set $\mathcal{M}$ of models from a set $\mathbf{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} | i = 1, \ldots, n\}$ of n input/output samples, where $\mathcal{X}$ and $\mathcal{Y}$ are respectively the input and the output spaces.[1] The empirical risk $R(m)$ associated to a model m is then evaluated as

$$R(m) = \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i)) \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function, and $\ell(y, m(\mathbf{x}))$ is the loss of predicting $m(\mathbf{x})$ when observing y. The selected model is then the one minimizing (1), that is

$$m^* = \arg\min_{m \in \mathcal{M}} R(m). \tag{2}$$

Another way to see the model selection problem that will be useful in this paper is to assume that a model m_l is said to be better than m_k (denoted $m_l \succ m_k$) if

$$R(m_k) - R(m_l) > 0, \tag{3}$$

or in other words if the risk of m_l is lower than the risk of m_k. Given the relation $\succ$ on $\mathcal{M}$, Equation (1) then simply amounts to take as best model the maximal element of $\succ$, or in case of equality due to indifference, one of the maximal model chosen arbitrarily.

In this work, we are however interested in the case where data are partial, that is where general samples are of the kind $(\mathbf{X}_i, Y_i) \subseteq \mathcal{X} \times \mathcal{Y}$. Here and in the rest of this paper, capital letters are used for partial data and small letters will denote precise one, and bold letters will represent vectors and Cartesian products of feature values. When the data is partial, Equations (1), (2) and (3) are no longer well-defined, and can be extended in multiple different ways. Two of the most

[1] As $\mathcal{X}$ is often multi-dimensional, we will denote its elements and subsets by bold letters.