Robust and light-weight overlay management for decentralized learning

Contact: François Taïani (francois.taiani@irisa.fr) – Group ASAP - IRISA / Inria Rennes

June 1, 2017

1 Context: Big Data, Machine Learning, and Decentralization

A growing number of companies are extracting value from the digital data produced by our modern society using *Machine learning* (ML) techniques. Most of these companies rely today on centralized or tightly coupled ML systems hosted in data centers or in the cloud [5, 6]. 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.

To address these limitations, this PhD proposes to explore an alternative approach inspired by peer-to-peer networks in which users control their own system, and only exchange a limited amount of information to construct local machine learning models. This strategy is more amenable to preserving user privacy, and respecting the constraints possibly imposed on sensitive data-sets (such as health records, or personal financial data), and holds the potential for highly scalable and robust learning systems. This project aims to study the challenges raised by this strategy in terms of distribution and overlay management.

2 Research Objectives

Ideally, a decentralized machine learning system should deliver the best learning at a minimal costs (for instance in order to be able to execute on constrained personal devices) while providing a high level of privacy protection. These different goals are inherently in tension, and one of the PhD's aims will be to explore to which extend they can be balanced using techniques borrowed from distributed computing and machine learning.

One of the PhD's starting point is to take inspiration from the existing work on highly scalable decentralized mechanisms, such as epidemic (aka gossip) protocols [1, 4] and self-organizing overlays [2, 3, 8]. These systems are fully decentralized in that they do not rely on any central point of coordination. Instead, each participating machine (also called peer or node) only possesses a partial knowledge of the rest of the system, and interact with a small number of other peers. These systems use stochastic interactions to overcome node and network failures, while delivering a high level of performance at scale.

Our objective is to investigate how machine learning could execute on such decentralized systems. A machine learning task can frequently be expressed as an optimization problem, in which the optimized "variable" is a model capturing the relationship between the inputs and outputs of the task [7]. This project assumes that the nodes of a decentralized system each have access to a part of the data to be learned (e.g. a user's preferences, of a hospital's records), and wish to solve related learning tasks. The key challenge consists in deciding which data should be exchanged by whom in order to achieve a given level of *learning quality*, *resource consumption*, and *privacy protection*.

To achieve this vision, we envisage in particular to explore in the context of this PhD the following two lines of research :

- We would like to study how biased decentralized sampling techniques might be able to rapidly bootstrap learning tasks while avoiding a broad exploration of the set of peers.
- In a second phase, we would like to explore how self-organizing overlays must be adapted to allow learning peers to rapidly and efficiently identify and contact desirable other peers for their learning tasks. We expect in particular that existing protocol will need to be adapted to insure that the "routing field" on which these protocols are based is sufficiently continuous and strongly connected to ensure the convergence of the learning process.

The PhD is funded by ANR (French Research Agency) within the PAMELA research project on decentralized learning, and will take place in collaboration with the project's partners, and more particularly the Inria Teams MAGNET from Lille[1], and CIDRE from Rennes[2].

3 Candidate profile

We are looking for applicants with a strong background in either distributed systems or/and machine learning, and a strong interest in developing their skills and expertise in both areas. Theory and algorithms as well as design and implementation considerations are of importance in this thesis, and therefore a good theoretical background but also the ability to prototype and validate results in practice are of importance.

4 Application

Applicants should apply, in the first instance, by sending electronically a CV (up to two A4 pages), a short statement outlining their research interests and motivation for a embarking on a Ph.D. (up to half an A4 page), a recent grade transcript (when applicable), and the names and addresses of at least two academic referees to François Taïani (francois.taiani@univ-rennes1.fr).

A rolling deadline applies.

References

[1] X. Bai, M. Bertier, R. Guerraoui, A.-M. Kermarrec, and V. Leroy. Gossiping personalized queries. In *EDBT'2010*, 2010.

[2] Marin Bertier, Davide Frey, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy. The gossple anonymous social network. In *Middleware'2010*, 2010.

[3] Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. T-man: Gossip-based fast overlay topology construction. *Comp. Netw.*, 53(13), 2009.

[4] Márk Jelasity, Spyros Voulgaris, Rachid Guerraoui, Anne-Marie Kermarrec, and Maarten van Steen. Gossip-based peer sampling. *ACM TOCS*, 25, 2007.

[5] Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th USENIX Symp. on Op. Sys. Design and Impl. (OSDI 14)*, pages 583–598, Broomfield, CO, October 2014. USENIX Association.

[6] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Learning continuous-time information diffusion model for social behavioral data analysis. In *ACML*, ACML '09, pages 322–337, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized collaborative learning of personalized models over networks. *CoRR*, abs/1610.05202, 2016.

[8] Spyros Voulgaris and Maarten van Steen. Epidemic-style Management of Semantic Overlays for Content-Based Searching. In *Eur. Conf. on Par. and Dist. Computing (EuroPar)*, pages 1143–1152, 2005.

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

[2] `http://www.rennes.supelec.fr/ren/rd/cidre/`