EPSRC CDT in Distributed Algorithms

PhD Project: Scalable Online Machine Learning

University of Liverpool

PhD Student: Andrew Millard
Project Partner: GCHQ

Supervisors:
Simon Maskell - University of Liverpool
Simon Goodchild, STFC Hartree

Project Description

I will be co-supervised and work alongside GCHQ and the project relates to extending the state-of-the-art to enable machine learning to fully capitalise on the information present in never-ending data streams. The additional data that arrives over time contains information that should facilitate improved machine learning. Not using this information gives rise to consistent yet surprising errors: this typically occurs when the training data is small relative to the algorithm's empirical experience. Concept drift can also occur: the passage of time also provides scope for the phenomena that give rise to the data to change. The result of concept drift is that, even if the phenomena of interest do not change, because the statistical environment changes, the performance of the machine learning is prone to degrading. Since the quantity of historic data is ever growing, given finite data storage and computational resources, innovative techniques are needed to summarise the information present in data and currently pertinent without requiring all the raw data ever received to be stored.

The proposed solution involves three novel components. First, to reduce the storage and computation that would otherwise be required, the pertinent data received up to the current time will be summarised in an adaptive tree-based data structure. This definition of this data structure will build on previous work on Approximate Bayesian Computation and involve approximating the information present in the raw data with summaries. Second, to ensure concept drift is catered for, these summaries will explicitly relate to the time-derivatives of the parameters that the machine learning is attempting to estimate. Finally, to maximise performance, previous related work involving variational inference, which will be extended to consider the aforementioned data structures, will also be adapted to consider numerical Bayesian inference.

The approach will be applied to real-world datasets involving combinations of: near-constant parameters for which concept drift is not relevant (e.g. related to rare events of interest); parameters that fluctuate smoothly over long timescales (e.g. diffusive spread of memes); sudden shifts in concepts (e.g. new memes appearing). Such datasets are anticipated to involve large and continually growing text corpuses (e.g. social media).

Go to the EPSRC CDT In Distributed Algorithms website.