A Data-processing Method Applied to the Detection of Emotional Breakpoints in Social Media

Runze Mao

Abstract: Studies have shown that significant changes often occur in social emotions unexpectedly, which increases the difficulty analyzing and predicting the emotions. Currently, few algorithms can address this problem. In this paper, we propose three improvements to an existing algorithm. Experiments show that these improvements help the algorithm fit better into real-world noisy data, and successfully find out the changes which match with the real events.

Keywords: Denoising, piecewise regression, sentiment analysis.

1 Introduction

Social media has become significantly helpful in fields like social psychology and communication owing to the vast information it contains. The users' comments and retweets can reveal their emotional states and opinions on a specific topic. By analyzing these data, we may extract the historical dynamics of the overall emotions expressed by the society, from which we can identify the characteristics of these emotions and even forecast future trends. Research into these aspects is part of the field called sentiment analysis.

However, when analyzing the emotional dynamics, it is usually hard to fit the data using traditional regression methods because of the fact that emotional intensity of social media users would evolve discontinuously at some tipping points. In other words, there are significant changes in emotions at some special time points, which we prefer to call **emotional breakpoints.** The emotional intensity changes remarkably near emotional breakpoints, but keeps steady between two adjacent breakpoints.

The cause of emotional breakpoints is still uncertain. According to L. Servi and S. B. Elson [1], psychologists and persuasion researchers study issues like human choices, while some other studies focus on the relations among the media users. Both ways consider the external events as the cause of emotional breakpoints. In M. Tsytsarau *et al.*'s work [2], the authors study the dynamics of news volume, which is much like emotional dynamics, and take the internal causes into consideration.

Since the causes of emotional breakpoints are unknown, it is challenging to detect the breakpoints, that is to say, to identify when they occur. Currently, applicable models are still deficient. L. D. Servi [3] proposed a piecewise-regression model, based on the assumption that emotional dynamics evolve continuously with *unknown* parameters until an interrupt occurs *unexpectedly*, and enter another continuous evolution which is independent of the former one after the interrupt. Motivated by this work, we polished it so that it can be applied to more kinds of noisy data, with a data-processing method which makes the breakpoints more detectable.

The rest of our paper will be organized in this way. In Section 2, we formalize the most important notations used in this paper for convenience. In Section 3, several previous work related to the detection of emotional breakpoints will be introduced, and their disadvantages will be illustrated. Then we show our data-processing method in Section 4. A mathematical demonstration is given in Section 5. In Section 6, we apply the method to two independent datasets, respectively, to detect the emotional breakpoints.

2 Formalism

To make our demonstration more coherent, we have to formalize the usage of some notations in our paper, which are summarized in Table 1. And a more detailed explanation is given below.

N	Length of data
Y	Time series of emotional intensity
Z	The handled data
M	The number of regions Y is partitioned into
$\vec{w}$	The emotional breakpoints vector
λ	The regularization constant
$\vec{\theta_j}$	The regression parameters of the j-th region

Table 1: Notations used in this paper

In our problem, Y is the time series of emotional intensity of length N, starting from 0 to $N-1$. The term *region* is defined as a part of Y lying between two adjacent emotional breakpoints. For example, two adjacent emotional breakpoints at t_1 and t_2, respectively, specifies the region $t_1, t_1+1, \ldots, t_2-1$, namely, the interval $[t_1, t_2)$. We are intended to partition Y into M regions using $M+1$ breakpoints, with the first at 0 and the last at N. The entire breakpoints compose the breakpoints vector $\vec{w}$, which is $(w_0, w_1, \ldots, w_M)$. For each region, we have to calculate the regression parameters $\vec{\theta}$ to fit the data within the region.

3 Related Work

In this section, two existing methods applied to the dectection of emotional breakpoints will be introduced. Their drawbacks will be discussed.

3.1 Piecewise Regression

3.1.1 Basic Idea

L. D. Servi在 [3]中提出了一种分段回归模型，它基于这样一种思路：

使用分段函数去拟合数据，每一段都是一个独立的回归。设定的断点不同，则分段方法不同，那么回归得到的误差也不同。使得各段回归的后向误差之和最小的分段方法，确定了最优解对应的断点所在。用数学的方式来表达，设问题的解空间为：

$$\mathcal{W} = \left\{(w_0, \ldots, w_M) \mid 0 = w_0 < \ldots < w_M = N\right\}$$

分段回归方法所求的解即为：

$$\vec{w}^* = \underset{\vec{w}\in\mathcal{W}}{argmin} \sum_{j=1}^{M} C(w_{j-1}, w_j)$$

其中，$C(w_{j-1}, w_j)$是模型在区间$[w_{j-1}, w_j)$上回归的后向误差，计算方法如下。令：

$$X_j = \begin{pmatrix} 1 & w_{j-1} & \ldots & w_{j-1}^d \\ 1 & w_{j-1}+1 & \ldots & (w_{j-1}+1)^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & w_j - 1 & \ldots & (w_j - 1)^d \end{pmatrix}$$

$$\vec{y}_j = (Y_{w_{j-1}}, \ldots, Y_{w_j-1})^T$$

则有：

$$\vec{\theta}_j = (X_j^T X_j)^{-1} X_j^T \vec{y}_j$$

$$C(w_{j-1}, w_j) = \|\vec{y}_j - X_j\vec{\theta}_j\|_2 \tag{1}$$

$d = 0$时，就是用常数去拟合；$d = 1$时，即为线性回归；$d = 2$时，即为二次回归；以此类推。通常取$d = 0$或$d = 1$，这是因为从经验上来看，社会情感的变化是呈线性的，而且这样模型会更简单，避免过拟合。

该方法一个应用的例子如图1所示。如果我们设定了参数$M = 2$, $d = 1$，即已确定使用线性回归将数据分为两段，算法最终会将断点设置在第6个时间点前，并对两边的数据分别进行拟合。而从直觉上来说，这种做法也是合理的，因为情感强度明显在第6个时间点有一个急剧的上升，极有可能在这个时刻社交媒体用户的情绪产生了突变。

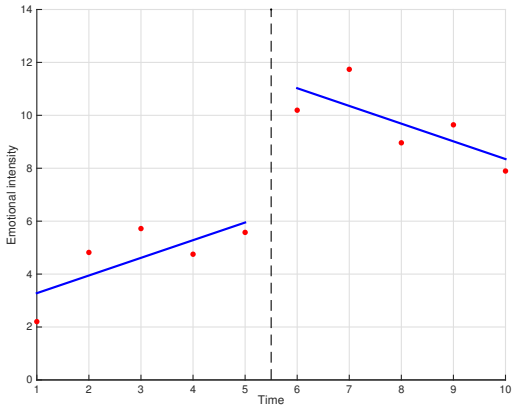

Figure 1: An example of piecewise linear regression

3.1.2 Solution

使用遍历法求解最优的$\vec{w}$，则算法复杂度为$O(N!)$。作者提出了一个动态规划算法，可以将复杂度减少到$O(MN^2)$。定义$F(T, k)$为将区间$[0, T)$划分为k个region的最优解对应的误差，则：

$$F(T, k) = \min_{x<T} \left\{F(x, k-1) + C(x, T)\right\} \tag{2}$$

通过$F(T, k)$的值，可以逆向求解出最优解：

$$w_j = \underset{x<w_{j+1}}{argmin} \left\{F(x, j) + C(x, w_{j+1})\right\} \tag{3}$$

可以通过添加正则化项的方法，使算法自动确定断点个数。注意到，之前的算法不允许前后两个断点重合，即必须满足$w_{j-1} < w_j$。现在将约束放松为$w_{j-1} \le w_j$，当$w_{j-1} = w_j$时，断点就减少了一个。我们希望断点尽量少，这样可以避免过拟合，所以当$w_{j-1} < w_j$时，对$eq(3)$的目标函数加入惩罚系数λ。于是，$eq(3)$被改写为：

$$w_j = \underset{x\le w_{j+1}}{argmin} \left\{F(x, j) + C(x, w_{j+1}) + \lambda \cdot \mathbb{I}(x \neq w_{j+1})\right\} \tag{4}$$

其中$\mathbb{I}(p) = \begin{cases} 1, & \text{p is true} \\ 0, & \text{p is false} \end{cases}$.

完整的算法如Algorithm 1所示。

3.1.3 Drawbacks

然而，在我们的实验中，我们发现，实际的数据集并不会这么理想。我们使用网上的Meme数据集，以及我们自己抓取的来自中国新浪微博的数据，发现社交媒体用户对于特定话题的情绪呈现出“钉子形”，如图2所示。这种数据带有严重的噪声，不呈线性，上述的算法无法应用。为了解决这个问题，我们提出了一种数据处理方式，用它来预处理数据，可以使上述算法在现实数据集上的效果得到显著提升。

Algorithm 1 Piecewise Regression Algorithm

Require: Y, M, λ
Ensure: $\vec{w}$

$N \leftarrow \text{length}(Y)$
function EmotBrkp(Y, M, λ)
  for i, j from 0 to N **do**
    Compute $C(i, j)$ according to $eq(1)$
  end for
  for T from 1 to N **do**
    for k from 1 to M **do**
      Compute $F(T, k)$ according to $eq(2)$
    end for
  end for
  for j from $M - 1$ to 1 **do**
    Compute w_j according to $eq(4)$
  end for
end function

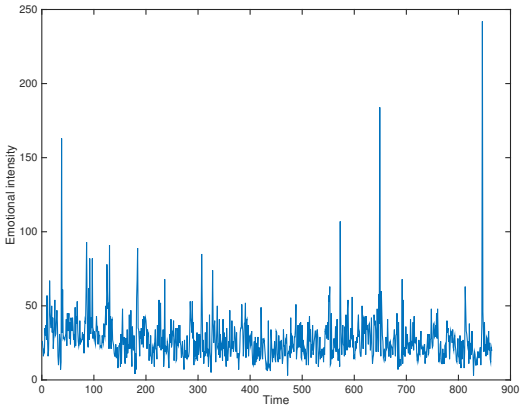

Figure 2: An example of real-world emotional dynamics from Meme dataset

3.2 Correlation Between News and Emotions

M. Tsytsarau等人在 [2]中研究了两个问题：一是事件与新闻之间的关系；二是新闻与情感之间的关系。对于第一个问题，他们采用了一种信号处理的方式。他们认为新闻发布量是媒体本身的响应函数与外部事件重要性卷积的结果。基于这个假设，他们将媒体的响应函数分为6种，并将事件分为两种，分别建立对应的数学模型。对于第二个问题，他们仅仅计算了情感强度与新闻量之间的相关函数：

$$\rho(s, n, \delta) = \frac{Cov[n(t), s(t+\delta)]}{\sigma_s \sigma_n}$$

其中，n为新闻发布量，s为情感强度，δ为时延(time lag)。首先，这种方法并没有直接计算出断点的位置，还要再挑选使得ρ函数值较大的δ，才能得到断点所在，而如何挑选，又牵涉到其它的问题。其次，情感强度的变化是否与新闻发布量存在相关性，这一点还有待更多的实验去证明。

4 Proposed Method

这一节中，我们以递进的方式，依次介绍三个版本的数据处理方式。我们用这些方法处理3.1节中提到的“钉子形”数据，将其转化成类线性分布的数据。然后就可以使用 [3]中的算法，对断点进行检测，收到显著效果。为方便起见，我们依然用Y表示原本的数据，而用Z表示处理过的数据。

4.1 Linear Addition

一种简单的思路是将Y累加。即：

$$Z_j = \sum_{i=1}^{j} Y_i$$

或：

$$Z_j = \begin{cases} Y_0 & (j = 0) \\ Z_{j-1} + Y_j & \text{(otherwise)} \end{cases}$$

请注意，为了防止数值溢出，我们先对Y进行了min-max归一化，即对于每个Y_j：

$$Y_j = \frac{Y_j - \min(Y)}{\max(Y) - \min(Y)}$$

原始数据中，情感强度只有在断点处会非常大，而在非断点处很小。可以预想到，这样处理后，Z会在断点处急剧上升，而在两个断点之间缓慢地线性上升。通过这种方式，原始数据变为了近似于分段线性函数的样子。然后我们再采用3.1节介绍的算法，也许就能检测出断点。

以图2中所示的数据为Y举例，经过处理后，Z如图3(a)所示。然而，数据整体呈现为一个线性函数，而不是分段的线性函数，断点“消失不见”了。这种方法存在重大的缺陷。

4.2 Linear Difference

分析linear addition方法效果不佳的原因，不难发现，是因为信息的冗余。由于断点远比非断点少，所以Z在非断点处的增长过多，从而掩盖了其在断点处的增长。解决方法是减少Z在非断点处的增长，令：

$$Z_j = \begin{cases} Y_0 & (j = 0) \\ Z_{j-1} + |Y_j - Y_{j-1}| & (otherwise) \end{cases}$$

为防止数值溢出，首先计算出所有的$|Y_j - Y_{j-1}|$，并对它们进行min-max归一化。

结果如图3(b)所示。Z已经不像图3(a)中呈现一个总体的线性，而是有一些较明显的突变点。这是因为在非断点处，前后两个数据点的差$|Y_j - Y_{j-1}| \approx 0$；而在断点附近，由于$Y_{\text{bkp}} \gg Y_{\text{non-bkp}}$，所以$|Y_j - Y_{j-1}| \gg 0$。这样，$Z$在断点处的增长比起在非断点处的增长，要显著的多。

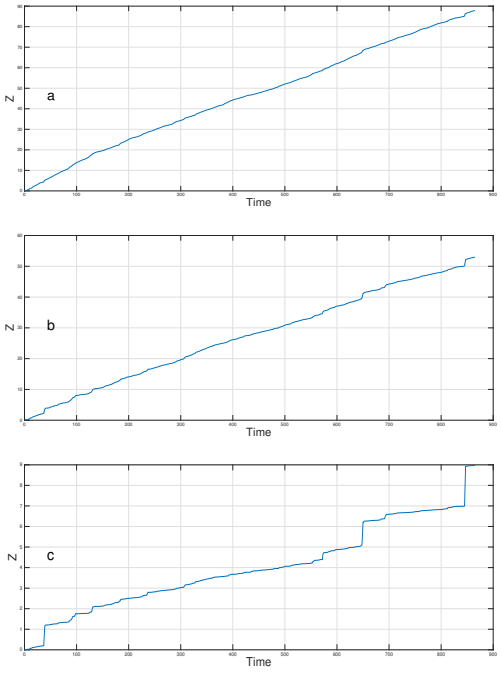

Figure 3: Results of three methods. Linear addition(a). Linear difference(b). Quadratic difference(c).

4.3 Quadratic Difference

Linear difference已经可以在实际的有噪声的数据集上取得不错的效果，但还可以改进。令:

$$Z_j = \begin{cases} Y_0 & (j=0) \\ Z_{j-1} + |Y_j - Y_{j-1}|^2 & (otherwise) \end{cases}$$

与linear difference相同，先对于$|Y_j - Y_{j-1}|$进行归一化。

结果如图3(c)所示。可以发现，此时Z中的断点已经非常明显，甚至可以靠肉眼观察出来。这是因为二次函数的性质。设α为断点附近的$|Y_j - Y_{j-1}|$，而β为非断点处的$|Y_j - Y_{j-1}|$。则有$\frac{\alpha}{\beta} > 1$，故$(\frac{\alpha}{\beta})^2 > \frac{\alpha}{\beta}$。平方后，$Z$在断点处与在非断点处的增长的比值变大了。故$Z$在断点处的增长占全部增长的比例增加，断点处的突变更加明显，断点也更容易检测出。

5 Mathematical Proof

6 Experimental Evaluation

6.1 Meme Dataset

6.2 Chinese Microblog

7 Conclusion

References

[1] L. Servi, S. B. Elson. A Mathematical Approach to Gauging Influence by Identifying Shifts in the Emotions of Social Media Users[J]. IEEE Transactions on Computational Social Systems, 2014, 1(4):180-190.

[2] M. Tsytsarau, T. Palpanas, M. Castellanos. Dynamics of news events and social media reaction. In SIGKDD, 2014:901-910.

[3] L. D. Servi. Analyzing social media data having discontinuous underlying dynamics[J]. Operations Research Letters, 2013, 41(6):581-585.