Extended Abstract: An Interpretable Semi-supervised Classifier using Rough Sets for Amended Self-labeling*

Isel Grau[1], Dipankar Sengupta[1,2], Maria M. Garcia Lorenzo[3], and Ann Nowe[1]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Belgium
[2] Centre for Cancer Research and Cell Biology, Queen's University Belfast, UK
[3] Department of Computer Science, Universidad Central de Las Villas, Cuba

This document is an extended abstract of the paper accepted at the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), held from the 19th to the 24th of July, 2020, Glasgow, United Kingdom.

Gathering data examples for training a machine learning classifier in a real-world scenario is often simple. However, the process of assigning labels to the examples can be costly in terms of money, time, or effort. In such scenarios, we might obtain datasets with more unlabeled than labeled data. Semi-supervised classification [4] techniques arise from the need to address this problem using both labeled and unlabeled data for training a classifier. The aim is to increase the classifier's generalization ability compared to a supervised classifier that only uses the available labeled data.

On the other hand, an increasing requirement observed in machine learning is to obtain not only precise models but also interpretable ones. End users often demand an insight into how an algorithm arrives at a particular outcome and needs an explanation of the decisions. A certain degree of global interpretability can be obtained using more transparent techniques as proxies for solving a task [1]. We refer to intrinsically interpretable models (e.g., linear regression, decision trees or decision lists) as white boxes, as opposed to the less interpretable black-box ones. Grey-box models use white boxes as surrogates for distilling previously trained black boxes. The grey boxes attempt to explain the domain by approximating the predictions produced by a black-box classifier, in an intrinsically interpretable structure.

In this paper, we explore the performance of the *self-labeling grey-box* (SlGb) [2]. In the SlGb, we use a black-box classifier to predict the decision class of the unlabeled instances, while a surrogate white box is used to build an interpretable predictive model, based on the whole instance set. The aim is to outperform the base white-box component using only the available labeled data,

* This work was supported by the IMAGica project, financed by the Interdisciplinary Research Programs and Platforms (IRP) funds of the Vrije Universiteit Brussel; and the BRIGHTanalysis project, funded by the European Regional Development Fund (ERDF) and the Brussels-Capital Region as part of the 2014-2020 operational program through the F11-08 project ICITY-RDI.BRU (icity.brussels).

while maintaining a good balance between performance and interpretability. The SlGb approach's performance largely depends on the black-box classifier's prediction capability when classifying unseen instances. In the context of self-labeling, the classification mistakes can reinforce themselves if no amending procedure is used during self-training. Therefore, we explore the effect of two amending procedures for assigning more importance to more reliable instances before training the surrogate white box, avoiding the propagation of errors or inconsistent information. The first strategy is based on class membership probabilities provided by the black box in the self-labeling. The second strategy aims to correct the inconsistency in the labels in the enlarged dataset by computing the certainty of the classification based on the Rough Set Theory (RST) [3] inclusion degree measure.

The experiments show that the choice of a white box and amending is relevant for the size of the structure. SlGb produces simpler models when using decision lists instead of a C4.5 decision tree as surrogate white boxes, even when no amending is performed. However, the amending procedures help further increase the simplicity without affecting the prediction rates by giving more importance to confident instances in the self-labeling. Especially RST based amending looks more promising since it does not need the black-box base classifier to provide calibrated probabilities. Furthermore, RST-based amending could be the right choice for a given case study where the uncertainty coming from inconsistency is high, even on the available labeled data. The study varying the number of unlabeled instances and labeled instances together shows that even when the number of labeled instances is not that scarce, the SlGb is able to leverage unlabeled instances for increasing the performance. Another conclusion is that adding unlabeled instances does not make the interpretability worse compared to adding more labeled instances. This evidences that the RST-based amending avoids that the SlGb generates more rules from inconsistent instances. Finally, the experimental comparison shows that our SlGb method outperforms the state-of-the-art self-labeling approaches, yet being far more simple in structure than these techniques.

References

1. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
2. Grau, I., Sengupta, D., Lorenzo, M.M.G., Nowe, A.: Interpretable self-labeling semi-supervised classifier. In: Proceedings of the IJCAI/ECAI 2018 2nd Workshop on Explainable Artificial Intelligence. pp. 52–57 (2018)
3. Pawlak, Z.: Rough sets. International Journal of Computer & Information Sciences **11**(5), 341–356 (1982)
4. Zhu, X., Goldberg, A.: Introduction to semi-supervised learning. Morgan & Claypool Publishers (2009)