Universiteit Leiden
The Netherlands

Massively collaborative machine learning

Rijn, J.N. van

1 Introduction

1.1 Introduction

We are surrounded by data. On a daily basis, we are confronted by many forms of it. Companies try to spread their commercials by means of billboards, commercials and online advertisements. We have instant access to our friends' social lives using services as Facebook and Twitter, and we can obtain information about countless topics of interest by means of websites such as Wikipedia. In most cases, this is a double-edged sword. Companies and governments also collect information about us. For example, most websites store information about our browsing behaviour, banks know most about our financial transactions, and telecom providers even have access to our exact whereabouts, as our GPS coordinates are shared by our mobile phones.

Data is also gathered for scientific purposes. Large sensor networks and telescopes measure complex processes, happening around us on Earth or throughout the Universe. Any estimation of the amount of data that is being produced, transferred and gathered would be pointless, as it will be outdated some moments after publication.

All this data is valuable for the information, knowledge and eventually wisdom we could obtain from it. We could identify fraudulent transactions based on financial data, develop new medicines based on clinical data, or locate extraterrestrial life based on telescope data. This process is called *learning*. The scientific community has created many techniques for analysing and processing data. A traditional scientific tasks is modelling, where the aim is to describe the data in a simplified way, in order to learn something from it. Many data modelling techniques have been developed, based on various intuitions and assumptions. This area of research is called *Machine Learning*.

However, all data is different. For example, data about clinical trials is typically

very sparse, but well-structured, whereas telescopes gather large amounts of data, albeit initially unstructured. We cannot assume that there is one algorithm that works for all sorts of data. Each algorithm has its own type of expertise. We have only little knowledge about which algorithms work well on what data.

The field of Machine Learning contains many challenging aspects. The data itself is often big, describing a complex concept. Algorithms are complex computer programs, containing many lines of code. In order to study the interplay between these two, we need data about the data and the algorithms. This data is called *meta-data*, and learning about the learning process itself is called *meta-learning*. It is possible to gain knowledge about the learning process when there is sufficient meta-data. Some effort has been devoted to building a large repository of this experimental data, called the 'open experiment database' [153]. It contains a large amount of publicly available Machine Learning results. This way, existing experimental data can be used to answer new research questions. Although this has proven extremely useful, there is still room for improvement. For example, sharing experiments was difficult: while all experimental data was accessible to the public, contributing new results towards the experiment database was only practically possible for a small circle of researchers. Furthermore, sensibly defining the types of meta-data that are being stored would expand the range of information and knowledge that can be obtained from the data. For example, storing all evaluation measures per cross-validation fold enables statistical analysis on the gathered data, and storing the individual predictions of the algorithms enables instance-level analysis. Our aim is to build upon the existing work of experiment databases, and demonstrate new opportunities for Machine Learning and meta-learning.

Our contributions are the following. We have developed an online, open experiment database, called 'OpenML'. This enables researchers to freely share their data, algorithms and empirical results. This significantly scales up the size of typical machine learning and meta-learning studies. Implementing algorithms and modelling data are both time-intensive tasks. Instead of setting up the experiments themselves, researchers can now simply look up the results by querying the database, covering a much larger set of experiments. OpenML automatically indexes and organizes the uploaded meta-data, allowing researchers to investigate common questions about the performance of algorithms, such as which parameters are important to tune, what the effect is of a given data property on the performance of algorithms, or which algorithm works best on a certain type of data.

We have demonstrated the effectiveness of this collaborative approach to meta-learning with two large-scale studies that were not practically feasible before. The first study covers the data stream setting, which contains some challenging real-world aspects: large amounts of data need to be processed at high speed and learned models

can become outdated. In this work, we created a novel approach that, while processing the stream, dynamically changes the modelling algorithm when another algorithm seemed more appropriate. The study covered 60 data streams, which to the best of our knowledge, is the largest meta-learning study in the data stream literature to date.

The second study covers a conventional meta-learning task, where the goal is to find an algorithm that adequately models the dataset. However, it is also important to find that algorithm as fast as possible. Indeed, whenever such an algorithm is recommended, it can be tested (e.g., using cross-validation) and if its performance is not sufficient, another one can be tried, but this can be a very slow process. This study showed that there are techniques that trade off performance and run time. If one is willing to settle for an algorithm that is almost as good as the absolute best algorithm for that dataset, the run time can be decreased by orders of magnitude.

The remainder of this thesis is organised as follows. Chapter 2 introduces some basic aspects about Machine Learning, and introduces some well-known model types. Chapter 2 surveys common meta-learning techniques. It approaches meta-learning both from a learning and a search perspective. Chapter 4 describes the online experiment database on which we collect experimental results. Chapter 5 describes the first study that demonstrates the use of OpenML for data streams. Chapter 6 describes a second study that shows how meta-learning techniques can trade off accuracy and run time. Chapter 7 concludes and points to future work.

1.2 Publications

The different chapters of this thesis are based on the following peer-reviewed publications:

- J. N. van Rijn, B. Bischl, L. Torgo, B. Gao, V. Umaashankar, S. Fischer, P. Winter, B. Wiswedel, M. R. Berthold, and J. Vanschoren. OpenML: A Collaborative Science Platform. In *Machine Learning and Knowledge Discovery in Databases*, pages 645–649. Springer, 2013 (Chapter 4)

- J. N. Van Rijn, V. Umaashankar, S. Fischer, B. Bischl, L. Torgo, B. Gao, P. Winter, B. Wiswedel, M. R. Berthold, and J. Vanschoren. A RapidMiner extension for Open Machine Learning. In *RapidMiner Community Meeting and Conference*, pages 59–70, 2013 (Chapter 4)

- J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren. Algorithm Selection on Data Streams. In *Discovery Science*, volume 8777 of *Lecture Notes in Computer Science*, pages 325–336. Springer, 2014 (Chapter 5)

- J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren. Towards Meta-learning over Data Streams. In J. Vanschoren, P. Brazdil, C. Soares, and L. Kotthoff, editors, *Proceedings of the 2014 International Workshop on Meta-learning and Algorithm Selection (MetaSel)*, number 1201 in CEUR Workshop Proceedings, pages 37–38, Aachen, 2014 (Chapter 5)

- J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014 (Chapter 4)

- J. N. van Rijn, S. M. Abdulrahman, P. Brazdil, and J. Vanschoren. Fast Algorithm Selection using Learning Curves. In *Advances in Intelligent Data Analysis XIV*, pages 298–309. Springer, 2015 (Chapter 6)

- J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren. Case Study on Bagging Stable Classifiers for Data Streams. In *Proceedings of the 24th Belgian-Dutch Conference on Machine Learning (BeNeLearn 2015)*, 6 pages, 2015 (Chapter 5)

- J. N. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren. Having a Blast: Meta-Learning and Heterogeneous Ensembles for Data Streams. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 1003–1008. IEEE, 2015 (Chapter 5)

- J. N. van Rijn and J. Vanschoren. Sharing RapidMiner Workflows and Experiments with OpenML. In J. Vanschoren, P. Brazdil, C. Giraud-Carrier, and L. Kotthoff, editors, *Proceedings of the 2015 International Workshop on Meta-Learning and Algorithm Selection (MetaSel)*, number 1455 in CEUR Workshop Proceedings, pages 93–103, Aachen, 2015 (Chapter 4)

- J. Vanschoren, J. N. van Rijn, and B. Bischl. Taking machine learning research online with OpenML. In *Proceedings of the 4th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 1–4. JLMR.org, 2015 (Chapter 4)

A full list of publications by the author can be found on page 155 of this thesis.