Kleinberg のバースト検知

稲毛惇人

2018 年 6 月 22 日

1　はじめに

時系列データにおいてイベントが急激に増加した，イベントの集中的な発生状態をバーストと呼ぶ．バーストを自動的に検出することができれば，結果としてバースト発生時の状況を効率よく解析することができるようになる．

Kleinberg は論文中に，2 つのバースト検知の手法を示しており，1 つ目は時々刻々と発生する連続的な時系列のイベントに対する手法，2 つ目は単位時間毎に発生したイベントを数え上げた離散的な時系列のデータに対する手法を示している [1]．どちらも 2 状態の有限オートマトンを利用し，前者は各イベントが到着した時間間隔の長さを利用してバースト検知を行うが，後者は全データ数，イベント発生数をそれぞれ利用してバースト検知を行い，その代表例として全文書中から特定のキーワードがどれだけ含まれているかを利用したバースト検知が可能になる．

今回は連続型のみ説明する．

2　イベント発生間隔による連続型バースト検知

2.1　指数分布

ほとんどの場合，メッセージなどのランダム到着を表現するためによく用いられるのが指数分布である．指数分布とは， ランダムなイベントの発生間隔を表すことができる分布であり，確率密度関数が式 1 で表されるような連続型確率分布を，**平均 μ の指数分布**という．x は連続型確率変数である．

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \qquad (x \geq 0) \tag{1}$$

2.2 メッセージ到着

メッセージ到着時間間隔を生成するための最も単純なモデルは，2.1 節で述べたように指数分布に基づく．メッセージは確率的に放出されるので，メッセージ i と $i+1$ との時間間隔 x は，パラメータ α に対する指数確率密度関数 $f(x) = \alpha e^{-\alpha x}$ となる．このモデルにおけるギャップの期待値は（指数分布の期待値と同じ）α^{-1} であるため，α をメッセージ到着率と呼ぶ.

2.3 2 状態モデル

バースト検知における状態は，定常状態，バースト状態の 2 つに限定されるため，q_0, q_1 の 2 状態を持つ確率的オートマトン A を考える．A が状態 q_0 のとき，メッセージは低速で放出され，確率密度関数 $f_0(x)$ に従って独立に分布される連続メッセージ間のギャップ x をもつ．$f_0(x)$ を式 2 に示す.

$$f_0(x) = \alpha_0 e^{-\alpha_0 x} \tag{2}$$

A が状態 q_1 にあるとき，$f_0(x)$ に従って独立に分布されるよりもギャップが短い間隔でメッセージが放出される．$f_1(x)$ を式 3 に示す.

$$f_1(x) = \alpha_1 e^{-\alpha_1 x} \tag{3}$$

最後に，メッセージ間で，A は，確率 $p_q \in (0, 1)$ で状態を変化させ，以前の放出および状態の変化とは無関係に，確率 $1 - p_q$ で現在の状態にとどまる．A は状態 q_0 で始まり，各メッセージ（最初のメッセージを含む）が放出される前に，A は確率 p_q で状態を変化させる．次に，メッセージが送出され，次のメッセージまでの時間間隔は，A の現在の状態に紐付いた分布に従う.

この 2 状態モデルを拡張し，$n+1$ 個のメッセージが到着したときの，メッセージ間隔 $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ を決定してみる．ここで，メッセージ間隔 x_i は必ず正を取る．x_k に対する状態を q_{i_k} とおくことで，$\boldsymbol{x}$ に対する状態を，状態シークエンス $\boldsymbol{q} = (q_{i_1}, q_{i_2}, ..., q_{i_n})$ と表現することができ，状態シークエンス $\boldsymbol{q}$ の条件付き確率を求めるために，ベイズの定理や事後確率最大化のベイズ決定法を用いることで実現できる.

2.3.1　2 状態オートマトンのコスト

各状態シークエンス $\boldsymbol{q}$ は ギャップの間隔に渡って確率密度関数が式 4 の形で導かれる．

$$f_{\boldsymbol{q}}(\boldsymbol{x}) = f_{\boldsymbol{q}}(x_1, ..., x_n) = \prod_{t=1}^{n} f_{i_t}(x_t) \tag{4}$$

もし， $q_{i_t} \neq q_{i_{t+1}}$ となるインデックスが $b = i_t$ で導かれるような，$\boldsymbol{q}$ における状態遷移数であるとき，その $\boldsymbol{q}$ の確率を式 5 に示す．ただし，A は状態 q_0 から開始するため，$i_0 = 0$ とする．

$$p(\boldsymbol{q}) = \left(\prod_{i_t \neq i_{t+1}} p_q \prod_{i_t = i_{t+1}} 1 - p_q \right) = p_q^b (1 - p_q)^{n-b} = \left(\frac{p_q}{1 - p_q} \right)^b (1 - p_q)^n \tag{5}$$

これから，条件付き確率 $p(\boldsymbol{q}|\boldsymbol{x})$ が求まるため，$p(\boldsymbol{q}|\boldsymbol{x})$ を 式 6 に示す．

$$p(\boldsymbol{q}|\boldsymbol{x}) = \frac{p(\boldsymbol{q}) f_{\boldsymbol{q}}(\boldsymbol{x})}{\sum_{\boldsymbol{q}'} p(\boldsymbol{q}') f_{\boldsymbol{q}'}(\boldsymbol{x})} = \frac{1}{\sum_{\boldsymbol{q}'} p(\boldsymbol{q}') f_{\boldsymbol{q}'}(\boldsymbol{x})} \left(\frac{p_q}{1 - p_q} \right)^b (1 - p_q)^n \prod_{t=1}^{n} f_{i_t}(x_t) \tag{6}$$

式 6 を最大化することによって，事後確率最大化を行うことができる．また，式 6 を最大化することは，自然対数をとって最小化することと等価であるため，$-\log(p(\boldsymbol{q}|\boldsymbol{x}))$ をベイズの定理を用いて式 7 に示す．

$$\begin{aligned} -\log(p(\boldsymbol{q}|\boldsymbol{x})) = \log \left(\sum_{\boldsymbol{q}'} p(\boldsymbol{q}') f_{\boldsymbol{q}'}(\boldsymbol{x}) \right) + b\log \left(\frac{1 - p_q}{p_q} \right) \\ -n\log(1 - p_q) + \left(\sum_{t=1}^{n} -\log(f_{i_t}(x_t)) \right) \end{aligned} \tag{7}$$

式 7 の第 2 項と第 4 項は $\boldsymbol{q}$ に無関係な変数であるため，$\boldsymbol{x}$ が与えられたときの $\boldsymbol{q}$ に対するコストとして表される．よって，式 7 を最小化するために最小化すべきコスト $c(\boldsymbol{q}|\boldsymbol{x})$ を式 8 に示す．

$$c(\boldsymbol{q}|\boldsymbol{x}) = b\log \left(\frac{1 - p_q}{p_q} \right) + \left(\sum_{t=1}^{n} -\log(f_{i_t}(x_t)) \right) \tag{8}$$

オートマトン A に対して， b を変化させることによって，A を現在の状態に固定する “慣性” として扱うことができるハイパーパラメータである．つまり，状態変化のしやすさを調整できる．

2.4 無限状態モデル

期間の長さ T に渡って到着する $n+1$ 個のメッセージ間隔を考える．メッセージが T に完全に均等な間隔で到着した場合，長さ $g = \frac{T}{n}$ のギャップで到着する．高強度バーストは長さ $\hat{g}$ よりもずっと小さいギャップに近づいている．つまり，可能なバーストの全範囲を捕捉するために，任意に小さいギャップサイズに対応できるような状態を持つ **無限状態オートマトン** を考えるべきである．ここで，前節までに説明してきたように，基本的な目標はコストが最小の状態 $\boldsymbol{q}$ を見つける手順と同様に，コストモデルを利用する．

2.4.1 無限状態オートマトンのコスト

ここで，完全に均等な間隔で到着したときの到着率 $\alpha_0 = \hat{g}^{-1} = \frac{n}{T}$ を伴い，指数分布の確率密度関数 f_0 を持つ定常状態 q_0 を有するオートマトンを考える．そのとき i $(i > 0)$ に対して，到着率 α_i を伴い f_i を持つ 状態 q_i が存在する．ハイパーパラメータ s を用いて α_i を式 9 に示す．

$$\alpha_i = \hat{g}^{-1} s^i \qquad (s > 1) \tag{9}$$

言い換えれば，状態 $q_0, q_1, ...$ から幾何学的に減少する到着間隔のギャップをモデル化する際に，i がより大きい値であるほどメッセージ到着の予想間隔が大きくなるように α_i が存在するということである．また，すべての i, j について，q_i から q_j への状態遷移にかかるコスト $\tau(i, j)$ が存在する．ここで，低強度バーストから高強度バーストに移行するコストについては, i, j の数値の差に比例する用に $\tau(\cdot, \cdot)$ を定義するが，高強度バーストから低強度バーストに降下するときのコストは 0 である．ハイパーパラメータ γ を用いて i, j によるコスト $\tau(i, j)$ を式 10 に示す．

$$\tau(i,j) = \begin{cases} (j-i)\gamma & (j > i) \\ 0 & (j < i) \end{cases} \qquad (\gamma > 0) \tag{10}$$

このオートマトンは，紐づくハイパーパラメータ s および γ とともに，$A^*_{s,\gamma}$ で表される．メッセージ到着間の正のギャップ $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ が与えられると，2.3.1 節で述べたものと同様に，最小化すべきコストを式 11 に示す．

$$c(\boldsymbol{q}|\boldsymbol{x}) = \left(\sum_{t=0}^{n-1} \tau\left(i_t, i_{t+1}\right) \right) + \left(\sum_{t=1}^{n} -\log\left(f_{i_t}\left(x_t\right)\right) \right) \tag{11}$$

状態 $q_0, q_1, ...$ に関して，$q_0, q_1, ...$ は無限に続くため，コストの最小値が明確に定義されていることを自動的に宣言することはできない．これまでのように，第 1 項を最小化す

ることは，状態遷移数が少ないこと，およびいくつかの異なる状態にのみ遷移することと一貫しており，第 2 項を最小化することは到着間隔に近い到着率を有する状態を遷移することと一致する．どちらの最小化も本質は，状態をあまり変化させることなく可能な限りギャップの間隔を追跡することである．

また，パラメータ s, γ に関しては，s がスケーリングパラメータと呼ばれ，各状態間距離がどの程度離れているかを調整するハイパーパラメータである．γ は，より高強度バーストへの状態遷移のコストが増えるため，バースト検知の感度といえる．原論文では，s は状態の離散的な間隔値が実数値のギャップを追跡することができる “解像度” (分解能ともいえる) とされ，γ はオートマトンが状態を変えることができる容易さを制御するとされている．

2.5 最小コスト状態シーケンス q の計算

メッセージ到着間の正のギャップ $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ が与えられたとき，$A^*_{s,\gamma}$ における状態シーケンス $\boldsymbol{q} = (q_{i_1}, ..., q_{i_n})$ のコスト $c(q|x)$ を最小化する アルゴリズムの問題について考える．最小値が明確に定義され，それを計算する手段を得るためには，オートマトンのバーストレベル状態数 $q_0, q_1, ...$ が自然数 k であるように有限の制約を定義することが有用である．そして，自然数 k に対して $q_0, q_1, ..., q_{k-1}$ を $A^*_{s,\gamma}$ から求め，得られた k 状態バーストレベルオートマトンを $A^k_{s,\gamma}$ で表す．2 状態オートマトン $A^2_{s,\gamma}$ は，前述の確率的 2 状態モデルと本質的に等価であることに留意されたい．$A^*_{s,\gamma}$ における最小コスト状態シークエンス $\boldsymbol{q}$ の計算量は，有限の制約に依存するが，それは計算可能であるらしい．原論文では難しいことではないとしている．

3　個人的なまとめ

- p_q は 0 か 1 であるため，式 5 は二項分布に従う．
- おそらく，$p(\boldsymbol{x}|\boldsymbol{q}) = f_{\boldsymbol{q}}(\boldsymbol{x})$ である．これは式 4 より確かに $\boldsymbol{q}$ が与えられたときの $\boldsymbol{x}$ に対する現在の状態の正しさである．
- 式 6 は，尤度関数 $L(\theta) = p(\boldsymbol{q}|\boldsymbol{x})$ として表すことで，それ以降の流れが最尤推定とほとんど等価にできる．最尤推定を用いるならば，対数尤度関数を偏微分することで θ を求めることができるため，式 4 を偏微分して極小値を求めれば，コストの式 8,10 が一瞬で求まる．
- 現在のバーストレベルの状態 q_o, q_1(無限状態オートマトンならば $q_0, q_1, ...$) が尤もらしい $q_0, q_1(q_0, q_1, ...)$ に近ければ近いほど，$f_{i_t}(x_t)$ は 1 に近い数値を表す．

参考文献

[1] J. Kleinberg, Bursty and hierarchical structure in streams, In *Proc. 8th SIGKDD*, pp.91-101, 2002.

[2] 株式会社 ALBERT potter, http://tech.albert2005.co.jp/391/, 2018/6/21 アクセス