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 the beginning or end of significant trends of a time series. There are two main tasks in tackling this problem. The first is identifying when a time series has a collective outlier that is accurately discernible from noise and not simply a contextual outlier. The second is we would like regime changes to be detected as soon as possible.

Over the past decade, time series classification has seen extremely competitive results using simple nearest neighbour techniques [2]. Fundamentally it involves a distance calculation between two sequences, followed by a one nearest neighbour classifier. The most common distance functions used for this is the Euclidean distance or some variant thereof.

1.2 Previous Work

[make this more related to the exact problem you are solving not a literature review]

Detecting abrupt changes in a time series is a common task in time series analysis and signal processing. Many outlier detection methods exist such as . A closely related topic is change point detection which focuses on discerning when possible regime shifts occur in a time series. Here the regime shifts indicate collective outliers and are considered out of place when compared to another regime in the time series.

Broadly speaking there are two main categories that methods fall into, parametric and non-parametric modelling. Parametric models attempt to model the data directly through parametrization of the phenomenon being measured. This makes sense for data that are based on well-understood theoretical models or natural phenomenon. When no theoretical models exist and data is generated in then a non-parametric model is suitable.

Another point brought up by dog [1] is the fact that

1.3 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:

1.3.1 Synthetic datasets

1.3.2 Public datasets

1.3.3 Private dataset

1.4 Latent Source Model

The set of examples $\mathcal{Y}$ is the set of possible labels $0, 1$ to denote trend or no trend. Therefore, the learner will attempt to learn a model using the training set that is

defined as $\mathcal{X} \times \mathcal{Y} : \mathcal{S} = ((x_1, y_1), (x_2, y_2)...(x_m, y_m))$. The term training set is used loosely here as no model parameters are actually learned.

Chen et al. [3] describe a *latent source* model for detecting sudden trends on twitter. The model is a data-driven, non-para metric classifier that is capable of classifying an observed, streaming time series as either a trending time series or not a trending time series. They construct the time series using the counts of a twitter string (or hash-tag) over time, binned at specific intervals. The approach taken is based on the assumption that there are fundamental behaviours that generate tweets when they trend compared to when they do not trend. These unknown fundamental behaviours are called *latent sources*.

In this supervised learning setting, data is labelled with either “trend" or “no trend". Each labelled data point is referred to as a reference signal or latent source and is denoted as r. The entire a set of reference signals is denoted as R. Therefore, we have a binary classifier where a new data point is classified into sets $R+$ or $R-$, reflecting the set of trending signals and non-trending signals respectively.

Nikolov [5] specifies an observation, s, is generated by a latent source r if s is a noisy version of r and proposes the following stochastic model:

$$\mathbb{P}(s \text{ generated by } r) \propto e^{-\lambda d(r,s)} \tag{1}$$

The distance function used is the sum of squares Euclidean distance:

$$d(r, s) = \sum_{i=1}^{N} (r_i - s_i)^2 \tag{2}$$

Here the observed signal s and the reference signal r are of length N. Note that the squared Euclidean distance is not a metric as it does satisfy the triangle inequality. However, as the authors point out, the function d may be replaced by any function that is symmetric, positive definite, and convex. Once the distance is computed, the weight of its vote is determined by:

$$W(r, s) = e^{-\lambda d(r,s)} \tag{3}$$

Where the scaling parameter λ can be thought of as a “sphere of influence" that allows the user to tune the relative importance of a similar or dissimilar time series. For example, a large value of λ generates very small weights for elements of R that are very different from s.

Finally, the weights are summed across all the items in the trending set, $R+$, and divided by all time series in the non-trending set, $R-$, to create a final metric η:

$$\eta(s) = \frac{\mathbb{P}(+|s)}{\mathbb{P}(-|s)} = \frac{\sum_{r \in R+} W(r,s)}{\sum_{r \in R-} W(r,s)} \tag{4}$$

The estimated classification of an observed time series, s, can then be defined as:

$$\hat{L}(s) = \begin{cases} +, & \text{if } \eta(s) > \theta. \\ -, & \text{if } \eta(s) \leq \theta. \end{cases} \tag{5}$$

Where θ defines the threshold for classification and is typically is set to 1. However, it can be tuned depending on the use case. For example, in cases where false positives are costly, the value of θ can be set to greater than 1.

1.5 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] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.

[3] George H Chen, Stanislav Nikolov, and Devavrat Shah. A latent source model for nonparametric time series classification. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1088–1096. Curran Associates, Inc., 2013.

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

[5] Stanislav Nikolov. Trend or no trend: a novel nonparametric method for classifying time series, 2012.