Decentralized Machine Learning under Constraints

Aurélien Bellet
aurelien.bellet@inria.fr

Marc Tommasi
marc.tommasi@univ-lille3.fr

1 Introduction and general objectives

We are producing ever growing amounts of digital data, from sources as diverse as the collaborative and semantic Web, large-scale social networks, or mobile devices. This flood of information has recently spawned a new economy of both large players and nascent start-ups that seek to extract value from the above sources using *Machine learning* (ML) techniques.

To tackle this information deluge, most of these companies rely today on centralized or tightly coupled ML systems hosted in data centers or in the cloud. This is problematic as this concentration poses strong risks to the privacy of users, and limits the scope of ML applications to tightly integrated datasets under unified learning models. They also require oversized infrastructures for storage, computation and communication without relying on the increasing capabilities of modern small devices and local networks.

To address these limitations, we propose to follow an alternative approach, inspired by peer-to-peer networks, in which each user controls its own system. In this setting, the privacy of users can be better preserved and the usage of their devices more effectively leveraged, albeit under new constraints (for instance on communication). This project aims to study the challenges raised by this approach from a machine learning point of view. We will address both theoretical and practical directions, and expect to provide fundamental tools and concepts to revisit the machine learning paradigm by learning decentralized models on local datasets under various constraints on communication, energy, etc.

2 State of the art

Distributed learning has emerged as a key research topic to address the challenges posed by big data [2]. Substantial efforts have recently gone into new paradigms and architectures (Hadoop, MapReduce, Spark, GraphLab, ...) to help develop distributed versions of traditional data mining and machine learning algorithms. From a theoretical point of view, recent works have revisited fundamental learning techniques in a distributed context while taking into account communication costs or privacy [4, 1, 3, 6]. These works usually make the following assumptions: (i) they consider a single learning task (one seeks the model minimizing the loss over a global dataset, as in a centralized setting); (ii) the peers (machines) are part of a fixed network with a known and simple topology (typically a master node coordinating slave nodes); and (iii) the distribution of data in the network is fully controlled. These assumptions stand in stark contrast to *decentralized learning* in which data are "naturally" distributed among peers pursuing different learning objectives within a complex

open network, a situation which has been much less studied. We precisely aim at investigating this research gap, paying particular attention to communication costs.

3 Research program

We will first focus on the problem of decentralized model propagation: we assume that some users (nodes) in the network have learned a model from their local data, and we would like to propagate these models to other similar nodes through the graph. In a preliminary work [5], we have started to adapt a label propagation algorithm to the decentralized setting and to evaluate its performance for the propagation of models that can be averaged in a straightforward way (such as linear classifiers). This work has opened many important questions. For other types of models (e.g., decision trees, nonparametric models), one should carefully define how the information is propagated, and the influence of propagation on structured prediction models should also investigated. From a theoretical point of view, we are particularly interested in proving convergence rates and estimating the robustness of these approaches. Connections with on-line learning will also be explored.

We will then extend these techniques and theoretical results to go from propagating pre-trained models, as outlined above, to decentralized methods which simultaneously learn and propagate through the network. In this more demanding setting, it is necessary to consider tight constraints on resources (communication, energy usage, etc.) in the design of the learning algorithms to make them efficient enough for deployment in real networks and applications. How to integrate such constraints in the decentralized learning objective is an important issue. The decentralized paradigm also gives rise to the question of the propagation of constraints derived from domain knowledge, and the adaptation towards personalized learning. A last type of constraints comes from the fact that in many practical applications, users cannot be active at all times (for instance, their device is turned off or without a Wi-Fi connection). This motivates the design of robust learning methods that can deal with the absence of some peers during communication rounds, for instance by using redundancy or involving only a subset of users at each step.

4 Timeline

The candidate will be required to have a solid background in machine learning, statistics and algorithms. In the first 2 months, the student will study the state of the art and the relevant literature to investigate the ideas outlined above. We aim at finishing all research in 31 months to leave 3 months for completing all dissemination efforts (including the dissertation). Concrete applications, such as decentralized recommendation, will be pursued continuously throughout the doctoral research to provide inspiration and ensure applicability of the results.

5 Impact

This project has the potential to deliver a substantial impact. Scientifically, we explore a novel framework for machine learning which is in line with the current trends towards big data, distributed algorithms and privacy-friendliness in machine learning. We can thus expect a high interest from the community. From the socio-economic point of view, the project proposes to lay the foundations of a new approach to personalized on-line services, offering an alternative to the prevalent model

promoted by big companies. We plan to strengthen this economic impact through our existing collaborations with relevant companies (see Section 6).

Relevance to the call priorities This project fits very well with the priorities of the CPER 2015-2020 Data, in particular the axis "Intelligence des données et des connaissances" but also, to some extent, the axis "Internet des objets".

6 Research environment

This project will be carried out in the MAGNET Team[1] at INRIA Lille, in which all required competences are present. In particular, Marc Tommasi did research in several areas of graph-based machine learning, while Aurélien Bellet has worked on distributed learning and decentralized gossip protocols. They have recently co-supervised a preparatory Master project on decentralized learning, leading to a paper currently under review [5].

This project will also stimulate existing and emerging collaborations with other research groups on themes at the intersection between machine learning, distributed systems and privacy. In particular, MAGNET has collaborations on this topic with the MLIA team at LIP6 and the ASAP team at INRIA Rennes.

Industrial contact include snips (a startup developing privacy-friendly personal assistants for smartphones) and Mediego (an INRIA spin-off working on decentralized recommenders systems).

References

[1] M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. In *COLT*, 2012.

[2] R. Bekkerman, M. Bilenko, and J. Langford. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.

[3] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. In *NIPS*, 2012.

[4] S. Lattanzi, B. Moseley, S. Suri, and S. Vassilvitskii. Filtering: a method for solving graph problems in mapreduce. In *SPAA*, pages 85–94, 2011.

[5] P. Vanhaesebrouck, A. Bellet, and M. Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. Research report, INRIA Lille, Oct. 2016.

[6] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, 2013.

[1] https://team.inria.fr/magnet/