Operations Research Letters 41 (2013) 581–585

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Analyzing social media data having discontinuous underlying dynamics[☆]

L.D. Servi

The MITRE Corporation, 202 Burlington Road, Bedford, MA 01773, USA

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form
24 June 2013
Accepted 27 July 2013
Available online 11 August 2013

Keywords:
Breakpoints
Dynamic programming
Social media
LIWC
Dynamics of emotions
Twitter

ABSTRACT

This paper develops a real time algorithm which identifies times of emotional discontinuity as reflected in social media. The paper formulates the optimization problem to solve, develops an algorithm to solve it using dynamic programming, and illustrates the new method by analyzing mood shifts reflected in 380,000 Twitter messages related to one of the world's most popular soccer teams, *Manchester United*, during their 2011–12 season.

1. Introduction

Although physical objects follow well understood Newtonian physics and are free of discontinuous dynamics, emotional intensity dynamics evolve continuously with unknown dynamics until a tipping point or external event triggers a discontinuity. Classical tracking [6] assumes underlying continuous dynamics, an assumption that is structurally inappropriate when tracking emotions expressed in social media. Classical statistical control process (SPC) [7] methods, on the other hand, assume a known underlying continuous dynamics with known parameters and are designed to statistically detect discontinuities. However, SPC does not simultaneously estimate the dynamics while seeking to detect the discontinuities.

Motivated by the deficiency of applicable models, this paper proposes a more nuanced dynamic. Specifically, it hypothesizes a continuous dynamic with unknown parameters interrupted by jumps occurring at unexpected times followed by resumed continuity independent of the dynamics before the jump. The discontinuous jumps might be viewed as emotional resets caused by significant external events. Analysis seeking to track such systems amid noisy measurements must simultaneously smooth the noisy measurements while identifying the times of the discontinuity.

This section first formally presents a dynamical model along with a motivating example which will be explored in more detail in Section 4. This section also presents an exhaustive search algorithm to analyze such data. Section 2 describes a mathematically equivalent algorithm which is recursive and therefore lends itself to processing data in real time. Section 3 presents a number of extensions to the algorithm. Section 4 applies the algorithm to tracking emotions inferred from twitter messages related to the soccer team, *Manchester United*.

More specifically, this paper partitions L time series x_t^ℓ for $t = 1, 2, \dots, T$ and $\ell = 1, 2, \dots, L$ into M consecutive regions with breakpoints at times $\mathcal{M} = \{m_0, m_1, \dots, m_M\}$, where $m_0 = 0$ and $m_M = T$, which is most consistent with a specified model of the dynamics. If, for example, the time series arise from piecewise linear dynamical functions, then, for $j = 1, 2, \dots, M$, in the jth partition, the ℓth time series $x_t^\ell \approx \alpha_j^\ell + t\beta_j^\ell$ for $t = m_{j-1} + 1, \dots, m_j$. For this example, this paper identifies the most consistent value of m_j, α_j^ℓ, and β_j^ℓ for $j = 1, 2, \dots, M - 1$ and $\ell = 1, \dots, L$.

Before formulating and solving this problem, consider the following example. Fig. 1 illustrates data from 380,000 twitter messages corresponding to noisy measurements of the intensity of expressed anger, sadness, and 'positive emotions' (as defined by the LIWC dictionary [8,9,11]) for each day between September 1, 2011 and May 15, 2012. Here, there are 3 time series corresponding to the 3 different emotions, i.e., $L = 3$. Fig. 1, while in principle

[☆] This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

E-mail address: lservi@mitre.org.

0167-6377/$ – see front matter © 2013 The Author. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.07.008

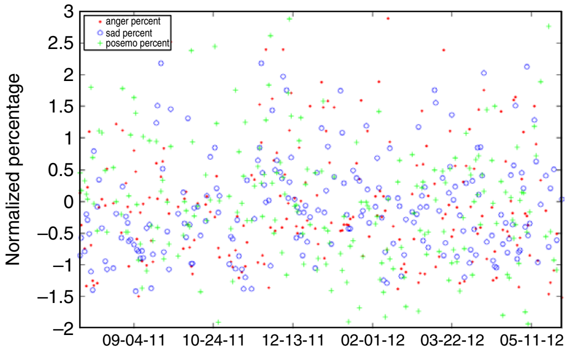

Fig. 1. Raw mood data from 380,000 Twitter messages having the hashtag #mufc.

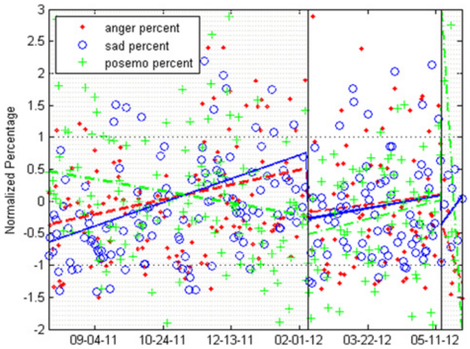

Fig. 2. Mood shift analysis of the Twitter messages showing a breakpoint on 2/6/2012 and 5/14/2012.

informative, is hard to interpret or glean knowledge from. Suppose each emotion follows a piecewise linear dynamical function between breakpoint times and discontinuously changes at the breakpoint times. Then, using the analysis and modeling assumptions of this paper, as discussed in more detail in Section 4, the data of Fig. 1 is most consistent with (i) there being $M = 3$ breakpoints on the dates illustrated in Fig. 2 and (ii) the smoothed linear dynamics of the three emotions between each of the three breakpoints are the equations corresponding to the straight lines illustrated in Fig. 2. Section 4 discusses why Fig. 2 might reasonably be considered consistent with real events associated with the breakpoint dates.

Proceeding more formally with more generality, suppose

$$\vec{x}_j^\ell(\vec{m}) \approx \mathbf{A}_j\vec{\theta}_j^\ell, \tag{1.1}$$

where

$$\vec{x}_j^\ell(\vec{m}) = (x^\ell_{m_{j-1}+1}, x^\ell_{m_{j-1}+2}, \ldots, x^\ell_{m_j})^T,$$

and $\mathbf{A}_j$ might represent a D degree polynomial model, e.g.,

$$\mathbf{A}_j = \begin{pmatrix} 1 & m_{j-1}+1 & \cdots & \cdots & (m_{j-1}+1)^D \\ 1 & m_{j-1}+2 & \cdots & \cdots & (m_{j-1}+2)^D \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & m_j & \cdots & \cdots & m_j^D \end{pmatrix},$$

(or perhaps instead a model with periodic functions), where $D = 0$ and $D = 1$ are the cases of most interest. Here, $\vec{\theta}_j^\ell$ is an unknown vector to be determined and the dependency of $\mathbf{A}_j$ on $\vec{m}$ and D are suppressed for notational clarity.

A measure of the consistency of the model in the jth breakpoint region is defined as

$$d(m_{j-1}, m_j, \vec{w}) = \vec{w}^T\vec{r}_j(\vec{m}), \tag{1.2}$$

where $\vec{r}_j(\vec{m})$ is an $L \times 1$ vector whose ℓth component is

$$[\vec{r}_j(\vec{m})]_\ell = \min_{\vec{\theta}_j^\ell} \left\| \vec{x}_j^\ell(\vec{m}) - \mathbf{A}_j\vec{\theta}_j^\ell \right\|_P^2, \tag{1.3}$$

$\vec{w} = (w_1, w_2, \ldots, w_L)$ represent weights reflecting the relative importance of the different time series with $\sum_\ell w_\ell = 1$, $\mathbf{P} = \text{diag}(p_1, p_2, \ldots)$ reflects the relative importance of the data points within a time series (due to the discounting of time or due to differing precision of the measurements of that data perhaps induced by a differing number of underlying raw measurements for some data points), and $\|x\|_P = \sqrt{\sum_t p_t x_t^2}$. Eq. (1.3) could be interpreted as finding the maximum likelihood estimate for the case of $x_j^\ell \sim \mathcal{N}(\mathbf{A}_j\vec{\theta}^\ell, \mathbf{P})$.

Finding each component of $\vec{r}_j(\vec{m})$ entails solving a simple linear regression: the minimum $\vec{\theta}^\ell$ in (1.3) is

$$\hat{\theta}_j^\ell = (\mathbf{A}_j^T\mathbf{P}\mathbf{A}_j)^{-1}\mathbf{A}_j^T\mathbf{P}\vec{x}_j^\ell(\vec{m}) \tag{1.4}$$

if the dimension of $\vec{x}_j^\ell(\vec{m})$, $m_j - m_{j-1}$, is more than D. If not, there are an infinite number of ways $\vec{r}_j(\vec{m})$ can achieve 0 in (1.3) and a variation of (1.4) must be used.

For the special case of finding a piecewise constant function in each breakpoint region, i.e., $D = 0$, if $\mathbf{P} = \mathbf{I}$, then (1.2) and (1.4) simplify to $\hat{\theta}_j^\ell = \sum_{t=m_{j-1}+1}^{m_j} x_t^\ell/(m_j - m_{j-1})$, i.e., the average value, and $d(m_{j-1}, m_j, \vec{w}) = \sum_{\ell=1}^L w_\ell \left(\sum_{t=m_{j-1}+1}^{m_j}(x_t^\ell - \hat{\theta}_j^\ell)^2\right)$, i.e., a weighted sample variance.

More generally, assuming $\mathbf{P}$ is diagonal, $D < M$, then from (1.2)–(1.4) computing $d(m_{j-1}, m_j, \vec{w})$ requires $O(LT^2)$ multiplications as the evaluation of norm in (1.3) requires $O(T^2)$ and must be computed for L different values of j.

A measure of the quality of the model over all of the breakpoint regions is, for $i = M$,

$$S_i(\vec{m}, \vec{w}) = \sum_{j=1}^{i} d(m_{j-1}, m_j, \vec{w}). \tag{1.5}$$

Hence, finding the best breakpoints, $\vec{m}^*(\vec{w})$, entails solving

$$\vec{m}^*(\vec{w}) = \arg\min_{\vec{m}\in\mathcal{M}} S_M(\vec{m}, \vec{w}), \tag{1.6}$$

where

$$\mathcal{M} = \{(m_0, m_1, \ldots, m_M) : 0 = m_0 \le m_1 \le m_M = T\}.$$

Eq. (1.5) requires $O(MLT^2)$ multiplications since it entails evaluating $O(M)$ values of j which each requiring $O(LT^2)$ multiplications to evaluate (1.4). Hence, (1.6) could be computed using an exhaustive search requiring $O(MLT^M)$ multiplications. Note that (1.6) is equivalent to

$$\vec{m}^*(\vec{w}) = \arg\min_{\vec{m}\in\mathcal{M}} \left[\sum_j \sum_\ell w_\ell \min_{\vec{\theta}_j^\ell} \left\| \vec{x}_j^\ell(\vec{m}) - \mathbf{A}_j\vec{\theta}_j^\ell \right\|_{\mathbf{P}}^2\right].$$

2. A fast recursive algorithm for finding the breakpoints

While the exhaustive search approach (1.6) entails computing the measure $S_i(\vec{m}, \vec{w})$ for all $O(T^M)$ possibilities and finding a faster recursive approach is possible using the following dynamic programming argument motivated by [4] which, in turn draws on [2,5].