THESIS TITLE - TO BE DETERMINED

Tyler Manning-Dahan

A thesis
in
The Department
of
Engineering and Computer Science

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science
Concordia University
Montréal, Québec, Canada

April 2019

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Tyler Manning-Dahan**

Entitled: **Thesis Title - To be determined**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

____________________________________ Chair

____________________________________ Examiner

____________________________________ Examiner

____________________________________ Examiner

____________________________________ Supervisor

Approved ____________________________________

Chair of Department or Graduate Program Director

____________ 20 ______ ____________________________

Abstract

Thesis Title - To be determined

Tyler Manning-Dahan

Text of abstract.

Acknowledgments

I would like to thank my supervisor Dr. Jia Yuan Yu for accepting me into his lab and pushing the limit of my intellectual understanding to places I have never thought possible.

I would also like to thank all my colleagues at DRW who have helped me learn about the finance industry over the past two years. I am especially grateful to Yves, Neil and Laura, who have been very patient with me and have me taught me a lot about investigating financial market microstrucutre, writing clean code and building robust statistical models.

Lastly, I would like to thank my fiancée, Tanya, who initially pushed me to go back to school while I was still young and had the opportunity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Recent advances in neural network architecture and utilization of large amounts of data have led to astonishing results in fields such as image recognition, natural language processing, and speech recognition [add citations]. Many modern techniques such as long short term memory neural networks and convolution neural networks have been applied to time series forecasting and, while the results are promising, it remains to be seen whether such techniques can significantly outperform classical statistical methods [4]. This opens the door for novel ideas to be tried and benchmarked against existing, classical methods.

In this thesis, we concern ourselves with abrupt regime changes in time series that indicate a change in the underlying distribution of a time series. There are several constraints when tackling this problem that need to taken into consideration when choosing a model.

The first is the nature of the data for the particular use case that the model will be used in. If the data is being generated by processes that have no fundamental or well-understood process then modelling particular probability distributions becomes intractable and does not apply. Therefore, models that do not model particular probability distributions or even assume that there is one to model must be used. Methods that fall into this category are known as *non-parametric* models and are more flexible for data that do not easily fall into a well known structure.

The second consideration is to determine what in what scenario the data is analysed. Some algorithms are off-line or batch algorithms in that they are applied in an ex post fashion after the dataset has been completely acquired. The first attempts are change point detection using control charts by Shewart were done in an on-line fashion as the data is streamed in a near, real-time fashion. In the literature, on-line methods of change point detection are also sometimes called sequential change point detection. For this thesis, we will use the terms interchangeably.

The third consideration that is a consequence of detecting change points in an on-line setting is the presence of outliers. This a particularly important concern for on-line methods that are looking for changes in the underlying distribution of points, not a single anomalous data point

The fourth consideration is determining if there are multiple change points or to assume there is only a single change point to detect. This is a more relevant consideration for off-line change point detection, but could also be relevant for the on-line case if a situation arises where the window of time series under consideration may contain more than one change point.

1.2 Previous Work

In 2016, James et al. [3] proposed using energy statistics to test the significance of a change point that is robust to anomalies. They point out that their work is the first to accurately detect change points with fast time to detection while not being affected by extreme outliers that are not change points. They compare their E-divisive with medians algorithm to the parametric PELT technique.

1.3 Problem Formulation

In the 2017 survey done by Aminikhanghahi and Cook [1], the authors give the following definition of a *change point*: A change point represents a transition between different states in a process that generates the time series data.

Generally speaking, the fundamental change point problem is the following:

- $H_0 : P_{X_m} = P_{X_{m+k}}$
- $H_A : P_{X_m} \neq P_{X_{m+k}}$

1.3.1 Performance Measures

Add information about what performance measures will be used to determine the utility of algorithm. E.g. time to detection, F1 measure, etc.

1.4 Data Sources

Inspired by twitter trend detection research, the ultimate goal is to test our contribution on labelled twitter count data that have some mix of trending and non-trending sequences. Since this data would be costly to acquire and label. This will not be used to test the performance of this work. Instead, the following sources of data will be used:

1.4.1 Synthetic datasets

1.4.2 Private dataset

1.5 Review of 2017 Garreau PhD

This PhD is an extension on the 2012 paper by Arlot et al. [2] that proposes a type of kernel change point detection. Their non-parametric method is concerned with the off-line setting of detecting change points. They tackle the possibility of multiple change points.

Fundamentally, their method is the kernel version of the following least squares optimization problem:

The benefits of this kernel change point detection is that it operates on any kind of data for which a kernel that properly reproduces a Hilbert space can be applied.

His method is theoretically sound and tested but has not been implemented on very many real datasets nor compared with other benchmark methods. In an application setting, the user would use some training set to calibrate the kernel and the penalty parameters, to then be tested with appropriate accuracy measures on some out of sample data set.

Possible extensions include testing his method on real world data sets and providing a comparison with other methods. Another extension is making an on-line version of his method. A harder extension would be automatic selection of a kernel based on the particular statistical moment that a user is interested in.

1.6 Our Contributions

Lay out the contributions of this work.

Chapter 2

Chapter 2

2.1 Chapter 2 Section

2.1.1 Chapter 2 Subsection

Subsection Test

Bibliography

[1] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[2] Sylvain Arlot, Alain Celisse, and Zaıd Harchaoui. Kernel change-point detection. *arXiv preprint arXiv:1202.3878*, 6, 2012.

[3] Nicholas A James, Arun Kejariwal, and David S Matteson. Leveraging cloud data to mitigate user experience from 'breaking bad'. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3499–3508. IEEE, 2016.

[4] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889, 2018.