Detecting Online Conversations Going Viral

Time-series aware evaluation of change detection algorithms

Matt Chapman
matthew.chapman@student.uva.nl

Spring 2017, 15 pages

Supervisor: Evangelos Kanoulas, Universiteit van Amsterdam
Host organisation: Buzzcapture International, http://www.buzzcapture.com

Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Master Software Engineering
http://www.software-engineering-amsterdam.nl

Contents

Abstract

Write abstract

Chapter 1

Problem Statement & Motivation

1.1 Problem Statement

Within the domain of change detection algorithms there are a number of available methods for evaluating the efficiency and accuracy of a given algorithm, depending on how the problem is framed. The different problem framings available are, for example:

are these definitions correct?

Classification Wherein the algorithm result fits into one of two (or more) classes, for example correct or incorrect.

Clustering Wherein the algorithm results are formed into clusters and evaluated using clustering metrics.

Partitioning Like clustering, but?

How is this different from clustering?

Retrieval Wherein the results of the algorithm are scored as a *retrieval problem*, where detection relevance is taken into account, perhaps along with some form of temporal or redundancy penalty.

With this research, it is intended to address this dichotomy between approaches, and further suggest an evaluation approach that is suitable for use when attempting to evaluate change detection algorithms applied to time-series data specifically.

While change detection itself has been around since the 1930's (and *online* change detection since the 1950's) it is still a field that attracts new thoughts and approaches - one example of which being Ginsberg et al.'s "Detecting influenza epidemics using search engine query data." [Gin+09].

Various attempts have been made to evaluate the myriad approaches to change detection ([BNZ14] for example) but each of these attempts tend to frame the problem somewhat differently, or do not take into account change detection in time-series data.

This research intends to address the following research questions:

RQ1 Are there deficiencies in existing methods for evaluating change detection algorithms?

RQ2 Do certain measures perform better as an evaluation method when applied against changes detected in a data stream with certain properties?

RQ3 Is there room for adjustment in existing measures, such that they can be made more effective for the evaluation of change detection algorithms?

this needs expansion, I think

1.2 Motivation

Change detection first came about as a quality control measure in manufacturing, and methods within this domain are generally referred to as *control charts*. Since the inception of approaches such as CUSUM that provide the possibility for on-line evaluation of continuous data streams, change detection has grown as a field. With applications such as epidemic detection, online reputation management and infrastructure error detection, change detection is hugely useful.

This particular research is motivated specifically by the online reputation management sector. The business hosting this research project (Buzzcapture International [http://www.buzzcapture.com]) is a Dutch online reputation management company that provides services to other businesses throughout europe. Chief among these is the BrandMonitor application, which, among other features, provides a rudimentary notification system for clients that is triggered once there is an increase in conversation volume of $\%n$. It is the intention of this research to provide a robust evaluation method for change detection algorithms such that an approach that is most effective for this particular use case can be selected and implemented.

Chapter 2

Research Method

2.1 Existing Approaches

As briefly discussed in the introduction to this thesis, there are a number of pre-existing approaches for the evaluation of change detection methods. Here some examples are detailed:

F1 Score This measure is utilised for testing accuracy in problems of binary classification. It considers two different measures, *presision* and *recall*. The F1 score can be described in general terms as follows:

cite this?

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.1}$$

recall is computed as the number of correct positive results, divided by the number of positive results that should have been detected. *precision* is computed as the number of correct positive results divided by the number of all possible positive results.

As the F1 score is a binary classification measure, it can only be used to test the precision of an algorithm in a single domain, that is, was the change detected or not.

Rand Index This measure is for computing the similarity between two clusters of data points. It is generally considered an accuracy score for clustering mechanisms. The Rand Index is defined as:

$$R = \frac{a + b}{a + b + c + d} \tag{2.2}$$

Where: define

2.2 Evaluation Pipeline

2.2.1 Introduction

The method of evaluating the approaches will be developed using R. R is a combined language and environment created for the purpose of statistical computing [R C15]. Thanks to the availability of the `changepoint` package in R[KE14], it is possible for me to experiment with different change detection approaches without needing to implement them from scratch, myself.

R also allows for the generation of test data using method calls such as `rnorm()`, which has considerably sped up my workflow compared to carrying out the same operations in Python. The R package `ggplot2` [Wic09] also provides a number of powerful tools for data visualisation directly in R.

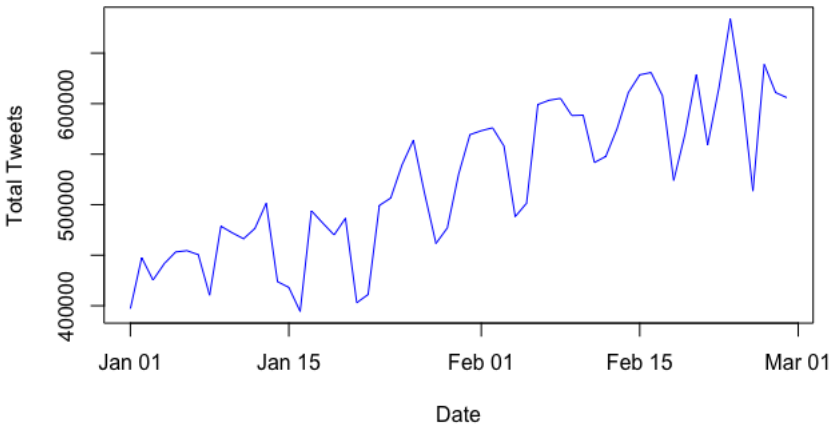

Figure 2.1: Twitter Data Set

2.2.2 Changepoint

`changepoint` is a powerful R package that provides a number of different change detection algorithms, along with various approaches to penalty values. `changepoint` offers change detection in mean, variance and combinations of the two, using

2.3 Data Preparation

There are two data sets that will be used in this research. The first is a collection of 30.9M Dutch language tweets collected between 01/01/2017 and 28/02/2017. The data will be filtered for certain terms in the tweet 'body' to generate collections of tweets that would be considered 'relevant' to a particular business or individual. In this way, I will be able to simulate BrandMonitor queries myself, without the need of the software itself.

Make graph prettier

The final result of the prepared datashall be 5 different subsets of varying size and nature. For example, data sets with large changes in variance (noise), and extremely large/small changes in mean.

finish getting data together

2.3.1 The Nature of Twitter Data

It is important, as part of this research, to understand the nature of the data being studied.

Figure 2.2 shows the distribution of twitter postings mentioning "ING" between 01/01/2017 and 28/02/2017. At first glance, this data has a number of interesting change points that we may wish to detect. Firstly, we shall examine what kind of distribution this data falls under. Carrying out the Shapiro-Wilk normality testgives us a *pvalue* of 0.0001037, well below the normally accepted 0.05. We can further see that this data set does not fit a normal distribution by generating a Q-Q plot of the data using R. 2.3 shows the completed QQ plot, which demonstrates the lack of normal distribution.

cite

The second type of data that will be used is entirely simulated. This will be achieved using R's facilities for generating sets of random data that fits a normal distribution. In this way it will be possible to demonstrate how (if at all) change detection algorithms handle data with different distribution models differently.

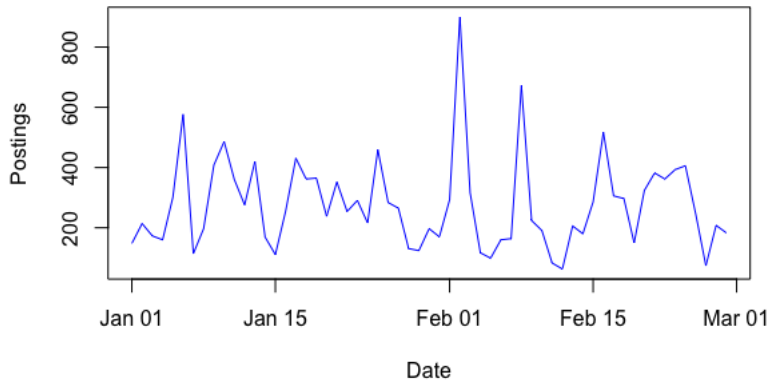

Figure 2.2: "ING" postings by date

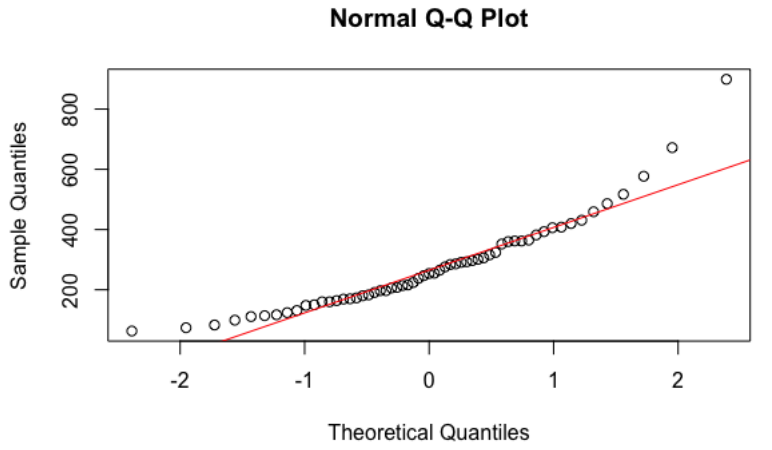

Figure 2.3: Q-Q plot of ING postings

2.4 Scoring Metrics

Example function:

$$f(relevance, temporal\_penalty, redundancy\_penalty) \tag{2.3}$$

start designing scoring methods

Possible relevance measure, where t_0 is the earliest a spike can be detected and t_n is the time that the signal returns to normal. $f(x)$ describes the function of the curve:

$$\int_{t_n}^{t_0} f(x)dx \tag{2.4}$$

start desigining relevance measure

planning to do something with the area under the curve for relevance

Chapter 3

Background & Context

This research was primarily borne out of reading "A Brief Comparison of Algorithms for Detecting Change Points in Data" by Buntain, Natoli, and Zivkovic[BNZ14]

Chapter 4

Research

Chapter 5

Results

Chapter 6

Analysis & Conclusions

Bibliography

[AF13] Leman Akoglu and Christos Faloutsos. "Anomaly, event, and fraud detection in large network datasets". In: *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13* (2013), p. 773. DOI: 10.1145/2433396.2433496. URL: http://dl.acm.org/citation.cfm?id=2433396.2433496.

[Alv+11] Foteini Alvanaki et al. "EnBlogue: emergent topic detection in Web 2.0 streams". In: *Proc. ACM SIGMOD International Conference on Management of Data* (2011), pp. 1271–1274. ISSN: 07308078. DOI: 10.1145/1989323.1989473. URL: http://doi.acm.org/10.1145/1989323.1989473%7B%5C%%7D5Cnpapers2://publication/uuid/44E009D1-8574-49C1-A0AC-C2B247FE9043.

[Ami+09] Enrique Amigó et al. "A comparison of extrinsic clustering evaluation metrics based on formal constraints". In: *Information Retrieval* 12.4 (2009), pp. 461–486. ISSN: 13864564. DOI: 10.1007/s10791-008-9066-8.

[BN93] M Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. 1993. ISBN: 0-13-126780-9. DOI: 10.1016/0967-0661(94)90196-1. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.6896%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

[BPP07] S. Bersimis, S. Psarakis, and J. Panaretos. "Multivariate statistical process control charts: An overview". In: *Quality and Reliability Engineering International* 23.5 (2007), pp. 517–543. ISSN: 07488017. DOI: 10.1002/qre.829.

[BNZ14] Cody Buntain, Christopher Natoli, and Miroslav Zivkovic. "A Brief Comparison of Algorithms for Detecting Change Points in Data". In: *Supercomputing*. 2014. URL: https://github.com/cbuntain/ChangePointDetection.

[Das+09] Tamraparni Dasu et al. "Change (detection) you can believe in: Finding distributional shifts in data streams". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5772 LCNS (2009), pp. 21–34. ISSN: 03029743. DOI: 10.1007/978-3-642-03915-7_3.

[DDD05] Frédéric Desobry, Manuel Davy, and Christian Doncarli. "An online Kernel change detection algorithm". In: *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 2961–2974. ISSN: 1053587X. DOI: 10.1109/TSP.2005.851098.

[Dow08] Allen B. Downey. "A novel changepoint detection algorithm". In: *Applied Microbiology and Biotechnology* (2008), pp. 1–11. arXiv: 0812.1237. URL: http://arxiv.org/abs/0812.1237.

[FP99] Tom Fawcett and Foster Provost. "Activity monitoring: Noticing interesting changes in behavior". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* 1.212 (1999), pp. 53–62. ISSN: 09258574. DOI: 10.1016/j.ecoleng.2010.11.031. URL: http://portal.acm.org/citation.cfm?id=312195.

[GP04] Pedro Galeano and Daniel Peña. "Variance Changes Detection in Multivariate Time Series". 2004.

[Gin+09] Jeremy Ginsberg et al. "Detecting influenza epidemics using search engine query data." In: *Nature* 457.7232 (2009), pp. 1012–4. ISSN: 1476-4687. DOI: 10.1038/nature07634. URL: http://www.ncbi.nlm.nih.gov/pubmed/19020500.

[KS09] Yoshinobu Kawahara and Masashi Sugiyama. "Change-point detection in time-series data by direct density-ratio estimation". In: *Proceedings of the 2009 SIAM International Conference on Data Mining* (2009), pp. 389–400.

[KBG04] Daniel Kifer, Shai Ben-david, and Johannes Gehrke. "Detecting Change in Data Streams". In: *Proceedings of the 30th VLDB Conference* (2004), pp. 180–191. ISSN: 00394564. DOI: 10.1016/0378-4371(94)90421-9. arXiv: 9310008 [cond-mat].

[KE14] Rebecca Killick and Idris A. Eckley. "changepoint: An R Package for Changepoint Analysis". In: *Journal of Statistical Software* 58.3 (2014), pp. 1–19. DOI: 10.18637/jss.v058.i03. URL: http://www.jstatsoft.org/v58/i03/.

[Kul+05] Martin Kulldorff et al. "A space-time permutation scan statistic for disease outbreak detection". In: *PLoS Medicine* 2.3 (2005), pp. 0216–0224. ISSN: 15491277. DOI: 10.1371/journal.pmed.0020059.

[LS99] Tze Leung Lai and Jerry Zhaolin Shan. "Efficient recursive algorithms for detection of abrupt changes in signals and control systems". In: *IEEE Transactions on Automatic Control* 44.5 (1999), pp. 952–966. ISSN: 00189286. DOI: 10.1109/9.763211.

[MJ12] David S. Matteson and Nicholas A. James. "A nonparametric approach for multiple change point analysis of multivariate data". In: *Submitted* 14853 (2012), pp. 1–29. ISSN: 0162-1459. DOI: 10.1080/01621459.2013.849605. arXiv: 1306.4933.

[PRG10] Anita M Pelecanos, Peter a Ryan, and Michelle L Gatton. "Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease." In: *BMC medical informatics and decision making* 10.1 (2010), p. 74. ISSN: 1472-6947. DOI: 10.1186/1472-6947-10-74. URL: http://www.biomedcentral.com/1472-6947/10/74.

[Qah+15] Abdulhakim A. Qahtan et al. "A PCA-Based Change Detection Framework for Multidimensional Data Streams". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* (2015), pp. 935–944. DOI: 10.1145/2783258.2783359. URL: http://dl.acm.org/citation.cfm?id=2783258.2783359.

[R C15] R Core Development Team. *R: a language and environment for statistical computing, 3.2.1*. R Foundation for Statistical Computing. Vienna, Austria, 2015. ISBN: 3-900051-07-0. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3. URL: https://www.r-project.org/.

[SV95] D. Siegmund and E. S. Venkatraman. "Using the generalized likelihood ratio statistic for sequential detection of a change-point". In: *The Annals of Statistics* 23.1 (1995), pp. 255–271. ISSN: 0090-5364. DOI: 10.1214/aos/1176324466.

[Sin+05] T. Sing et al. "ROCR: visualizing classifier performance in R". In: *Bioinformatics* 21.20 (Oct. 2005), pp. 3940–3941. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti623. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti623.

[TR05] Alexander G Tartakovsky and Boris L Rozovskii. "A Nonparametric Multichart CUSUM Test for Rapid Intrusion Detection". In: *Proceedings of Joint Statistical Meetings* (2005), pp. 7–11.

[Tar+06] Alexander G. Tartakovsky et al. "Detection of intrusions in information systems by sequential change-point methods". In: *Statistical Methodology* 3.3 (2006), pp. 252–293. ISSN: 15723127. DOI: 10.1016/j.stamet.2005.05.003.

[TGS14] Dang-Hoan Tran, Mohamed Medhat Gaber, and Kai-Uwe Sattler. "Change Detection in Streaming Data in the Era of Big Data: Models and Issues". In: *ACM SIGKDD Explorations Newsletter - Special issue on big data* 1 (2014), pp. 30–38. ISSN: 1931-0145. DOI: 10.1145/2674026.2674031.

[Wic09] Hadley Wickham. *Elegant Graphics for Data Analysis*. Vol. 35. July. Springer-Verlag New York, 2009, p. 211. ISBN: 9780387981406. DOI: 10.1007/978-0-387-98141-3. arXiv: arXiv:1011.1669v3. URL: http://had.co.nz/ggplot2/book.

[WJ76] Alan S. Willsky and Harold L. Jones. "A Generalized Likelihood Ratio Approach to the Detection and Estimation of Jumps in Linear Systems". In: *IEEE Transactions on Automatic Control* 21.1 (1976), pp. 108–112. ISSN: 15582523. DOI: 10.1109/TAC.1976.1101146.

[Xu+11] Yi Xu et al. "Pattern change discovery between high dimensional data sets". In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* (2011), p. 1097. DOI: 10.1145/2063576.2063735. URL: http://dl.acm.org/citation.cfm?doid=2063576.2063735.

ToDo Notes