https://doi.org/10.31449/inf.v46i4.4455 Informatica **46** (2022) 583–584

Semi-supervised Learning for Structured Output Prediction

Jurica Levatić
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
E-mail: jurica.levatic@ijs.si

Thesis Summary

Keywords: semi-supervised learning, predictive clustering trees, predicting structured outputs

Received: October 4, 2022

This article presents a summary of the doctoral dissertation of the author on the topic of semi-supervised learning for predicting structured outputs.

Povzetek: Članek predstavlja povzetek doktorske disertacije avtorja, ki obravnava temo polnadzorovanega učenja za napovedovanje strukturiranih vrednosti.

1 Introduction

In contrast to traditional supervised machine learning methods, which use only labeled data, semi-supervised methods additionally use unlabeled data. Due to laborious annotation procedure, labeled data are a limited asset in many real-life problems, which can hinder the predictive performance of algorithms. Unlabeled data, on the other hand, are often much easier to obtain. Semi-supervised learning (SSL) [1] aims to exploit unlabeled data to achieve better performance than can be achieved by labeled data alone.

Structured output prediction (SOP) is concerned with predicting structured, rather than scalar values, such as multiple classes/variables, hierarchies or sequences [2]. Such outputs are encountered in many applications of predictive modeling. Compared to SSL for primitive outputs, SSL for SOP received much less attention in the scientific community, although the need for SSL is even stronger there: Obtaining labels of structured data is even harder. Furthermore, this field lacks interpretable methods and methods that can handle various SOP tasks.

2 Methods and evaluation

In the thesis [3], to overcome the aforementioned issues, we extend the predictive clustering (PC) framework towards SSL. The PC framework [4] is implemented using predictive clustering trees (PCTs) which can efficiently handle various SOP tasks. We propose two classes of semisupervised methods stemming from the PC framework that can handle the following SOP tasks: multi-target regression, multi-label classification and hierarchical multi-label classification.

The first class of methods is based on the self-training paradigm - it uses its own most reliable predictions in the learning process. We propose a self-training method for multi-target regression based on ensembles of predictive clustering trees [5]. To the best of our knowledge, this is currently one of the very few general-purpose semi-supervised methods for this type of structured output. Since the reliability of predictions in the context of multi-target regression was not studied before, we propose two different reliability scores for predictions based on intrinsic mechanisms of ensemble methods. Furthermore, we propose an algorithm for automatic selection of the appropriate threshold on reliability scores.

The second class of methods we propose is based on the extension of the variance functions of predictive clustering trees in order to accommodate both labeled and unlabeled examples [6, 7]. This enables to build semi-supervised predictive clustering trees that can exploit unlabeled examples while preserving the appealing characteristics of supervised trees, such as interpretability and computation efficiency. Semi-supervised predictive clustering trees are general in terms of the type of the structured output: They can predict different types of structured outputs: multiple target variables and hierarchically structured classes. We propose parametrization of semi-supervised predictive clustering trees by which it is possible to control the amount of supervision, i.e., the learned models can range from fully unsupervised to fully supervised.

We perform an extensive empirical evaluation of the proposed methods on a wide range of datasets from different domains and with different types of structured output. We analyze the influence of the amount of labeled data to the performance of the proposed methods, as we all various aspects of their practical usability, such as, interpretability, computational complexity, and sensitivity to parameters.

3 Discussion and Conclusions

The thesis contributes to the field of SSL for SOP with two classes of global semi-supervised methods for structured output prediction: self-training for multi-target regression

[6] and semi-supervised predictive clustering trees [6, 7]. The empirical evaluation showed that the proposed methods outperform their supervised counterparts on a number of datasets from different domains and with different types of structured outputs.

The self-training approach offers a state-of-the-art predictive performance on multi-target regression problems, while producing black-box models and with the cost of increased computational complexity (due to iterative training of the base model) as compared to supervised random forests. Semi-supervised predictive clustering trees, on the other hand, produce readily interpretable models, which are often considerably more accurate than the corresponding supervised models for structured outputs. The semi-supervised predictive clustering trees (and ensembles thereof) also exhibit attractive predictive performance on machine learning tasks with primitive outputs, i.e., classification and regression.

We also perform two case studies demonstrating the practical usability of the proposed semi-supervised methods: (1) We show that the proposed semi-supervised methodology is well-suited for quantitative structure-activity relationship modeling, i.e., prediction of biological activity of chemical compounds [8]; (2) We demonstrate on the problem of water quality prediction that semi-supervised predictive clustering trees can efficiently learn from partially labeled data [9].

There are a number of possible directions to continue the work presented in the thesis, such as extending the proposed methods to other structured output prediction tasks, such as time-series classification or sequence learning, or utilising the proposed methods to develop feature ranking for semi-supervised and unsupervised learning.

References

[1] Chapelle, O., Schölkopf, B., Zien, A. (2006). *Semi-supervised learning*. Cambridge, Massachusetts: MIT Press.

[2] G. Bakır, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S. Vishwanathan (2007) *Predicting structured data*, The MIT Press.

[3] J. Levatić (2017) *Semi-supervised learning for structured output prediction*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.

[4] H. Blockeel (1998) *Top-down induction of first order logical decision trees*, PhD Thesis, Katholieke Universiteit Leuven, Belgium.

[5] J. Levatić, M. Ceci, D. Kocev, S. Džeroski, (2017) Self-training for multi-target regression with tree ensembles, *Knowledge-based systems*, 123:41–60

[6] J. Levatić, D. Kocev, M. Ceci, S. Džeroski, (2018) Semi-supervised trees for multi-target regression, *Information Sciences*, 450:109–127

[7] J. Levatić, M. Ceci, D. Kocev, S. Džeroski, (2017) Semi-supervised classification trees, *Journal of Intelligent Information Systems*, 49(3):461–486

[8] J. Levatić, M. Ceci, T. Stepišnik, S. Džeroski, D. Kocev, (2020) Semi-supervised regression trees with application to QSAR modelling, *Expert Systems with Applications*, 158:113569

[9] S. Nikoloski, D. Kocev, J. Levatić, D. P. Wall, S. Džeroski, (2021) Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland, *Ecological Informatics*, 61:101161