October 12, 2015

Genomic Time-Series Modeling Using Recurrent Neural Networks

Arya Iranmehr AIRANMEHR@UCSD.EDU
Vineet Bafna VBAFNA@CS.UCSD.EDU
Ali Akbari ALAKBARI@ENG.UCSD.EDU

Abstract

The advent of Next Generation Sequencing (NGS) has made it possible to study genomic data throughout time. This modern paradigm, Evolve-and-Resequence (E&R), enables us to make more accurate and robust inferences, i.e. estimate model parameters, using multiple observations along generations. In this paper, we consider the recently repopularized Recurrent Neural Networks (RNN) to model the genomic time series of E&R. In fact, RNN is used as a generative model which for a initial estate and a choice of model parameter generates a sequence. Parameter estimation procedure involves a (non-convex) optimization of least square loss between observed sequence data and RNN-generated sequence with respect to model parameter. Backpropagation-in-time, is effectively used to compute gradients of objective function, and stochastic gradient descent with momentum algorithm is used for optimization. Experimental study on simulated data shows RNN provides significantly more accurate and robust estimates in shorter times.

1. Introduction

Until very recently, biological data analysis has been considered processing a snapshot of data. However, the emergence of NGS and related technologies has made it possible to not only create larger datasets but also to measure multiple observations of the same quantity in the course of time. In many cases, such as population genetics, it is of the great interest to model the evolutionary process and make inferences, predictions and retrospective studies. Indeed, a random process is better explained by time series data than a single observation.

In addition to inexpensive data availability, over last two decades, a large amount of efforts is dedicated to High-Performance Computing (HPC), which re-popularized and rebranded computationally intensive algorithms such as Neural Networks. The first properly proposed neural network model to exploit full potential of multi layer neural networks published by Hinton and Salakhutdinov (2006); Hinton et al. (2006) and its spectacular performance on image processing problems immediately spawned the field of Deep Neural Networks (DNN), aka Deep Learning. Shortly, DNNs has made breakthroughs in generative models Sutskever et al. (2011), speech processing Hinton et al. (2012), natural language processing Collobert and Weston (2008), etc the tasks that they are originally indented to accomplish Krizhevsky et al. (2012).

Terhorst et al. (2015) Stephan et al. (2006) Hudson (2002) Peng and Kimmel (2005) Sutskever (2013) Bergstra et al. (2010) (Rumelhart et al., 1986)

2. Methods

3. Experiments

4. Discussions

References

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU Math Expression Compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Others. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Richard R Hudson. Generating samples under a WrightFisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, February 2002.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Bo Peng and Marek Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, September 2005. doi: 10.1093/bioinformatics/bti584.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

Wolfgang Stephan, Yun S Song, and Charles H Langley. The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. *Genetics*, 172(4):2647–2663, April 2006.

Ilya Sutskever. *Training recurrent neural networks*. PhD thesis, University of Toronto, 2013.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

Jonathan Terhorst, Christian Schlötterer, and Yun S Song. Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLoS Genet*, 11(4):e1005069, 2015.