A Case for Guided Machine Learning

Florian Westphal[1(✉)], Niklas Lavesson[1,2], and Håkan Grahn[1]

[1] Blekinge Institute of Technology, Karlskrona, Sweden
{florian.westphal,hakan.grahn}@bth.se
[2] Jönköping University, Jönköping, Sweden
niklas.lavesson@ju.se

Abstract. Involving humans in the learning process of a machine learning algorithm can have many advantages ranging from establishing trust into a particular model to added personalization capabilities to reducing labeling efforts. While these approaches are commonly summarized under the term interactive machine learning (iML), no unambiguous definition of iML exists to clearly define this area of research. In this position paper, we discuss the shortcomings of current definitions of iML and propose and define the term guided machine learning (gML) as an alternative.

Keywords: Guided machine learning · Interactive machine learning · Human-in-the-loop · Definition

1 Introduction

With the continuing advances in machine learning, the decisions taken by machine learning algorithms have more and more impact on everyday life. Therefore, it is important that users of these algorithms, as well as people affected by these decisions can understand and trust the used algorithms. One common way to achieve this is interactive machine learning (iML) [12], which interactively involves users in the training process. This can help users to better understand the decisions taken by the machine learning algorithm and therefore increase trust in those algorithms. Furthermore, it enables users to adjust the algorithm's behavior to their needs. Thus, making the benefits of machine learning available to the wider public, leading to a democratization of machine learning.

While characterizations and definitions of iML have been provided in survey papers by Amershi et al. [1], Bertini and Lalanne [2] and Holzinger [10], we argue that all of these definitions are ambiguous to some degree and thus include or exclude more approaches than intended. As an unambiguous definition is important to define a research area clearly and to help identify relevant work easily, we propose a new definition for iML, which avoids the identified ambiguities. Furthermore, we argue that the word *interactive* in iML is unintentionally broad and propose the term *guided machine learning* (gML) for this area instead.

Published by Springer Nature Switzerland AG 2019
A. Holzinger et al. (Eds.): CD-MAKE 2019, LNCS 11713, pp. 353–361, 2019.
https://doi.org/10.1007/978-3-030-29726-8_22

We discuss the issues with current definitions of iML in Sect. 2. In Sect. 3, we propose our definition of gML and examine its implications for other fields of research within machine learning, and in Sect. 4, we summarize the main points of this position paper.

2 Interactive Machine Learning

In the following, we clarify basic machine learning terminology in Sect. 2.1, describe the difficulty to distinguish iML from general machine learning (ML) in Sect. 2.2 and discuss the shortcomings of different attempts to establish this distinction in Sects. 2.3 and 2.4. Furthermore, we argue that the term *interactive* is too broad to describe what is currently considered as iML in Sect. 2.5.

2.1 Machine Learning

Machine learning is concerned with algorithms that learn from data. Mitchell [15] defines this learning as follows:

Definition 1. *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

In order to achieve this learning from experience, a machine learning algorithm builds a model from the training data. Thus, the model encodes the current state of the learner. Since solving tasks from the given class of tasks is based on the model, improvements of the model result in an improved performance.

2.2 Difference to Machine Learning

In order to establish iML as distinct field of research, any definition of iML needs to separate it clearly from existing fields. This is especially true since the name *interactive machine learning* does not in itself separate iML from ML. Based purely on the name, one could define iML as *algorithms that improve their performance at a task through interactions*. Here, it is important to note that this definition does not limit the type, source or target of the interactions. From Definition 1, we can see that improvement through *experience*, for ML, compared to improvement through *interactions*, for iML, is the only difference between this name based iML definition and the definition of ML. However, **this makes iML identical to ML**, since experience presupposes interaction, as we will show in the following.

Experience can only be gained through either practical action or observation. Gaining experience through practical action involves doing something and observing its effect, thus it involves interaction. For example, learning to play Tetris involves playing the game and observing the results of each decision taken; hence, interacting with the Tetris world. **Therefore, gaining experience through practical action clearly requires interaction.**

Furthermore, gaining experience through observation requires this observation to be active. This means that it is not sufficient to gather observational data, but this data needs to be processed to gain experience. However, this processing of observational data requires a certain level of interaction with this data. For example, it is not sufficient to stare at someone playing Tetris to learn how to play, instead hypotheses need to be formed based on current observational data to direct attention to gain more relevant data and to extract relevant observations. **Therefore, gaining experience through observation clearly requires interaction.**

In order to avoid this overly broad definition of iML, common definitions of iML require the presence of a human or non-human interaction partner, which distinguishes iML from ML.

2.3 Human-in-the-Loop

The most common restriction on the interaction partner is to require this partner to be human. For example, Bertini and Lalanne describe iML as follows: *"Interactive Machine Learning is an area of research where the integration of human and machine capabilities is advocated, beyond scope of visual data analysis, as a way to build better computational models out of data. It suggests and promotes an approach where the user can interactively influence the decisions taken by learning algorithms and make refinements where needed."* [2]. This clearly separates iML from ML by requiring a human user to interact with the learning algorithm. However, one potential issue with this approach is that it introduces a certain degree of ambiguity, since **it is unclear if an iML algorithm**, according to this description, **should still be considered as such if the human is replaced with a program simulating human interactions.** This uncertainty is problematic for a definition of iML, since it leaves the classification of an algorithm as iML algorithm up for interpretation. **In this way, it may be possible that important iML approaches are not recognized as such and are overlooked by the iML community.**

2.4 Non-human Agents

One way to avoid this ambiguity caused by requiring human interaction partners is to extend the iML definition to non-human partners. This is done, for example, in Holzinger's definition of iML [10]:

Definition 2. *Interactive machine learning is concerned with algorithms that can interact with agents and can optimize their learning behavior through these interactions, where the agents can also be human.*

While this definition clearly sidesteps the ambiguity caused by limiting iML only to human interaction partners, it is actually even more ambiguous. **The main issue of this definition is that it requires an ambiguous distinction to be made between non-human agents and machine learning**

mechanisms, in order to distinguish iML from ML. In the following, we will illustrate the difficulty to make this distinction with the help of the operation of decision tree pruning as example.

Decision tree pruning is an operation performed on a decision tree to avoid overfitting and to improve the overall performance of the decision tree by removing training data specific subtrees. Han and Cercone [8], for example, propose the DTViz system, which allows human users to interactively construct and prune decision trees. Clearly, this system allowing users to perform tree pruning should be considered as iML system according to Definition 2. One algorithm, which could replace the user in this approach, is the pruning algorithm proposed by Kearns and Mansour [13]. This algorithm determines automatically for a given decision tree which subtrees should be pruned and thus performs the same task as the human user. Therefore, the use of this algorithm instead of a human user should not change the system's classification as iML system. This is the case, since the basic interaction, the system presents the current decision tree and receives pruning decisions from the agent, is still the same. However, the same can be said about the approach proposed by Gelfand et al. [4], which integrates the tree pruning into the tree construction process. While the basic interaction stays the same, the pruning algorithm becomes part of the learning algorithm. Based on this, **one can either argue that this integrated approach should be classified as iML, since the interaction is basically the same as in the case of the DTViz system, or argue that it should not be classified as iML, since no real interaction is taking place, because the pruning algorithm is part of the learning mechanism**. Thus, it is not possible to unambiguously classify the presented pruning approach as iML algorithm.

The presented example illustrates that it can be difficult to distinguish between an agent interacting with a machine learning algorithm and a part of this learning algorithm. However, this distinction is necessary, since otherwise any learning mechanism and thus all of ML could be classified as iML, which would render the definition useless.

2.5 Interactive Learning

Apart from the problem of differentiating iML from ML, **another issue for any definition of iML is that the term *interactive* covers two different scenarios**. On the one hand, the interaction partner has an idea of the task the machine learning algorithm should perform and directs it towards this goal, while on the other hand the interaction partner may not have such a goal and may just interact with the algorithm without clear purpose. Clearly, both scenarios are covered by the term *interactive*, since an agent interacts with the learning algorithm in both cases. However, the former, directed scenario is arguably more interesting and most commonly considered in characterizations of iML. This is made clear, for example, in the previously mentioned description of iML by Bertini and Lalanne [2], as well as by Amershi et al. who state: *"As a result of these rapid interaction cycles, even users with little or no machine-learning expertise can steer machine-learning behaviors through low-cost trial and error*

or focused experimentation with inputs and outputs" [1]. **In order to obtain a focused definition of iML, which reflects this preference, any iML definition needs to explicitly exclude the undirected case.** An alternative solution, i.e., the replacement of the word *interactive* with *guided* will be discussed in the next section.

3 Guided Machine Learning

In this section, we propose *guided machine learning* (gML) as alternative to *interactive machine learning* (iML). In Sect. 3.1, we propose a definition for gML and discuss how this definition addresses the previously raised issues. Furthermore, we review the relationship between gML and other fields of research within machine learning in Sect. 3.2.

3.1 Proposed Definition

In Sect. 2, we have shown that, while it is important to distinguish iML from ML, it is difficult to do without introducing a certain degree of ambiguity. We argue that the ambiguity introduced by allowing non-human agents, as discussed in Sect. 2.4, is worse than when focusing only on human interaction partners. This is the case, since considering non-human agents either requires highly subjective judgments on whether or not to include a proposed algorithm or, if relaxed too much, may lead to including all of ML. Therefore, we propose to focus on human interaction partners.

In order to reduce the ambiguity introduced by limiting our definition to human users, we propose to require the presence of a user interface instead of a user for considering an approach as falling under our definition. In this way, the substitution of a real user with a program simulating user activity does not undermine the definition, while the required presence of a user interface sets a clear boundary between considered approaches and the rest of ML.

As mentioned in Sect. 2.5, the use of the word *interactive* covers two distinct scenarios of which only one is relevant. While this issue could be addressed in the definition, we propose to replace *interactive* with *guided*, since this captures the intended scenario in which a user interacts with a machine learning algorithm in order to improve its performance on a certain task. Therefore, we propose the term *guided machine learning* and define learning through guidance similar to Mitchell's definition for learning [15] as follows:

Definition 3. *A computer program is said to learn through guidance G from a human H with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves through actions performed by H, given that these actions are dependent on the program's current state and aim to achieve such an improvement.*

This definition captures the guidance aspect by requiring the human user to perform actions, which improve the performance of the machine learning

algorithm, with the goal to achieve this improvement. Furthermore, it captures the interactiveness of this guidance process by requiring the user's actions to be based on the current state of the algorithm. This state can be presented to the user directly in form of the current model or indirectly in form of the current performance on the task. In this way, it is possible for the user to iteratively provide guidance to the algorithm throughout the learning process. With this definition of learning from guidance, we can now define gML as follows:

Definition 4. *Guided machine learning is concerned with the design of interfaces for human users, which enable a user to perform actions, which allow a computer program to learn from this guidance.*

This definition clearly separates gML from ML by requiring the machine learning algorithm to learn from data collected iteratively through a user interface. It avoids to be dependent on the deployment context by not requiring the presence of a human or the presence of guidance actions, but instead focuses on the presence of an interface, which would allow a human to perform such actions.

One potential issue of the proposed definition is that ascertaining the required properties of the provided interface may be subjective. However, we would argue that checking if a provided interface presents the machine learning algorithm's current state in a way aimed at human understanding and allows actions to be taken in response to the presented information, which can improve the algorithm's performance, is reasonably objective.

3.2 Definition Consequences

In the following, we will discuss the relationship between gML, as defined in the previous section, and other research areas in ML.

Supervised Learning. The key defining feature of supervised learning is that supervised machine learning algorithms learn to perform their task from a labeled training data set. While these labels are generally provided by humans, supervised learning cannot be considered as gML, since data sets in supervised learning are available in full at the beginning of the training. In contrast, in gML, training labels are provided throughout the training process via a user interface and depend on the model's current state. One example for such guidance through the provision of more labeled data is the approach to pixel classification proposed by Fails and Olsen [3]. In their approach, Fails and Olsen allow users to view the current pixel labeling performance of a classifier and to provide more pixel labels to improve the performance. Apart from the plain pixel labels, the provided information also contains the implicit knowledge that the newly labeled pixels are more relevant to the learning process than a randomly selected set of pixels, which would be chosen by a supervised learning approach.

Unsupervised Learning. The overall goal in unsupervised learning is to extract information from a given data set without any form of human intervention or guidance. Thus, unsupervised learning is clearly unrelated to gML.

Reinforcement Learning. Reinforcement learning is concerned with learning how to act in a given situation not from a prescribed ideal action, as in supervised learning, but instead from a cumulative reward signal [18]. In general, reinforcement learning approaches cannot be considered as part of gML, since reinforcement learning requires only the collection of a reward signal, which does not necessarily presuppose a user interface. However, reinforcement learning ideas can be used in gML approaches, such as in the approach by Thomaz and Breazeal [19], which allows users to give reward signals to a virtual robot, in order to teach it to bake a cake. The reward signal provided for certain behavior depends on the robot's current state, since its actions are determined by it.

Active Learning. In active learning [17], an active learning algorithm selects the training samples, which should be used to train a machine learning algorithm based on the learning algorithm's current state. While this conditionality on the learner's state leads to a close relationship between active learning and gML, not all active learning approaches are also gML approaches, since active learning does not require a user interface. However, active learning is useful for designing user interfaces, since it can reduce the amount of data to be presented to a user. For example, the approach proposed by Heimerl et al. [9] for classifying text documents as relevant or irrelevant uses active learning to solicit user labels.

Adversarial Training. In adversarial training scenarios, learning is facilitated by the competition between two or more learners. Typical examples for these scenarios are Samuel's checkers program [16], which trained an algorithm to play checkers by playing against itself, as well as generative adversarial networks (GANs) [7], which train a generator to generate samples from a target distribution together with a discriminator for distinguishing between generated and real samples. While learning using adversarial training proceeds in an iterative feedback loop between the adversaries, it is clearly different from gML. This is the case, not only because most adversarial training does not use humans as adversary, which would be required for gML, but also because gML does not assume a competition between the user and the algorithm. In contrast, the guidance aspect requires the user to perform actions with the aim to improve the learner's performance.

Explainable Machine Learning. The main goal of explainable machine learning is to make decisions taken by an ML algorithm transparent, understandable and explainable [6]. This can be achieved either through post-hoc explanations, which are generated on demand for a particular decision, or through ante-hoc explanations, which arise naturally from the used model [11]. While explainable ML does clearly not belong to gML, it is an important aspect in the interface design for gML approaches. In particular ante-hoc systems are useful for gML, since they are directly interpretable by users and should therefore make it easier for them to guide the learning process. This connection between explainable ML

is even clearer in the concept of causability [11], which requires the provided explanation to reach a certain causal understandability. This is interesting for gML, since a causal explanation of a taken decision should enable users to guide the learning process more efficiently and more easily.

Machine Learning Environments. While not directly a research area in ML, machine learning environments, such as Weka[1], may appear to be closely related to gML, since they provide a user interface, which can be used to choose learning algorithm and model hyperparameters, which can improve the algorithm's performance on a task. However, the difference between those environments and gML approaches is that they normally rebuild the previous model from scratch with the newly chosen hyperparameters. This is different from gML, which assumes an update rather than a rebuild of the model. One approach, which blurs this line between machine learning environments and gML is human-guided machine learning (HGML) [5], which allows users to interact with an automated ML (AutoML) system. These interactions can be concerned with the input data, for example in form of feature or instance selection, with the model development, such as model selection or parameter settings, or with the model interpretation, for example in form of model assessment or parameter comparison. Such a system could be considered as a gML approach, if the user actions are used as input to teach a meta-learner to find a suitable configuration for the AutoML system. One other approach, which may bridge the gap between gML and machine learning environments is explanatory debugging, as proposed by Kulesza et al. [14]. The main idea of explanatory debugging is to provide users with an explanation of the algorithm's current performance, which can help them to modify the model accordingly.

4 Summary

In this paper, we have discussed various issues of existing descriptions and definitions of iML. We have argued that iML needs to be defined based on the presence of an interaction partner to distinguish it from general machine learning. However, we have also shown that requiring the presence of a human or non-human interaction partner leads to certain ambiguities. Furthermore, we have pointed out that the word *interactive* in iML may lead to the inclusion of approaches commonly not considered as part of iML. We have addressed these issues by proposing *guided machine learning* and defining it in a way, which avoids the identified sources of ambiguity.

Acknowledgements. The authors would like to thank Huynh Khanh Vi Tran for valuable discussions about possible gML definitions, as well as the anonymous reviewers for their useful comments.

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

[1] https://www.cs.waikato.ac.nz/ml/weka/.

References

1. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. AI Mag. **35**(4), 105–120 (2014)
2. Bertini, E., Lalanne, D.: Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In: Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, pp. 12–20. ACM (2009)
3. Fails, J.A., Olsen Jr., D.R.: Interactive machine learning. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, pp. 39–45. ACM (2003)
4. Gelfand, S.B., Ravishankar, C.S., Delp, E.J.: An iterative growing and pruning algorithm for classification tree design. In: Conference Proceedings, IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 818–823 (1989). https://doi.org/10.1109/ICSMC.1989.71407
5. Gil, Y., et al.: Towards human-guided machine learning. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 614–624. ACM (2019)
6. Goebel, R., et al.: Explainable AI: the new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
8. Han, J., Cercone, N.: Interactive construction of decision trees. In: Cheung, D., Williams, G.J., Li, Q. (eds.) PAKDD 2001. LNCS (LNAI), vol. 2035, pp. 575–580. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45357-1_61
9. Heimerl, F., Koch, S., Bosch, H., Ertl, T.: Visual classifier training for text document retrieval. IEEE Trans. Visual Comput. Graphics **18**(12), 2839–2848 (2012)
10. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. **3**(2), 119–131 (2016)
11. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisciplinary Rev. Data Min. Knowl. Discovery, e1312 (2019)
12. Holzinger, A., et al.: Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl. Intell., 1–14 (2018)
13. Kearns, M.J., Mansour, Y.: A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998, pp. 269–277. Morgan Kaufmann Publishers Inc. (1998). http://dl.acm.org/citation.cfm?id=645527.657457
14. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 126–137. ACM (2015)
15. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
16. Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM J. Res. Dev. **3**(3), 210–229 (1959). https://doi.org/10.1147/rd.33.0210
17. Settles, B.: Active learning. Synth. Lect. Artif. Intell. Mach. Learn. **6**(1), 1–114 (2012)
18. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
19. Thomaz, A.L., Breazeal, C.: Teachable robots: understanding human teaching behavior to build more effective robot learners. Artif. Intell. **172**(6–7), 716–737 (2007)