Grey-Box Model: An ensemble approach for addressing semi-supervised classification problems

Isel Grau IGRAUGAR@VUB.AC.BE; IGRAU@UCLV.EDU.CU
Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuani km 5 1/2, 54830 Santa Clara, Cuba

Dipankar Sengupta DIPANKAR.SENGUPTA@VUB.AC.BE
Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

Maria Matilde Garcia Lorenzo MMGARCIA@UCLV.EDU.CU
Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuani km 5 1/2, 54830 Santa Clara, Cuba

Ann Nowe ANN.NOWE@VUB.AC.BE
Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

Keywords: ensemble classifiers, semi-supervised classification, interpretability

Abstract

In this paper, we propose a novel and interpretable *grey-box* ensemble using a self-labeled approach for semi-supervised classification problems. The prospective *grey-box* ensembles a more interpretable *white-box* model with a *black-box* technique. This scheme could guide the comparatively data expensive *white-box* component with the results from the more accurate *black-box* part. We evaluate the proposal in an inductive learning setting showing good performance in partially labeled datasets.

1. Introduction

Researchers in machine learning and related domains are primarily concerned with obtaining accurate prediction methods, fitting a wide range of real-world problems. Regrettably, the most successful classifiers are *black-box* models, providing no insights into the reasoning process associated with the decision model. However, in some domains where machine learning models are applied (e.g. bioinformatics), the transparency in their predictions is crucial (Robnik-Sikonja & Kononenko, 2008). Also, in some real-world classification tasks it is easier to obtain unlabeled than labeled data because it requires less effort, expertise or time. This class of machine learning problems, known as *semi-supervised classification*, can be tested into two settings: transductive or inductive learning. One of the most successful approaches in semi-supervised classification is based on supervised learning and is denoted as self-labeled techniques (Triguero et al., 2015). The proposal of accurate semi-supervised classifiers providing interpretability remains an open problem since the trade-off between accuracy and transparency is difficult to achieve. In the next section, we introduce an interpretable *grey-box* ensemble that is capable of outperforming the prediction accuracy of the base-line *white-box* classifier when tested on an inductive setting of partially labeled datasets.

2. *Grey-box* ensemble architecture

In order to build an interpretable semi-supervised classifier, we design a *grey-box* ensemble model. In the proposed architecture, we explicitly allow the *black-box* component to generate extra labeled data for the *white-box* component, since the *black-box* models can more reliably classify unseen data when trained on limited data. More explicitly, the labeled data is first provided to the *black-box* component for training purposes. Thereafter, the unlabeled data is processed by the trained *black-box* obtaining an additional set of la-

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

beled data. Further, the originally labeled data along with the extra labeled data are provided to the *white-box* model for training, obtaining an interpretable classifier with a likely enhanced performance in semi-supervised problems, compared to a *white-box* model trained only with the initially labeled data (Fig. 1).

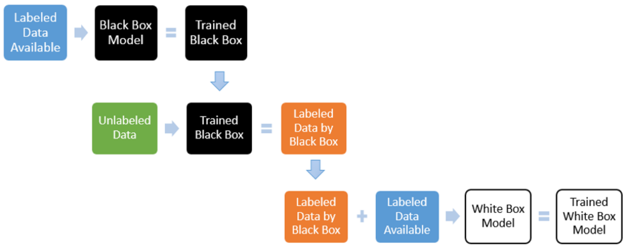

Figure 1. General architecture of the *grey-box* model.

What are we expecting from this approach? The inclusion of the *black-box* classifier (e.g., Random Forests (Breiman, 2001)) increases the accuracy when assigning the decision class to unlabeled data, whereas the the *white-box* (e.g., Fuzzy Cognitive Maps (Kosko, 1986)) allows interpreting the reasoning process over both original and predicted labeled data. Our hypothesis is that the *grey-box* model should be able to outperform the accuracy of the *white-box* classifier when the amount of labeled data is limited.

3. Numerical simulations

For the validation of our hypothesis, we have tested several *black-box* and *white-box* models as components of the ensemble, against a benchmark of partially labeled datasets with regards to an inductive setting (separate test set). As *black-box* models, we have used: Random Forests, Multilayer Perceptron and Support Vector Machines; since these techniques are known to be the most likely successful approaches for addressing a real-world problem in terms of accuracy (Fernández-Delgado et al., 2014). While for the *white-box* component we have studied different kinds of interpretability: Decision Tree (if-then rules), Bayesian Networks (Bayesian reasoning) and Fuzzy Cognitive Maps (causality between variables). In order to study the behavior of the proposal against different sizes of partially labeled datasets, we use learning curves where the X axis indicates the size of labeled data set provided to the learning algorithm and the Y axis is the kappa value achieved by the model. We use kappa statistic as performance measure since it is more robust on imbalanced datasets (Triguero et al., 2015). In our simulations, we studied the performance of the proposed architecture using 121 benchmark datasets from the UCI Machine Learning Repository (Lichman, 2013). All datasets are divided in train and test set for validation. We simulate different quantities of unlabeled data by using random re-sampling without replacement, splitting the training set in unlabeled and labeled sets. The percentage of labeled instances ranges from 5% to 100% of the training data.

Fig. 2 shows the average learning curve for a *grey-box* model using Random Forest and Decision Tree (the best performing combination). Observe that when only 5% to 70% of the data is labeled, the performance of the *grey-box* is considerable better than the *white-box*, i.e. the *black-box* component is boosting the accuracy of the whole model. Although the *black-box* model is still better than the proposed *grey-box*, it is not interpretable. Also, it is observed that the performance of the *grey-box* and the *white-box* behave comparably as the amount of labeled data approaches the utilization of complete training set as labeled data. This behavior is however expected in our experiments since increasing the number of labeled data is equivalent to reducing the number of unlabeled instances to be labeled by the *black-box* classifier. Therefore, we obtained an architecture where we keep the interpretability of the *white-box* (a Decision Tree in this case) while boosting its performance in presence of partially unlabeled datasets.

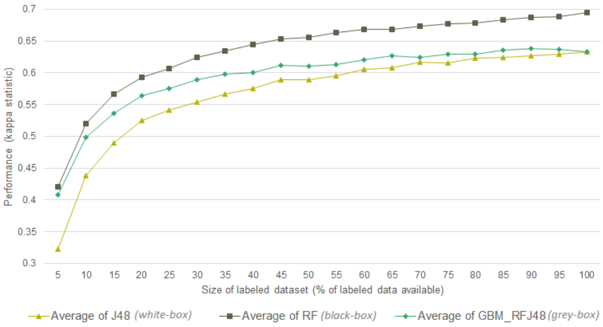

Figure 2. Learning curves for the studied models.

This model will be tested in a bioinformatics context: the pathogenesis classification of Arrythmia Syndromes (Antzelevitch et al., 2005). This classification task emerges from BRIDGEIRIS project, where a total of 23278 variants are being processed (Grau et al., 2015). Due to the time cost and large number of instances, only a small part will be manually classified and used as labeled dataset. The proposed interpretable semi-supervised technique will not only help clinicians by the elimination of manual process, but also on gaining insights in the relations between variant features and their pathogenic classification.

References

Antzelevitch, C., Brugada, P., Borggrefe, M., Brugada, J., Brugada, R., Corrado, D., Gussak, I., LeMarec, H., Nademanee, K., Riera, A. R. P., et al. (2005). Brugada syndrome: report of the second consensus conference endorsed by the heart rhythm society and the european heart rhythm association. *Circulation*, *111*, 659–670.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*, 3133–3181.

Grau, I., Daneels, D., Dooren, S. V., Bonduelle, M., Farid, D. M., Croes, D., Nowé, A., & Sengupta, D. (2015). Gevact: Genomic variant classifier tool. *BeNeLux Bioinformatics Conference* (p. 1). Antwerp, Belgium.

Kosko, B. (1986). Fuzzy cognitive maps. *International Journal of man-machine studies*, *24*, 65–75.

Lichman, M. (2013). UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences.* http://archive.ics.uci.edu/ml.

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *Knowledge and Data Engineering, IEEE Transactions on*, *20*, 589–600.

Triguero, I., García, S., & Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, *42*, 245–284.