Interactive Teaching with Unknown Classes of Bayesian Learners

Carla Guerra[1,2] , **Francisco S. Melo**[1,2] , **Manuel Lopes**[1,2]
[1]INESC-ID
[2]Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
carla.guerra@gaips.inesc-id.pt, fmelo@inesc-id.pt, manuel.lopes@tecnico.ulisboa.com

Abstract

In this work we empirically explore an interactive approach for machine teaching with classes of students. We use interactivity to overcome the common mismatch between the knowledge the teacher has about the students and the students themselves. We analyze a specific situation where the students learning algorithm is known but the corresponding parameters are not. We focus on the case of Bayesian Gaussian learners, where the lack of knowledge regarding the students parameters significantly deteriorates the performance of machine teaching. With a multi-learner setting we also investigated the best way to consider the class - as a whole or divided in partitions accordingly to the students priors. The results of an user study have shown that, regardless of considering partitions or not, the interactive approach increases the learning performance of the class (reducing the teaching dimension) when compared to the non-interactive approach.

1 Introduction

A significant amount of teaching relies on providing examples. Therefore, the learning efficiency can be greatly improved if the teacher selects the examples that are more informative for each particular learner. A lot of research on education systems can be found in the literature. However, many treat the human learner as a black-box function, not considering the learning model of the student [Patil *et al.*, 2014; Nkambou *et al.*, 2010; Koedinger *et al.*, 1997; Davenport *et al.*, 2012; Clement *et al.*, 2015].

Machine teaching (MT), on the other hand, 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 [Zhu, 2013; Zhu, 2015; Zhu *et al.*, 2018]. Thus, machine teaching treats the human learner as a "transparent box" [Zhu, 2015]. The optimal amount of training activities needed is known as the teaching dimension (TD) of that task. A smaller teaching dimension means less effort required both from the teacher and the student. Thus, minimizing the TD is the ultimate goal of machine teaching.

The main problem with this is its reliance on unrealistic assumptions. MT often assumes that the learner, or the learning algorithm, is completely known. This is a very strong assumption that does not hold in the general case, and certainly not in the case where the learner is a human. [Melo *et al.*, 2018] explicitly addressed the unavoidable mismatch between what the machine teaching system assume about the learner and the learner himself and proposed interactivity as the means to overcome it.

However, that work and most of MT research so far has focused on single learner scenarios. There are not yet many advances in machine teaching applied in a setting where the teacher must teach multiple learners although this is the reality in real-world class-rooms. The students have different backgrounds and prior knowledge, which the teacher must take into consideration when delivering the same lecture to everyone. [Zhu *et al.*, 2017] gave the first steps in order to generalize machine teaching to multiple learners, offering help on how to teach large classes. The authors defined a mini-max teaching criterion to assure the success of the worst student in the class and considered two types of learners: (1) Bayesian learners and (2) linear regression learners. Results have shown that the teaching dimension is higher as class diversity increases. The authors considered also other method where they partitioned the class into sections. They show cases where the optimal partitions allows to minimize the aggregate teaching dimension with no loss of performance on all learners. Curiously, they found that having only one learner per section, i.e. having the most personalized education, is not necessarily the optimal partition. But, once again, all these findings assume perfect knowledge about the learner.

Our goal in this work is to empirically explore the impact of that imperfect knowledge in a classroom scenario with multiple learners and how the interactive approach proposed by [Melo *et al.*, 2018] can help overcoming it. Within this problem we formulated two hypotheses:

- Hypothesis 1: Considering interactivity in the teaching process outperforms other non-interactive approaches. This hypothesis is supported by numerous pedagogical studies found in the literature suggesting that interactivity enhances students' engagement and learning [Alexander, 2018; Beauchamp and Kennewell, 2010; Kennewell and Beauchamp, 2007];

- Hypothesis 2: Dividing the class into partitions accordingly to the students priors and giving one sample per partition makes the learning process faster than considering the class as a whole. This hypothesis was inspired in the work by [Zhu *et al.*, 2017] with machine teaching with multiple learners (but no interactivity).

To confirm this hypotheses we conducted an user study with classes of students. We found that interactivity can be the means to close the gap between the student and teacher parameters. Also, partitioning the whole group into smaller groups, although increasing the effort to the teacher in each run, revealed to improve the overall performance when combined with interactivity.

2 Background

The interactive teaching approach proposed by [Melo *et al.*, 2018] with single learners extends the machine teaching algorithms in order to relax the assumptions on how well the learner is known. Consequently, they model the learner as a standard learning algorithm and the variability across learners is explained by different settings/parameters of the algorithm. The teacher could then estimate the necessary parameters using data from the learner. To obtain such data, the teacher acts as an active learner, whose learning task is to infer the parameters of the learner. They considered learning in the exponential family to illustrate how the proposed approach can be applied. They assumed that the learner has a prior over μ of the form $p_0(\mu) = N(\mu; \mu_0, \sigma_0^2)$. The teacher, on the other hand, assumes a prior $p_0'(\mu) = N(\mu; \mu_0', \sigma_0^2)$.

To teach the optimal mean, μ^*, the teacher provides samples to the learner determined by:

$$x_{n+1} = \frac{\sigma^2}{\sigma_n^2}(\mu^* - \mu_n') + \mu^* \tag{1}$$

The learner then updates its distribution, obtaining the posterior $p_{n+1}(\mu|x_{n+1}) = N(\mu; \mu_{n+1}, \sigma_{n+1}^2)$, with:

$$\sigma_{n+1} = \frac{\sigma^2\sigma_n^2}{\sigma^2 + \sigma_n^2} \tag{2}$$

and

$$\mu_{n+1} = \sigma_{n+1}^2\Big(\frac{\mu_n}{\sigma_n^2} + \frac{x_{n+1}}{\sigma^2}\Big) \tag{3}$$

The interactive process consists of querying the learner for its distribution over the correct hypothesis. Inverting the update process, the teacher can then estimate the original prior of the learner. Although this approach still relies on the strong assumption that the learner is learning from an exponential family, it no longer requires knowledge of the prior. Their results clearly show that the interactive approach is much faster than the classical machine teaching approach.

Following the described approach, we summarize the algorithm we used to compute the teaching set of samples when considering classes of learners instead of a single learner as follows:

Algorithm 1 Interactive Teaching with Classes of Students

Input: Class nth estimation of the mean, μ_n
Parameters: Correct hypothesis $N(\mu^*, \sigma^2)$; Class prior variance, σ_0^2
Output: Next sample, x_{n+1}

```
while Class Mean Error > 0 do
    Query the class for its current estimation for the mean of the distribution ←μ'_n
    Compute the sample to show next, x_{n+1} (using Equation 1)
    Compute the updated class variance, σ_{n+1}, after seeing the new sample (using Equation 2)
end while
return Next sample, x_{n+1}
```

3 User Study

In this section we present an empirical user study created to validate the hypotheses formulated related to our work. Mainly we want to test in a real-world scenario if the interactive approach proposed by [Melo *et al.*, 2018] is still faster when we extend the discussion of interactivity to multi-student settings. This raises several novel questions regarding the teaching process: how to interact with multiple students? How to deal with the individual differences between students? How much does the feedback from one student inform the teacher about the state of the class? Inspired by the work by [Zhu *et al.*, 2017] with multiple learners (but no interactivity), we explored different ways of considering the class - as a whole or partitioned.

3.1 Experimental Design

We used the same artificial problem created in [Melo *et al.*, 2018] where each participant has to estimate the mean monthly rent of an 1-bedroom apartment in a city A in the US. This could be any city, with a very high or a very small mean. In this way we assure that the prior knowledge of the students is low, avoiding it to dominate the information contained in our examples and be too slowly influenced by them. Looking at the learning updates (Equations 2 and 3), weaker priors lead to better learning. In this case, instead of teaching only one student, we considered groups of 10 students at the same time. In each run each student gives an answer (not shared with the rest of the group). After that, we give an example (said to be real) of an 1-bedroom apartment rented in that city to each student (which is, again, not shared with the others). This is repeated for 10 runs with every group. The optimal mean considered was 500 EUR with an assumed standard deviation of 300 EUR.

Given the amount of participants needed in each run simultaneously and to facilitate the interaction process, we developed a simple online interface where each of the participants could remotely send his answers and receive our samples (Figure 1). To note that even though each participant felt like his participation was individual, the samples were shown taking into consideration all the answers given by the group. This interface was developed using a Java server and a Python client.

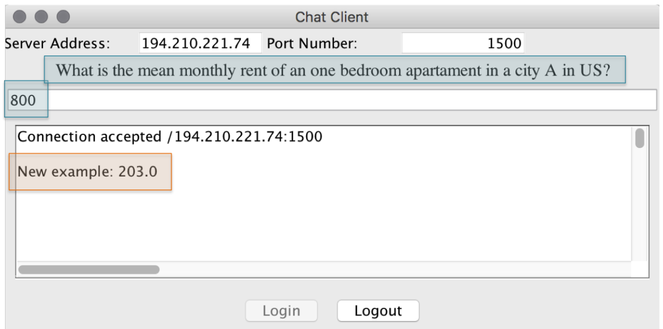
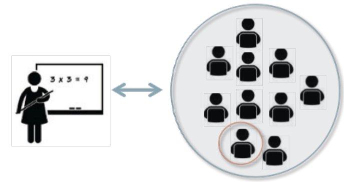


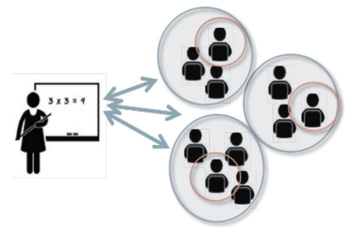
Figure 1: The interface used to send samples and receive responses from each user participating in the group study.

To address the teaching samples to the group we explored 3 different approaches:

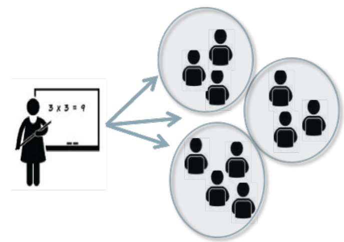
- Condition 1: Interactive Teaching with No Partitions (Figure 2a) - here we teach the whole group at the same time, considering in each iteration one randomly selected answer from the total group as the answer of the class. We then follow the algorithm proposed by [Melo *et al.*, 2018], where we take into account the considered answer of the class in each iteration and use it to calculate the sample showed using the equation given in Equation 1. We show the same sample to every participant. As before, in order to avoid showing a repeated value over and over again each time the participant maintains its estimate, we added a random noise to the sample showed (between -5 EUR and $+5$ EUR) to make it more believable. Also, when the learner estimate is too high and the algorithm calculates too low or even negative examples to show, we converted those values to 100EUR, considering it to be the minimum reasonable value expected for an one-bedroom apartment (this applies to all the conditions);
- Condition 2: Interactive Teaching with Partitions (Figure 2b) - instead of considering the whole-group, here we divide the group into smaller partitions. To do so we first ask each participant to select the interval where his estimate falls. We use that information to aggregate the participants into smaller groups (partitions) accordingly to this first answer. Based on the results of the previous studies, we defined 6 possible partitions (6 intervals where the first estimation can fall: $[0, 400[$, $[400, 500[$, $[500, 600[$, $[600, 700[$, $[700, 800[$ or $[800, +\infty[$). In each iteration and for each partition we consider one randomly selected answer as the answer of that partition. If we have n partitions, we will consider n answers. And we will then show n samples, one per partition. Thus, not every participant sees the same sample as in condition 1 - participants in different partitions see different samples (following Equation 1 and having into account the considered answer in each partition).
- Condition 3: Non-Interactive Teaching with Partitions (Figure 2c) - the group is partitioned as in condition 2. But, in contrast, the samples are presented to the students in each partition following Equation 1 without taking into account the answers given by the students, but rather assuming perfect knowledge about the student parameters. The μ_0 assumed in each partition was defined by the mean value of each partition interval, except for the first and last partitions where we defined μ_0 regarding what better adapts to the data obtained in the previous study by [Melo *et al.*, 2018] ($[0, 400[$: $\mu_0 = 350$, $[400, 500[$: $\mu_0 = 450$, $[500, 600[$: $\mu_0 = 550$, $[600, 700[$: $\mu_0 = 650$, $[700, 800[$: $\mu_0 = 750$, $[800, +\infty[$: $\mu_0 = 1000$).



(a) No partitions - Interactive



(b) Partitions - Interactive



(c) Partitions - Non Interactive

Figure 2: Diagram with the 3 approaches considered.

3.2 Participants

This study was conducted in two universities, and there was a total of 239 participants: 80 participants (8 groups with 10 participants each) were on condition 1 (Interactive Teaching with No Partitions), 79 (7 groups with 10 participants and 1 with 9) on condition 2 (Interactive Teaching with Partitions) and 80 (8 groups with 10 participants) on condition 3 (Non-Interactive Teaching with Partitions). The male percentage was 71%. We detected 14 outliers across the participants (with their first estimate below or above the first or third quartile, respectively, by $1.5\times$ inter-quartile range), which we excluded.

3.3 Results

The results confirmed our two hypothesis regarding the approach with partitions vs. considering the class as whole and also the use of interactivity instead of non-interactive approaches.

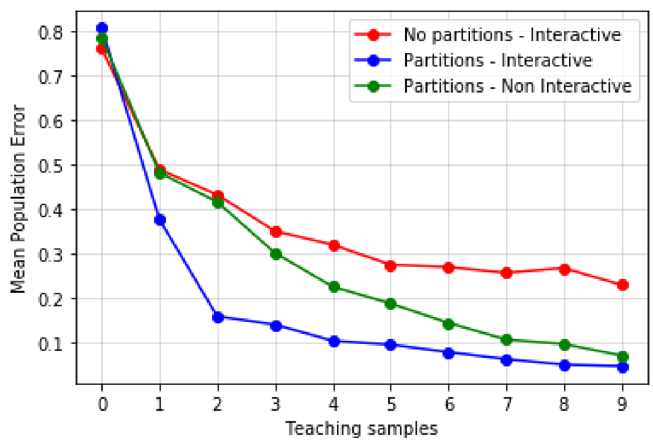


Figure 3: Comparison of the learning performance in the interactive condition when teaching multiple learners.

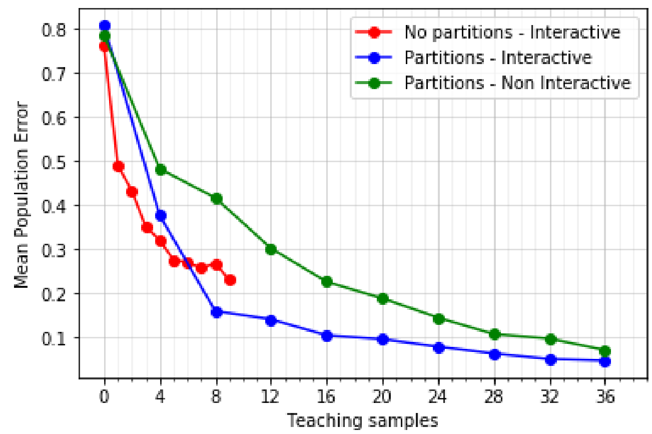


Figure 4: Comparison of the learning performance in the interactive condition when teaching multiple learners, considering the extra cost of teaching more partitions.

Partitions vs. Whole group The first result regards the impact of considering partitions or the group as a whole. As Figure 3 shows, the conditions with partitions seem to have a better performance than the one with no partitions.

However, when a teacher decides to partition the group, his teaching effort (the number of samples needed to teach) increases with the number of partitions considered. To take this into account, we multiplied the teaching samples of the conditions with partitions (blue and green curves) by the average number of partitions in all the groups acquired in each condition (around 4 in both of them) - Figure 4. Including this extra effort, it is not so obvious that the conditions with partitions have a better performance. We could see this is true when comparing the interactive cases - the difference is indeed statistically significant ($p-value = 8.4e-09$) when performing a Mann-Whitney analysis right at the 3rd iteration (where the teacher had shown 2 samples to the group in the no partition approach versus $2 \times 4 = 8$ samples in the conditions with partitions). This difference is still significant ($p-value = 0.04$) if we compare the end of the teaching process of the no partition approach (10^{th} iteration, where 9 samples were given) to the equivalent iteration in terms of teaching effort of the interactive with partitions approach (3^{rd} iteration, where 8 samples on average were given). In contrast, considering the extra effort, the non-interactive condition with partitions is no better than the (interactive) condition with no partitions - the latter one has indeed significantly lower error rates on its last iteration when comparing to the equivalent 3^{rd} iteration (in terms of teaching samples) in the non-interactive with partitions approach ($p-value = 0.03$).

Interactivity vs. Non-interactivity The previously mentioned results show that having partitions is not necessarily better and that one must also consider the factor of using interactivity in the teaching process. Indeed, with or without partitions, the interactive conditions showed the best performance, even when considering the extra teaching effort associated to the partitions (Figures 3 and 4). Performing a Mann-Whitney analysis we found that the interactive approach with partitions is a significantly faster approach. The difference to the non-interactive method is significant right in the 3rd iteration - $p-value = 0.007$) but towards the end (10^{th} iteration) the difference is no longer significantly different ($p-value = 0.19$), meaning that both methods were able to teach the desired mean, but the interactive one was clearly faster.

4 Conclusions

In this work we empirically investigate the impact that the strong assumptions usually made in classical machine teaching approaches can have in the learning performance of classes of students. To explore these issues we conducted an user study with a multi-learner setting. We inspected two ways to consider the class: as a whole or partitioned accordingly to the students priors. We could confirm that, by allowing the teacher to interactively assess the state of the class as proposed by [Melo *et al.*, 2018], the impact of the aforementioned mismatch is significantly mitigated. The results have shown that the interactive teaching approaches (with partitions or not) significantly outperforms the non-interactive alternative. One of our findings is that is better to divide the group into smaller partitions and have interactivity in the teaching process. However, if we can not have both (interactivity and partitions), the interactive approach without partitions performs better than the non-interactive approach with partitions.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

References

[Alexander, 2018] Robin Alexander. Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education*, 33(5):561–598, 2018.

[Beauchamp and Kennewell, 2010] Gary Beauchamp and Steve Kennewell. Interactivity in the classroom and its impact on learning. *Computers & education*, 54(3):759–766, 2010.

[Clement *et al.*, 2015] Benjamin Clement, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, 7(2):20–48, 2015.

[Davenport *et al.*, 2012] Jodi Davenport, Anna Rafferty, Michael Timms, David Yaron, and Michael Karabinos. Chemvlab+: Evaluating a virtual lab tutor for high school chemistry. In *Proceedings of the International Conference of the Learning Sciences (ICLS)*, 2012.

[Kennewell and Beauchamp, 2007] Steve Kennewell and Gary Beauchamp. The features of interactive whiteboards and their influence on learning. *Learning, Media and Technology*, 32(3):227–241, 2007.

[Koedinger *et al.*, 1997] Kenneth Koedinger, John Anderson, William Hadley, and Mary Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 02 1997.

[Melo *et al.*, 2018] Francisco S. Melo, Carla Guerra, and Manuel Lopes. Interactive optimal teaching with unknown learners. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2567–2573, Stockholm, Sweden, 2018.

[Nkambou *et al.*, 2010] Roger Nkambou, Jacqueline Bourdeau, and Valéry Psyché. Building intelligent tutoring systems: An overview. In *Advances in Intelligent Tutoring Systems*, pages 361–375. Springer, 2010.

[Patil *et al.*, 2014] Kaustubh R Patil, Xiaojin Zhu, Łukasz Kopeć, and Bradley C Love. Optimal teaching for limited-capacity human learners. In *Advances in neural information processing systems*, pages 2465–2473, 2014.

[Zhu *et al.*, 2017] Xiaojin Zhu, Ji Liu, and Manuel Lopes. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3588–3594, 2017.

[Zhu *et al.*, 2018] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint:1801.05927*, 2018.

[Zhu, 2013] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems*, pages 1905–1913, 2013.

[Zhu, 2015] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.