Knowledge-Based Systems 123 (2017) 41–60

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Self-training for multi-target regression with tree ensembles

Jurica Levatić[a,b,*], Michelangelo Ceci[c], Dragi Kocev[a,b,c], Sašo Džeroski[a,b]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[c] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 10 February 2017
Accepted 11 February 2017
Available online 12 February 2017

Keywords:
Semi-supervised learning
Self-training
Multi-target regression
Random forests
Reliability of predictions
Predictive clustering trees

ABSTRACT

Semi-supervised learning (SSL) aims to use unlabeled data as an additional source of information in order to improve upon the performance of supervised learning methods. The availability of labeled data is often limited due to the expensive and/or tedious annotation process, while unlabeled data could be easily available in large amounts. This is particularly true for predictive modelling problems with a structured output space. In this study, we address the task of SSL for multi-target regression (MTR), where the output space consists of multiple numerical values. We extend the self-training approach to perform SSL for MTR by using a random forest of predictive clustering trees. In self-training, a model iteratively uses its own most reliable predictions, hence a good measure for the reliability of predictions is essential. Given that reliability estimates for MTR predictions have not yet been studied, we propose four such estimates, based on mechanisms provided within ensemble learning. In addition to these four scores, we use two benchmark scores (oracle and random) to empirically determine the performance limits of self-training. We also propose an approach to automatically select a threshold for the identification of the most reliable predictions to be used in the next iteration. An empirical evaluation on a large collection of datasets for MTR shows that self-training with any of the proposed reliability scores is able to consistently improve over supervised random forests and multi-output support vector regression. This is also true when the reliability threshold is selected automatically.

1. Introduction

The major machine learning paradigms are supervised learning (e.g., classification, regression), where all the data are labeled, and unsupervised learning (e.g., clustering), where all the data are unlabeled. Semi-supervised learning (SSL) [1] examines how to exploit both labeled and unlabeled data, aiming to benefit from the information that unlabeled data bring. SSL is of a practical relevance because, in many real-world scenarios, labeled data are scarce due to a costly and/or time-consuming labelling procedure; while unlabeled data abound and are easy to obtain. For example, such scenarios are encountered in life sciences (gene function prediction, quantitative structure-activity relationship modelling), ecology (habitat and community modeling), multimedia (annotation of images and videos) and semantic web (categorization and analysis of text and web).

Intuitively, SSL yields best results when there are few labeled examples as compared to unlabeled ones (i.e., large-scale labelling is not affordable). Such a scenario is especially relevant for machine learning tasks with structured outputs where, due to the increased complexity of the output, labelling of the data is even more difficult. Consider, for example, the problem of natural language parsing, where the aim is to predict the parse tree that generates a given input sentence. To label the data, linguists need to determine the parse tree for input sentences. This is feasible for a few sentences, but for large number of sentences the process becomes very tedious and expensive: It took 2 years to manually construct parse trees for 4000 sentences of Penn Chinese Treebank [2]. At the same time unlabeled input sentences are readily available in vast amounts. Another prominent example comes from the ecological modelling domain, where some attribute values are easily available (e.g. temperature, humidity) whereas some other attribute values have to be manually collected/measured by experts and thus, can be the subject of the prediction process (e.g. water pollution in a river, or abundance of specific species which populate the river). Obviously, in the latter case, data collection is very expensive and time consuming, so only few observations can be obtained with limited resources [3].

In this study, we are concerned with SSL for the task of *multi-target regression* (MTR). MTR is a structured output prediction task

* Corresponding author.
E-mail addresses: Jurica.Levatic@ijs.si (J. Levatić), michelangelo.ceci@di.uniba.it (M. Ceci), Dragi.Kocev@ijs.si (D. Kocev), Saso.Dzeroski@ijs.si (S. Džeroski).

http://dx.doi.org/10.1016/j.knosys.2017.02.014
0950-7051/

where the goal is to predict multiple continuous target variables (also known as multi-output or multivariate regression). In many real-life problems, we are interested in simultaneously predicting multiple continuous variables. Prominent examples of this task come from ecology: predicting the abundance of different species occupying the same habitat [4], assessing different properties of forests [5], or estimating vegetation quality indices [6]. We argue that SSL, as for classical machine learning tasks, can lead to improved predictive capabilities also for MTR by leveraging the contribution of unlabeled examples and, at the same time, by exploiting the possible dependencies among the multiple target variables.

The handful of existing SSL methods for structured output prediction almost exclusively deal with discrete outputs. Here, a prominent work was done by Brefeld [7], who used the co-training paradigm and the principle of maximizing the consensus among multiple independent hypotheses to develop semi-supervised support vector learning algorithm for joint input-output spaces and arbitrary loss. Zhang and Yeung [8] proposed a semi-supervised method based on Gaussian processes for a task related to MTR: multi-task regression. In multi-task learning the aim is to predict multiple single-target variables with different training sets (in general, with different descriptive attributes) at the same time. The few existing SSL methods for MTR are highly specialized for individual applications. For example, Navaratnam et al. [9] have proposed a SSL method for MTR specialized for computer vision. On the other hand, SSL for single-target regression has received more attention in the past [10–13]. While it is possible to decompose a MTR problem into several (local) single-target ones and use such methods, there are several advantages of learning a global multi-target model over learning a separate local model for each target variable. Global models have better computational efficiency, typically perform better and overfit less than a collection of single-target models [14,15].

We propose a global SSL method for MTR. More specifically, we extend the self-training approach [16] to the task of MTR. The main advantage of this iterative SSL approach is that it can be "wrapped" around any existing (supervised) method. In the past, several studies have proposed supervised methods for solving the task of MTR directly and demonstrated their effectiveness [6,14,17,18]. We propose to use predictive clustering trees (PCTs), or more precisely, random forests [19] of PCTs for MTR, as base predictive models [14] for the self-training approach. PCTs are a generalization of standard decision trees towards predicting several types of structured outputs: tuples of continuous/discrete variables, hierarchies of classes, and time series.

The main principle of self-training is iterative usage of its own most reliable predictions for the unlabeled data as additional data in the training process. The most reliable predictions are selected by applying a threshold on the reliability scores of predictions. A good reliability scoring function assigns a high score to the predictions with low error and a low score to the predictions with high error. Obviously, a proper reliability scoring function is crucial for the success of self-training, since an error once made can reinforce itself in the subsequent iterations. However, developing a good reliability scoring function is not a simple task [20]: This has not yet been entirely resolved in the single-target regression and classification, and even less for the task of MTR.

In this paper, we propose and evaluate several reliability scoring functions for MTR, which are based on the mechanisms provided by ensemble learning. Namely, we use the variance of the votes of an ensemble and random forest proximities to estimate the reliability of predictions [19,20]. These reliability estimates are by-products of ensemble learning. Hence, they impose almost no additional computational overhead, as opposed to some other reliability estimates for regression [20]. This aspect is especially important in SSL, where we can expect to deal with huge amounts of unlabeled data and/or to re-train the model several times (as in self-training). In order to empirically determine the performance limitations of the proposed approach to self-training for MTR, we use oracle scoring (the best possible scoring function) and random scoring as benchmark scoring functions. Finally, we explore the influence of two strategies for merging per-target scores into a global score (normalized averaging and ranked averaging) on the performance of self-training.

We also consider the problem of automatically determining the threshold on the reliability scores of predictions. The thresholding is crucial for the selection of the unlabeled examples with the most reliable predictions. The selected unlabeled examples together with the predictions are then considered as training examples in the next iteration. To this end, we propose an automatic threshold selection algorithm for SSL that exploits the out-of-bag error obtained when learning the ensemble.

Our study investigates three important questions: (1) Can unlabeled data improve predictive performance on MTR tasks in a self-training setting? (2) Which reliability scoring function yields the best predictive performance in this setting? (3) Can we exploit the advantage introduced by the self-training setting for MTR when an automatic threshold selection algorithm is used? To address these questions, we perform experimental evaluation of self-training with the various reliability scoring functions using 9 MTR datasets from various domains. The evaluation reveals that self-training, coupled with any of the proposed reliability scoring functions, is able to outperform a supervised random forest and multi-output support vector regression (MSVR) [21,22]. In particular, all of the proposed reliability scoring functions performed better than random scoring. The best results, excluding the upper (oracle) limits on the performance of self-training, were achieved by using a reliability score based on the variance of the votes of ensemble members.

We summarize the major contributions of this paper as follows:

- A SSL method tailored for the task of MTR based on the predictive clustering framework for predicting structured outputs.
- Two reliability scoring functions for MTR predictions, two normalization strategies, and two strategies for merging per-target scores into a global score.
- Empirical determination of the upper bounds on the performance, i.e., the potential of SSL with self-training.
- An automatic threshold selection algorithm, i.e., a practical solution for exploiting the potential of SSL with self-training.
- Empirical evaluation of the proposed method on 9 MTR datasets.

An initial investigation of the proposed SSL method for MTR has been presented in a workshop paper [23,24]. We extend that work along six major dimensions. First, we propose a new reliability scoring function based on the random forest proximities, in addition to the one based on the variance of the votes of the ensemble members. Second, we propose four strategies for obtaining a single global score from the per-target scores. Third, we propose an oracle score to test the performance bounds of the self-learning paradigm. Fourth, we propose an algorithm for the automatic selection of the threshold for reliability of predictions. Fifth, we consider a more sophisticated stopping criterion for self-training, which automatically stops learning if the performance begins to degrade. Finally, the empirical evaluation is performed on a larger collection of MTR datasets.

The remainder of the paper is organized as follows. In the next section, we present the background of the work presented here. This includes a discussion on related work on the topics of MTR and SSL, and a brief description of the predictive clustering framework. Next, in Section 3 we describe the self-training approach and the proposed reliability scores for MTR. The experimental design