Exascale Matrix Factorization: Using Supercomputers and Machine Learning for Drug Discovery

Tom Vander Aa <tom.vanderaa@imec.be>, ExaScience Life Lab at imec, Belgium
Xiangju Qin, xiangju.qin@aalto.fi, Aalto University, Finland
João Neto, joaomlneto@tecnico.ulisboa.pt, INESC-ID, Portugal

In the ExCAPE European funded project we investigated the power of supercomputers to speed up drug discovery using machine learning. One of the machine learning algorithms studied in ExCAPE is Matrix factorization (MF). Matrix Factorization is a core machine learning technique for applications of collaborative filtering, such as recommender systems. In drug discovery it can used to predict the interaction between chemical compounds and protein targets. Known interactions are stored in a data matrix called Y. This matrix Y is factorized into a product of two matrices, such that $Y \approx X \times W$. The main task in such applications is to predict unobserved elements of a partially observed data matrix.

Bayesian matrix factorization (BMF) formulates the matrix factorization task as a probabilistic model, with Bayesian inference conducted on the unknown matrices X and W. Advantages often associated with BMF include robustness to over-fitting and improved predictive accuracy, as well as flexible utilization of prior knowledge and side-data. Finally, for application domains such as drug discovery, the ability of the Bayesian approach to quantify uncertainty in predictions is of crucial importance.

Yet BPMF is more computationally intensive and thus more challenging to implement for large datasets. Therefore, a high-performance parallel implementation of BPMF that is suitable for large-scale distributed systems was crucially needed. This implementation was developed and optimized during the ExCAPE project and was instrumental for the pharmaceutical partners of the project. It allowed them to discover new insights in compound-protein interaction thanks to the large-scale models built on datasets that were previously intractable to be processed.

In the PraceDays19 talk we will present the ExCAPE project, the BPMF technique, how the HPC infrastructure and the HPC implementations were crucial to reach insights and how these insights helped the pharma industry in their drug discovery process. We will also explain what research we are currently doing in the EPEEC project to improve scaling of Matrix Factorization even further, and how PRACE resources are helping us reach all of these goals.

Acknowledgments

This work has been supported by the EU H2020 FET-HPC projects EPEEC (contract #801051), ExCAPE (contract #671555).

References

- ExCAPE: http://excape-h2020.eu
- EPEEC: https://epeec-project.eu/
- BPMF: https://github.com/ExaScience/bpmf