Scalable Online Machine Learning

Joshua Murphy, Supervised by Professor Simon Maskell and Dr. Simon Goodchild (STFC Hartree Centre)

EPSRC Centre for Doctoral Training in Distributed Algorithms, University of Liverpool, Liverpool, UK

Background

As the worldwide volume of data grows, it becomes increasingly important to fully capitalise on the information present in never-ending data streams to avoid the consistent errors in algorithmic predictions that result when the training data is small relative to the algorithm's experience. Another problem that can occur is concept drift whereby the statistical environment varies over time which can degrade the accuracy of the predictive model. Current hardware is unable to store all incoming data which is another issue to consider.

My project partner GCHQ is interested in online monitoring, particularly in models that maintain a robust decision boundary which is good at filtering out data that is not of interest, thus preventing false alarms.

Aim: To create a model that can handle a constant stream of data which also maintains performance over time in a changing statistical environment.

Local Optima

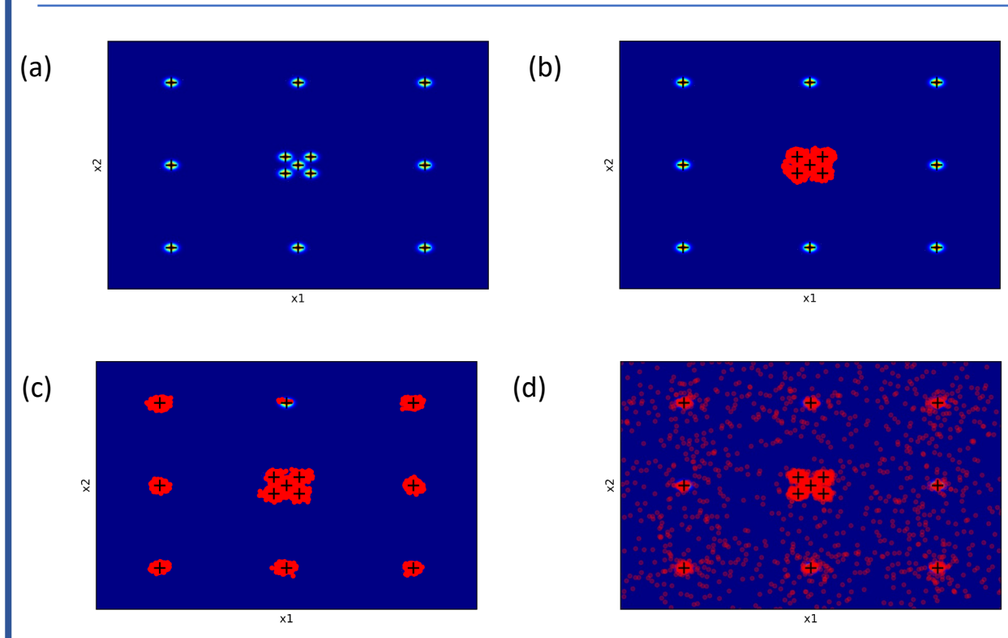

Fig. 1 (a) Probability density plot of a 2D Gaussian mixture model (GMM) with 13 equal weight modes. Modes are shown by crosses and samples by red dots. (b) Standard MCMC. (c) MCMC with mode jumping proposal. (d) SMC sampler with mode jumping proposal. Lower weighted samples are more transparent.

Monte Carlo methods are an effective sampling approach to performing Bayesian inference but can suffer on distributions with multiple local optima and saddle points. Using a mixture of proposal kernels in Markov Chain Monte Carlo (MCMC), with some proposals to sufficiently sample the characteristics of local modes and others to explore the wider space ('mode jumping proposals'), has shown promise [1]. However, MCMC is a relatively slow algorithm which wastes samples and is also limited in that it needs to conform to detailed balance. Fig. 1 (c) replicates the results of [1] and (d) shows my current work.

I plan to incorporate mixture proposals into the more flexible Sequential Monte Carlo (SMC) samplers and re-formulate the mode jumping proposals to be constructed using HMC rather than optimisation methods.

Bayesian Coresets

In scenarios with streaming data, it is infeasible to retain all the incoming data due to hardware limitations but it is necessary to preserve some past information. For example, if a model were to require re-training.

Bayesian coresets are weighted subsets of the data which are smaller than the original dataset. Currently many coreset formation algorithms are reliant on a time-consuming, clustering process or require carefully tailored, user input and are specific to certain inference algorithms [2,3]. I will look at building faster context-independent streaming coreset algorithms which build on existing work which uses variational inference or clustering methods.

Concept Drift

Over time the statistical environment a model is applied to can change from the data it was trained on which is known as concept drift. Drift can occur over varying timescales so it can be difficult to distinguish between true drift and mere outlying data resulting from random noise.

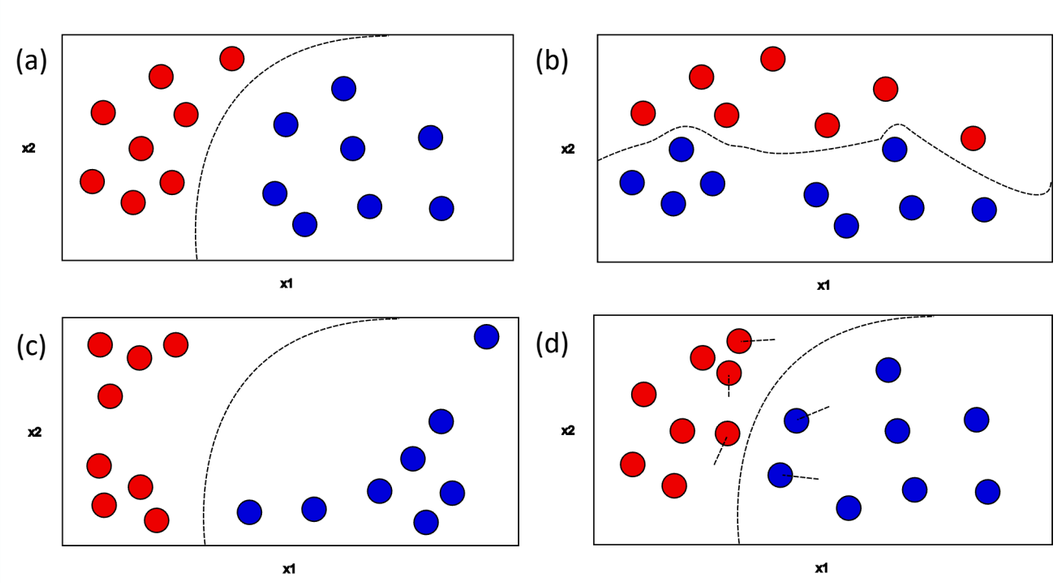

Fig. 2 (a) Original data. (b) Real concept drift. (c) Virtual concept drift. (d) Population drift.

Real concept drift is a change in the relationship between the target variable and input data (decision boundary altered). Virtual concept drift is a shift in the distribution of incoming data. Population drift refers to differences in the population from which future samples are drawn [4].

Most existing ways of tackling concept drift need the true result for comparison with the prediction but this is often only known much later. I plan to explore semi-supervised methods to ensure sufficiently rapidly labelled data to facilitate these methods.

References

[1] H. Tjelmeland and B. K. Hegstad, 'Mode Jumping Proposals in MCMC', *Scandinavian Journal of Statistics*, vol. 28, no. 1, pp. 205–223, 2001, doi: 10.1111/1467-9469.00232.

[2] T. Campbell and B. Beronov, 'Sparse Variational Inference: Bayesian Coresets from Scratch', *arXiv:1906.03329 [cs, stat]*, Oct. 2019, Accessed: Apr. 04, 2022. [Online]. Available: http://arxiv.org/abs/1906.03329.

[3] J. H. Huggins, T. Campbell, and T. Broderick, 'Coresets for Scalable Bayesian Logistic Regression', *arXiv:1605.06423 [cs, stat]*, Feb. 2017, Accessed: Mar. 29, 2022. [Online]. Available: http://arxiv.org/abs/1605.06423.

[4] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, 'A survey on concept drift adaptation', *ACM Comput. Surv.*, vol. 46, no. 4, p. 44:1-44:37, Mar. 2014, doi: 10.1145/2523813.

Engineering and Physical Sciences Research Council

UNIVERSITY OF LIVERPOOL

GCHQ