

Teaching Unknown Learners to Classify via Feature Importance

Extended Abstract

Carla Guerra
INESC-ID & Instituto Superior
Técnico, Universidade de Lisboa
Lisbon, Portugal
carla.guerra@gaiaps.inesc-id.pt

Francisco S. Melo
INESC-ID & Instituto Superior
Técnico, Universidade de Lisboa
Lisbon, Portugal
fmelo@inesc-id.pt

Manuel Lopes
INESC-ID & Instituto Superior
Técnico, Universidade de Lisboa
Lisbon, Portugal
manuel.lopes@tecnico.ulisboa.pt

ABSTRACT

In this work we introduce an interactive machine teaching approach that teaches a classification task to the learner. Our adaptive approach - Feature Importance Teaching (FIT) - does not assume perfect knowledge about the learner, as most machine teaching approaches do. It chooses, online, which sample to show next, as it updates the learner's model based on feedback from the student on the weights attributed to the features. We present simulated results where the student has a different prior knowledge from the one assumed by the teacher. The results have shown that our teaching approach can mitigate this mismatch and lead to a significantly faster learning curve than the ones obtained in conditions where the teacher randomly selects the samples or does not consider this feedback from the student.

KEYWORDS

Machine Teaching; Classification Tasks; Interactivity

ACM Reference Format:

Carla Guerra, Francisco S. Melo, and Manuel Lopes. 2021. Teaching Unknown Learners to Classify via Feature Importance: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021, IFAAMAS*, 3 pages.

1 INTRODUCTION

Machine teaching (MT) considers the problem of finding the smallest set of examples that allows a specific learner to acquire a given concept, explicitly considering a computational learning algorithm for the student [5, 7–9]. Since a significant amount of teaching relies on providing examples, the learning efficiency can be greatly improved if the teacher selects the examples that are more informative for each particular learner. MT, however, often assumes that the learner is completely known [2, 4, 6, 7]. This very strong assumption that the teacher has perfect knowledge about the learner is seldom (if ever) satisfied, particularly in the case where the learner is human.

In this work we propose considering interactivity in the teaching process as the means to solve this problem. The main research question addressed in this paper is, thus: *how can interactivity be used in machine teaching systems to improve the learning performance of students?* To answer this question we contribute with a novel interactive approach for a machine teaching system - Feature Importance-based Teaching (FIT) - that teaches classification tasks

based on feedback from the learner regarding its use of the features of the data. FIT asks the learner about the feature weights used during classification in order to estimate the learner's prior knowledge, and then selects the samples that minimize the expected future error based on a model closer to the real model of the learner.

We tested our teaching approach against two other approaches— one also interactive [3] and another non-interactive. In the simulations we assumed the student learning model is known by the teacher but we consider that the prior knowledge is different between student and teacher. The results obtained show that our approach teaches faster than the compared approaches.

2 THE FIT APPROACH

The problem addressed regards interactively teaching a classification task to one student by providing the smallest number of samples that guarantees he learns the task, i.e., that he reaches a small classification error. More formally, a sample is an entry from a labeled dataset, $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where each x_i is an M -dimensional feature vector, and $y_i \in \{1, \dots, C\}$ is its corresponding class label. We call teaching set, D_t , to the set composed by the teaching samples selected, which is a subset of D . To accomplish our goal we propose an interactive system that adaptively selects these samples based on information asked to the learner about the importance he attributes to each feature used to classify a given sample. We refer to this information as the learner feature weights, encoded as an M dimensional vector $W_l = [w_{f_1}, w_{f_2}, \dots, w_{f_M}]$, where each w_{f_i} is the relevance value assigned by the learner to the i^{th} feature when classifying a specific sample. Our goal with this information is to better model the learner and, therefore, choose samples that make him learn faster. The interaction process followed by the FIT teaching system is illustrated in Figure 1.

2.1 Interactively Updating the Learner's Model

To model the learner we consider two details: his learning model, θ_l , and his prior knowledge, PK_l . However, since we can not assume perfect knowledge about the learner, the teaching system considers a learner model θ_l and a prior knowledge PK_l . The learning model is assumed to follow a *Logistic Regression* (LR), thus we define the learner's model by a vector of parameters $\theta_l = [W_l, b_l]$, where $W_l \in \mathbb{R}^d$ is his weight vector and $b_l \in \mathbb{R}$ is a scalar offset term.

The learner feature weights, W_l , are interactively obtained from the student, which makes $W_t = W_l$. This allows us to estimate his prior knowledge, that we encode in the sample set PK_t . At each moment, the vector PK_t can be seen as summarizing the sample set that the student has observed up to that moment. It is composed by the samples the student had seen previously to teaching, $preD_t$

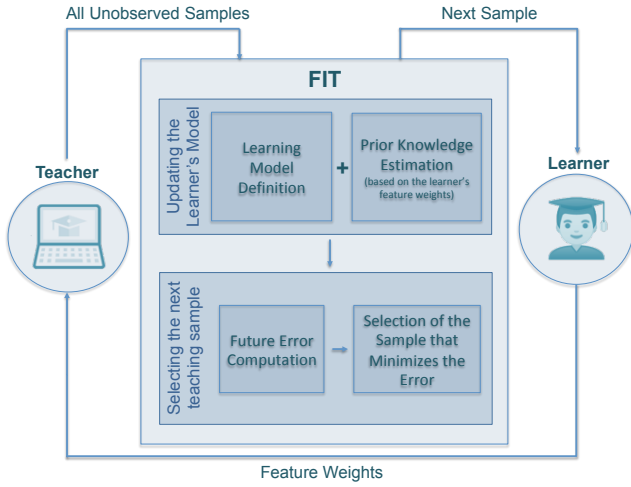


Figure 1: Diagram illustrating the interaction process in FIT.

(which are interactively inferred using the learner feature weights), plus the ones showed during teaching, D_t (adaptively chosen by the system). To estimate $preD_t$ we follow the approach of Liu et al. [4] that reconstructs such a training set from the parameter vector θ_t . To do this we need $n = 2 \frac{\lambda ||W_t||^2}{2\tau_{max}}$ training samples, where $x_k = x_+, y_k = 1$, for all $k \in \{1, \dots, \frac{n}{2}\}$ and $x_k = x_-, y_k = -1$, for all $k \in \{\frac{n}{2} + 1, \dots, n\}$. The samples x_+ and x_- are designed satisfying:

$$x_+^T W_t = t - b_t, \quad x_- = x_+ - 2tW_t ||W_t||^{-2}, \quad (1)$$

where the constant t is defined by $t = \tau^{-1} \frac{\lambda ||W_t||^2}{n}$.

The sample set $preD_t$ can be defined as $\{x_-, x_+\}$. The prior knowledge is then $PK_t = preD_t \cup D_t$.

2.2 Selecting the Teaching Samples

Based on the learner conditional distribution over the teaching set, $P_t(y|x)$, we finally select the next sample to show to the student, x_t , using the Expected Error Reduction algorithm - S_{eer} [3]. S_{eer} chooses the teaching sample which would lead to the greatest reduction in the future error of the student over the unobserved samples, D_u :

$$x_t = \operatorname{argmin}_{x_p} \sum_{x_i, y_i \in D_u} (1 - P_t^{+(x_p, y_p)}(y_i|x_i)) \quad (2)$$

Here, $P_t^{+(x_p, y_p)}(y_i|x_i)$ is the estimate of the student's conditional distribution if he was shown x_p and would label it correctly after have seen it.

3 RESULTS

We applied the proposed teaching approach on the problem of classifying a given food as being nutritious or not by looking at its labeling information. The data used to teach was taken from the Portuguese Food Composition Database [1]. To test our approach we simulated learners whose learning model equals a Logistic Regression (the same as assumed by the teacher), but with prior knowledges different between learner and teacher.

We compared our teaching approach against two other, considering 4 conditions: (1) Random, where the samples are randomly selected (non-interactive); (2) FIT (our approach), where the system selects the samples that minimize the expected future error, while inferring the learner's prior knowledge through interactivity on his feature weights vector; (3) FIT_1F, a version of our approach more appropriate for real human users, where only the most important feature is asked to the learner (instead of the complete weights vector); (4) EER, where the system follows the approach proposed in [3], with interactivity on the answers level and a different way of modelling the student that does not attempt to infer its prior knowledge.

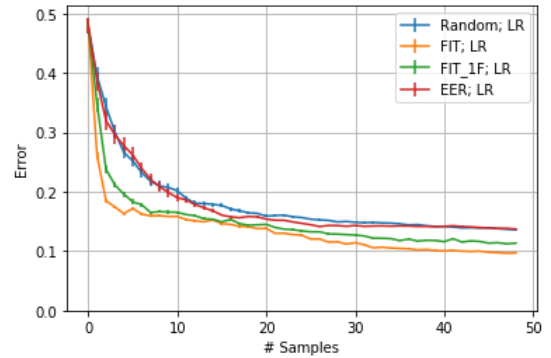


Figure 2: The resulting mean error curves over one hundred runs across conditions.

The results have shown that our approach outperforms the other approaches considered, even when the student gives only the most important feature (Figure 2). All the error curves start in a common state, but the FIT curves (conditions FIT and FIT_1F) converge significantly faster. The differences in the error distributions of both FIT approaches and the other approaches considered are significant right after 1 sample is shown when tested with a statistical Mann-Whitney U test. After 5 samples were shown, the FIT curves are already below a 20% error, while the other two approaches need around 15 samples to reach a similar error.

4 CONCLUSIONS AND FUTURE WORK

In this paper we propose a novel interactive teaching approach to teach classification tasks - FIT - that uses the importance attributed by the student to each feature used to classify a given sample. With this information it can estimate his prior knowledge and, therefore, consider a more realistic model for the student, overcoming the common mismatch between the learner and the teacher's assumptions for him. It then selects better teaching samples that improve the learning rate, as shown by the results obtained in simulations. In the future we would like to perform a study with human users to verify if the results still hold in the real world.

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and the FCT PhD grant with reference SFRH/BD/118006/2016.

REFERENCES

- [1] 2019. PortFIR webpage with the "Food Composition Table" (Version 4.1). <http://portfir.insa.pt/foodcomp/introduction>. Accessed: 2020-03-26.
- [2] Maya Cakmak and Manuel Lopes. 2012. Algorithmic and human teaching of sequential decision tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [3] Edward Johns, Oisín Mac Aodha, and Gabriel J Brostow. 2015. Becoming the expert-interactive multi-class machine teaching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2616–2624.
- [4] Ji Liu and Xiaojin Zhu. 2016. The teaching dimension of linear learners. *The Journal of Machine Learning Research* 17, 1 (2016), 5631–5655.
- [5] Francisco S Melo, Carla Guerra, and Manuel Lopes. 2018. Interactive Optimal Teaching with Unknown Learners. In *IJCAL* 2567–2573.
- [6] Adish Singla, Ilija Bogunovic, G Bartók, A Karbasi, and A Krause. 2014. Near-Optimally Teaching the Crowd to Classify. In *Proceedings of the International Conference on Machine Learning*, Vol. 1. 3.
- [7] Jerry Zhu. 2013. Machine Teaching for Bayesian Learners in the Exponential Family. In *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc., 1905–1913. <https://proceedings.neurips.cc/paper/2013/file/9c01802ddb981e6bcfbec0f0516b8e35-Paper.pdf>
- [8] Xiaojin Zhu. 2015. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [9] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. 2018. An overview of machine teaching. *arXiv preprint arXiv:1801.05927* (2018).