Deep forests with tree-embeddings and label imputation for weak-label learning

Pedro Ilídio[1,2], Ricardo Cerri[3], Celine Vens[1,2], and Felipe Kenji Nakano[1,2]

[1] KU Leuven, Campus KULAK, Department of Public Health and Primary Care
[2] Itec, imec research group at KU Leuven
{pedro.ilidio, celine.vens, felipekenji.nakano}@kuleuven.be
[3] Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação
cerri@icmc.usp.br

Abstract. This study proposes deep forest approaches that incorporate tree embeddings (TE) for weak-label learning, exploring new label imputation techniques to be applied after each layer. We introduce two novel imputation methods: Strict Label Complement (SLC) and Fluid Label Addition (FLA). SLC establishes fixed estimates for missing annotations as upper limits, while FLA adjusts imputation probabilities dynamically at each layer. The proposed models achieve comparable or superior performance to the state-of-the-art, highlighting the potential of TE and imputation methods in weak-label learning.

Keywords: Deep forest · Weak-label learning · Tree-embeddings · Multi-label classification

1 Introduction and method

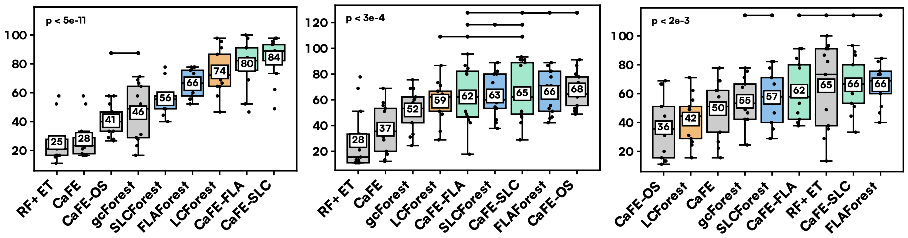
Weak-label learning consists of multi-label classification problems in which the negative label annotations are unreliable [5]. Deep forests [6] have been proven to be an effective algorithm in these settings [4]. A deep forest is a sequence of decision forests, in which each forest augments the feature space before the next one is trained. The current state-of-the-art technique for tabular weak-label learning [4] enhances deep forests with a label imputation procedure applied after each layer. However, the technique tends to overestimate of the number of missing positives and relies on inefficient criteria for model-length control. Tree-embeddings (TE) are feature augmentation techniques leveraging the structure of decision trees [3]. Recent advances [3] have shown that TE improve the performance of deep forests in the supervised case, but the result was not yet tested in the context of weak labels.

Our study [1] is thus concerned with two main objectives: i) investigating the benefits of employing TE for weak-label learning tasks; and ii) proposing new techniques for label imputation in deep forests. Two label imputation methods were proposed, named Strict Label Complement (SLC) and Fluid Label Addition (FLA). SLC determines fixed estimates for the number of missing annotations and uses them as upper limits for a label imputer model. FLA performs such estimates on each layer, adjusting the probabilities of imputation without setting an upper limit. We then introduce four new algorithms: CaFE-SLC and CaFE-FLA (which incorporate TE); and SLCForest and FLAForest (which do not).

2 Experiments

We performed 5-fold cross-validation on 13 multi-label datasets of various scientific domains. We only used the labels with at least 30 positive annotations. 30%, 50% and 70% of these positive annotations were randomly masked to generate weak label tasks.

For performance metrics considering binary outcomes (Fig. 1a), the models employing label imputation were consistently the top-performers. CaFE-SLC and CaFE-FLA outperformed SLCForest and FLAForest, demonstrating the efficiency of TE for weak-label settings. When scoring the predicted probabilities directly, our models consistently outperformed the main baseline, LCForest [4] (Fig. 1c). If no label masking is performed, imputation still significantly improved performance in some cases. For example, FLAForest performs better than CaFE [3] and gcForest [6] (Fig. 1b), suggesting the existence of unexpected missing positives.



(a) MCC, 70% missing (b) MCC, supervised (c) AP, 70% missing

Fig. 1: Percentile ranks for the Matthews Correlation Coefficient (MCC) and Average Precision (AP). Crossbars connect models that are not statistically distinguishable (p > 0.05, Wilcoxon signed ranks).

3 Conclusion

TE and label imputation proved to be effective strategies for deep forests applied to weak-label learning. Our models resulted in comparable or superior performance to the state-of-the-art in all the investigated settings. In future works, we will expand our methods to hierarchical multi-label classification [2] and partial multi-label learning.

Acknowledgments. The authors would like to thank the Flanders Government (AI Research Program), FWO (1235924N), and FAPESP, grant number 2022/02981-8. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

1. Ilídio, P., Vens, C., Cerri, R., Nakano, F.K.: Deep forests with tree-embeddings and label imputation for weak-label learning. In: Proceedings of the 2024 International joint conference on neural networks. pp. 1–8. Yokohama (2024)
2. Nakano, F.K., Lietaert, M., Vens, C.: Machine learning for discovering missing or wrong protein function annotations: a comparison using updated benchmark datasets. BMC bioinformatics **20**, 1–32 (2019)
3. Nakano, F.K., Pliakos, K., Vens, C.: Deep tree-ensembles for multi-output prediction. Pattern Recognition **121**, 108211 (2022)
4. Wang, Q.W., Yang, L., Li, Y.F.: Learning from weak-label data: A deep forest expedition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6251–6258. New York (2020)
5. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Science Review **5**(1), 44–53 (2018)
6. Zhou, Z.H., Feng, J.: Deep forest. National Science Review **6**(1), 74–86 (2019)