Teaching Explanations by Examples ⋆

Cèsar Ferri[1], José Hernández-Orallo[1], and Jan Arne Telle[2]

[1] vrAIn, Universitat Politècnica de València, Spain
{cferri, jorallo}@dsic.upv.es
[2] Department of Informatics, University of Bergen, Norway.
Jan.Arne.Telle@uib.no

Machine teaching is an emerging field that has recently attracted the general attention in AI [7]. Briefly, machine teaching can be considered as an inverse problem to machine learning. Concretely, the goal of machine teaching is to find the smallest (optimal) training set that –using a learning algorithm– produces a target model. Machine teaching has been applied in many different fields. For instance, in education the "learner" can be a human student, and the teacher has a target model (i.e. the educational goal). If we assume a cognitive learning model of the student, machine teaching can be employed to reverse-engineer the optimal training data. In other words, we obtain the data that is going to optimise the learning process for that student, like a personalised lesson.

However, most results in the machine teaching literature only apply to concept languages with examples that do not have structure. When confronted with richer languages, we find that we may teach a concept with a single example, but this example might be arbitrarily large. Looking for a more intuitive way of assessing the theoretical feasibility of teaching concepts for structured languages, in [5] we introduced *the teaching size* and obtained results for universal languages (e.g., Turing machine or natural language). We included an experimental validation of our method for teaching a universal language: the universal language P3, a simple language for string manipulation. When coupled with a strong bias for simplicity, we found the remarkable result that, in many cases, teaching a concept with examples led to shorter descriptions than giving the shortest rule-based or program-base transcription of the logic of the decision. For the first time, we showed both theoretically and empirically that teaching with examples is often more *efficient* than giving the concept itself.

In this work we propose to explore the use of machine teaching for providing explanations of AI models. Decades of converting black boxes into white boxes have not solved the problem of extracting comprehensible explanations to justify the decisions made by a model. Either these models oversimplify the problem or they are not assimilated by humans, or both. This is not only because the techniques in explainable AI ignore the psychology of the recipient (the *explananti*) but also because they ignore the way in which concepts can be easily transmitted from one language of representation to another. Machine teaching techniques can be employed to extract significant instances from AI systems and ML models that can be used to (1) give humans a better understanding of the behaviour

⋆ This research was supported by the EU (FEDER) and the Spanish MINECO (RTI2018-094403-B-C32), Universitat Politècnica de València (PAID-06-18), and the Generalitat Valenciana (PROMETEO/2019/098 and BEST/2018/027). J. Hernández-Orallo is also funded by FLI (RFP2-152).

of complex AI systems, and (2) provide reassurance that the learner has really identified the right concept, and not confused by an incorrect one, forced by the strong bias on explanation simplicity.

Let us consider the well-known *Monks1* problem [6]. This a simple binary problem with six categorical features, although here we consider a simplified problem with only the three relevant features: $V1, V2, V3$ with $(3, 3, 4)$ distinct values respectively. The class is positive (2) when $V1 = V2 \vee V3 = 1$, otherwise it is negative (1). This dataset has been widely used to show the limitations of machine learning methods that can not capture relations between features.

Consider we train a MLP network with the *Monks1* dataset. If we see this model as a black-box, a popular approach to explain it is to build a comprehensible model (e.g. a decision tree) from a surrogate dataset labelled using the black box model [3]. Figure 1 includes an unpruned J48 decision tree [4] of twelve leaves built from a surrogate dataset.

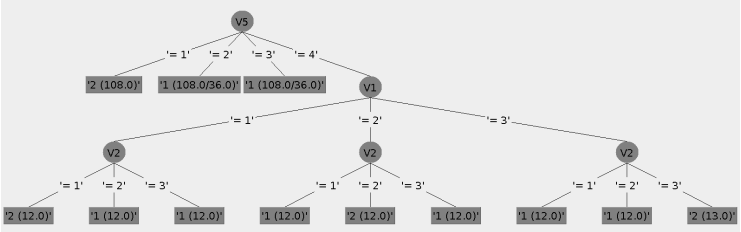

Fig. 1. An unpruned J48 tree extracted from a MLP model of the *Monks1* problem.

As we illustrated in [5] using a universal language, in some cases it is more effective to transmit a set of key instances instead of sending the model. Do we find the same situations when explaining concepts to humans? Do humans understand the *Monks1* concept better with the model of Figure 1, or is it more effective just teaching the concept by using a few set of insightful examples? In the *Monks1* case, the concept contains a relational pattern and illustrative examples can be more useful to humans to gain knowledge about the behaviour of model. For instance, the example of Table 1 could give humans a more comprehensible snapshot of the model's behaviour.

V1	V2	V3	Class	V1	V2	V3	Class
1	1	2	2	1	2	1	2
2	2	3	2	1	2	3	1
1	2	2	1				

Table 1. Illustrative examples of the *Monks1* problem

Let us compare the transmission cost of this theory with respect the estimated witness set of the problem. Consider a non-binary decision tree with n_f attributes and n_c classes, where each attribute i has $nf(i)$ distinct values. For each node we need a bit to indicate if it is a leaf or a branch. In the case of a branch we need to send the attribute of the split $log_2(n_f)$ bits. In the case of a leaf, we need

to send only the class. We set an order between receiver and sender in the tree traverse and also in the $nf(i)$ distinct values of all the attributes. Considering this coding, we need $1 + log_2(3)$ for the five branch nodes and $1 + log_2(2)$ for the twelve leaf nodes. Thus, the decision tree can be sent with 36.92 bits. In the case of the witness set, we have five examples, and we need $log_2(3) + log_2(3) + log_2(4) + log_2(2) = 6.17$ for each example, then 30.84 in total. Again, in this scenario, teaching a concept with examples led to a shorter description than giving the rule-based logic of the concept itself.

Machine teaching can be used to find the smallest set of examples from which a learning system (in this case humans) can induce the concept. For that reason we need to model the learning bias of humans, which is very different from the bias of machine learning approaches. For instance, humans are specially good at capturing relational concepts or considering negation.

We propose to explore different machine teaching techniques in order to, given a target concept, extract different versions of witness sets. An experimental evaluation using human understanding of the proposed witness sets could be useful to discover which machine teaching methods are the most appropriate to be used as an instance-base teacher for explaining concepts to humans.

An alternative to be considered in the framework is an interactive scenario between learner and teacher, as in [2]. In this work, the authors propose an active teacher model that can query the learner for knowing the learner's status. The use of this knowledge allows the teacher to guide better the learner.

There are a few works that combine explanation and machine teaching already. In [1] the authors show that by leveraging explanations, teaching can be significantly accelerated. The authors propose the NOTES algorithm. The method is based on a formalism of the teaching problem as a two-stage decision-making process: the learner's attention model and her decision model.

References

1. Chen, Y., Aodha, O.M., Su, S., Perona, P., Yue, Y.: Near-optimal machine teaching via explanatory teaching sets. In: Proc. of the 21st International Conference on Artificial Intelligence and Statistics. vol. 84, pp. 1970–1978. PMLR (2018)
2. Dasgupta, S., Hsu, D., Poulis, S., Zhu, X.: Teaching a black-box learner. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 1547–1555. PMLR (2019)
3. Domingos, P.: Knowledge acquisition form examples vis multiple models. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 98–106. Morgan Kaufmann Publishers Inc. (1997)
4. Hornik, K., Buchta, C., Zeileis, A.: Open-source machine learning: R meets Weka. Computational Statistics **24**(2), 225–232 (2009)
5. Telle, J.A., Hernández, J., Ferri, C.: The teaching size: Computable teachers and learners for universal languages. In: Under review (2019)
6. Thrun, S.B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., et al.: The monk's problems a performance comparison of different learning algorithms. Tech. rep. (1991)
7. Zhu, X.: Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 4083–4087. AAAI'15, AAAI Press (2015)