A Comparison Between Offline and Online Multivariate Change-Point Detection Algorithms

Cody Buntain
cbuntain@cs.umd.edu

Christopher Natoli
natoli@uchicago.edu

Miroslav Zivkovic
M.Zivkovic@uva.nl

July 4, 2014

1 Research Abstract

Change point detection has been an active area of research for several decades, has spanned numerous disciplines, and has seen use in financial analysis, cyber security, climatology, and many other areas. With the advent of the Internet, virtual astronomical observatories, the Large Hadron Collider, genome sequencing, and other "big data" sources, however, we have already past the point where we can afford to store and analyze entire data sets in a reasonable amount of time. Furthermore, existing univariate approaches can no longer adequately capture the complex interactions present. Instead, online and streaming algorithms that process multivariate data are increasingly popular and are becoming more powerful. This departure represents a significant shift in change-point detection techniques, but relatively little research seems to exist regarding comparisons between older offline and newer online algorithms. This project seeks to bridge this gap between the old and new by comparing a set of change-point detection methods, containing both retrospective (i.e., offline) and streaming (online) approaches, across several data sets of increasing dimension and exploring the strengths and weaknesses among them.

More specifically, we plan to compare four existing change-point detection algorithms that focus on changes in covariance: two offline approaches based on the likelihood ratio test (LRT) and cumulative sums of squares (CUSUM), and two online methods using support vector machines (SVMs) and Gaussian processes [tentatively]. The expectation here is that offline methods will prove more powerful and accurate but will take much longer to process the data than online methods as the tradeoff. To obtain a more thorough picture of these techniques, we will further compare them across simulated data and three real data sets of increasing dimensionality: a bivariate sensor set from structural stress testing, a historical set of cryptocurrency prices across a small set of currency markets, and a high-dimensional social media sentiment data set. From there, we will then explore methods one might use to circumvent these algorithms to compare their weaknesses.

1.1 Algorithm List

Likelihood Ratio Test The LRT statistic as outline by Galeano and Peño [2].

Cumulative Sum of Squares The CUSUM statistic as outline by Galeano and Peño [2].

Kernel Change Detection (KD) An SVM-based, non-parametric algorithm for change detection proposed by Desobry et al. [1].

Gaussian Processes [tentative] [3]

1.2 Data Sets

Bridge Data Supplied bivariate sensor data for bridge structural stress analysis.

Historical Bitcoin Data Bitcoin price data from 2010 to present averaged across exchanges and for several different currencies (e.g., US Dollar and the Euro).

Twitter Sentiment Data Timed frequency information on a collection of sentiment-specific keywords within Twitter.

References

[1] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *Signal Processing, IEEE Transactions on*, 53(8):2961–2974, Aug. 2005.

[2] P. Galeano and D. Peña. Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*, 137(1):194–211, Jan. 2007.

[3] Y. Saatçi, R. D. Turner, C. E. Rasmussen, Y. Saat, R. D. Turner, and C. E. Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.