Ensembles for Predicting Structured Outputs

Dragi Kocev
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
E-mail: Dragi.Kocev@ijs.si Web: http://kt.ijs.si/DragiKocev

Thesis Summary

Keywords: ensemble methods, predictive clustering trees, predicting structured outputs

Received: February 7, 2012

This article presents a summary of the doctoral dissertation of the author on the topic of building ensembles for predicting structured outputs.

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

1 Introduction

In many real life problems of predictive modelling, the output is structured, meaning that there can be dependencies between classes or some internal relations between the classes (e.g., classes are organized into a tree-shaped hierarchy or a directed acyclic graph). These types of problems occur in domains such as life sciences (gene function prediction, drug discovery), ecology (analysis of remotely sensed data, habitat modelling), multimedia (annotation and retrieval of images and videos) and the semantic web (categorization and analysis of text and web). Having in mind the needs of the application domains and the increasing quantities of structured data, the task of “mining complex knowledge from complex data” was listed as one of the ten most challenging problems in data mining [5].

2 Methods and evaluation

In the thesis [2], we address the task of learning models for predicting structured outputs that take as input a tuple of attribute values and produce as output a structured object. In contrast to classification and regression, where the output is a single scalar value, in our case the output is a data structure, such as a tuple or a directed acyclic graph. We consider both global and local prediction of structured outputs, the first based on a single model that predicts the entire output structure and the latter based on a collection of models, each predicting a component of the output structure.

A variety of methods, specialized for predicting a given type of structured output, have been proposed [1]. However, many of them are computationally demanding and not suited for dealing with large datasets (especially large outputs). In the thesis, we propose to use predictive clustering trees (PCTs) [3] for efficient and accurate prediction of structured outputs. PCTs offer a unifying approach to dealing with different types of structured outputs. We extend PCTs in the direction of ensemble methods [4] to further increase their predictive performance.

In particular, we take the notion of an ensemble, i.e., a collection of predictive models whose predictions are combined, and apply it in the context of predicting structured outputs. We develop methods for learning different types of ensembles of PCTs for global and local prediction of different types of structured outputs. The different types of ensembles include bagging, random forests, random subspaces and bagging of subspaces. The types of outputs considered correspond to the different predictive modeling tasks, i.e., multi-target regression, multi-target classification, and hierarchical multi-label classification. Each of the combinations can be applied both in the context of global prediction (producing a single ensemble) or local prediction (producing a collection of ensembles).

Computational complexity analyses of the methods show that the global ensembles are the most efficient, especially random forests. The analyses also indicate that the proposed approaches are scalable to datasets which can be large along any of the following dimensions: number of attributes, number of examples, and size of the target. This is confirmed also by the empirical evaluation of the proposed methods on a large number of datasets.

3 Conclusion

The thesis makes several contributions to the area of machine learning and the respective application areas. First, we propose ensemble learning methods for predicting structured outputs that use PCTs as base predictive models. The proposed methods are general in terms of the type of the structured output: they support the tasks of predicting multiple continuous targets, predicting multiple discrete targets, and hierarchical multi-label classification.

Second, we perform an extensive empirical evaluation of the proposed methods over a variety of benchmark datasets. We construct ensembles of up to 1000 predictive models and select ensembles of 50 global predictive models as optimal in terms of predictive performance and efficiency.

Third, we compare the performance of ensembles of global models and single global models, as well as ensembles of local models. Both global and local ensembles perform better than the single model counterparts in terms of predictive power. Global and local ensembles perform equally well, with global ensembles being more efficient and producing smaller models, as well as needing fewer trees in the ensemble to achieve the maximal performance.

Fourth, we apply the proposed methods in three practically relevant domains. (1) We constructed models that assess vegetation condition from remotely sensed data and generated maps of the state of Victoria, Australia [6]. (2) On the task of hierarchical annotation of medical X-ray images, the ensembles of PCTs provided the best annotation results reported so far in the literature [8]. (3) Extensive experimental evaluation over several tasks of gene function prediction in three organisms showed that bagging of PCTs is superior to or competitive with state-of-the-art methods [7].

In the thesis, we also present some preliminary results that further explore the proposed paradigm of ensembles for structured prediction. We first discuss structured prediction for different types of structured outputs. Next, we propose a method for feature ranking in the context of structured outputs, based on random forests. Finally, we suggest a novel ensemble learning method that is based on the beam search strategy and can control directly the diversity in the ensemble.

References

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

[2] D. Kocev (2007) *Ensembles for predicting structured outputs*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.

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

[4] G. Seni, J. Elder (2010) *Ensemble methods in data mining: Improving accuracy through combining predictions*, Morgan & Claypool Publishers.

[5] Q. Yang, X. Wu (2006) 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making*, 5(4):597–604.

[6] D. Kocev, S. Džeroski, M. White, G. Newell, P. Griffioen (2009) Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, *Ecological Modelling*, 220(8):1159–1168.

[7] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, S. Džeroski (2010) Predicting gene function using hierarchical multi-label decision tree ensembles, *BMC Bioinformatics*, 11(2):1–14.

[8] I. Dimitrovski, D. Kocev, S. Loskovska, S. Džeroski (2011) Hierchical annotation of medical images, *Pattern Recognition*, 44(10-11):2436–2449.