2B-02

多項分布ストリームデータにおける異常区間検出の比較実験

山岸　祐己†　　斉藤　和巳†
† 静岡県立大学院 経営情報イノベーション研究科

1　はじめに

ストリームデータにおける異常検出手法は，Kleinberg のバースト検知 [1] をはじめとして幅広く研究がなされており [2, 3]，近年では異常（バースト）のモデル化の研究も進んでいる [4]．Kleinberg のバースト検知は，パラメータを多様に調整できることを考慮すると汎用性が高いように思えるが，パラメータを決定した時点で想定されるバースト（異常）区間の長さも固定されてしまうため，想定よりも長期的な異常，すなわち潜在的な異常を検出できない恐れがある．更に，これらのバースト検知や異常検出手法 [1, 2, 3] は，一般的に 1-カテゴリーの情報を扱うために設計されているので，複数カテゴリーの情報をもつストリームデータに対してはカテゴリー毎に独立して適応する必要があり，複数カテゴリー間の関係を考慮することが難しい．また，これらの手法は観測データの時刻間隔に依存するところが大きいため，観測時刻間隔がもともと一定になっているデータや，観測時刻間隔があまり変化しないデータには向いていない．よって我々は，上記の異常検出手法が扱うようなデータ形式に加え，カテゴリーが複数含まれていたり，観測時刻間隔が一定だったりするような多種データを扱うことを前提として，多様な区間長の異常（バースト）を検出する手法を提案する．提案手法は，代表的なバースト検知の研究 [1, 5] と同様の回顧的 (Retrospective) な観点から，データの確率分布に対する尤度最大化と尤度比検定に基づいて考案したものである．この手法におけるパラメータは，異常検出の感度を決定するものただ一つであるため，汎用性が高い設計となっている．本論文では，多様な区間長の意図的なバーストを含む人工データによる比較実験を行う．

2　問題設定

J-カテゴリーの情報を持つストリームデータについて，それらの確率分布における変化の異常と，異常の持続区間を推定することを考える．n 番目の観測ステップ t_n でのカテゴリー情報を s_n とすると，観測されたストリームデータ $\mathcal{D}$ は $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ のように表せる．ここで，$s_n =\in \{1, \cdots, J\}$ である．便宜上，s_n はカテゴリー $j \in \{1, \cdots, J\}$ の J-次元ベクトルダミー変数として

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

のように変換する．カテゴリー j が出現する確率 p_j が多項分布に従っていると仮定すると，ストリームデータ $\mathcal{D}$ に対する対数尤度は，出現確率のパラメータベクトル $\boldsymbol{p} = \{p_1, \cdots, p_J\}$ によって $\mathcal{L}(\mathcal{D}; \boldsymbol{p}) = \sum_{n=1}^{N} \sum_{j=1}^{J} s_{n,j} \log p_j$ のように計算できる．この対数尤度の最尤推定量は $\hat{p}_j = \sum_{n=1}^{N} s_{n,j}/N$ のように与えられる．

ここからは，このパラメータベクトル $\boldsymbol{p}$ が，確率分布の変化に従って K 箇所で分割された階段関数の形をとることを考える．すなわち，$k \in \{1, \cdots, K\}$ 番目の分割点の時刻 T_k $(t_1 < T_k < t_N)$ においてパラメータベクトルが $\boldsymbol{p}_k$ から $\boldsymbol{p}_{k+1}$ に切り替わることを仮定する．K 個の分割点を持つ集合 $C_K = \{T_1, \cdots, T_K\}$ とし，便宜上 $T_{k-1} < T_k, T_0 = t_1, T_{K+1} = t_N$ とする．更に，C_K による $\mathcal{D}$ の分割を $\mathcal{D}_k = \{n; T_{k-1} < t_n \leq T_k\}$, すなわち $\mathcal{N} = \{1, 2, \cdots, N\} = \{1\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{K+1}$ とし，$|\mathcal{D}_k|$ は $(T_{k-1}, T_k]$ における観測ステップ数とする．ここで，任意の $k \in= \{1, \cdots, K+1\}$ について $|\mathcal{D}_k| \neq 0$，つまり少なくとも一つの $t_n \in \mathcal{D}_k$ が存在することを条件とする．この時点で，この分割点検出問題は，部分集合 $C_K \subset \mathcal{T}$ の探索問題ということになる．ここで，$\mathcal{T}$ は観測ステップの集合 $\mathcal{T} = \{t_1, t_2, \cdots, t_N\}$ である．

分割点 C_K によって与えられる $\mathcal{D}$ の対数尤度は，パラメータベクトル集合 $\boldsymbol{P}_{K+1} = \{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{K+1}\}$ を用いて $\mathcal{L}(\mathcal{D}; \boldsymbol{P}_{K+1}, C_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}_k; \boldsymbol{p}_k)$ のように計算できる．つまり，この対数尤度の k と j についての最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{D}_k} s_{n,j}/|\mathcal{D}_k|$ となる．これらを代入すると，$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) = \sum_{k=1}^{K+1} \sum_{n \in \mathcal{D}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}$ が導ける．したがって，この分割点検出問題は $\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K)$ を最大化する C_K の探索問題に帰着できる．しかし，このままでは分割点集合 C_K の導入によってどれだけ対数尤度が改善したかという直接的な評価をすることができない．この問題において，分割を考慮しない，すなわちパラメータベクトルの変化が無いことを仮定したときの対数尤度からの改善度合いを評価することは重要となるため，尤度比最大化問題として目的関数を

Comparative Experiments of Anomaly Section Detection in Multinomial Distribution Data Streams
†Yuki YAMAGISHI　†Kazumi SAITO
†University of Shizuoka

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.

表 1: 3-カテゴリーの出現確率設定

	$p_{*,1}$	$p_{*,2}$	$p_{*,3}$
$\boldsymbol{p}_1$ ($\mathcal{D}_1$)	1/3	1/3	1/3
$\boldsymbol{p}_2$ ($\mathcal{D}_2$)	2/4	1/4	1/4
$\boldsymbol{p}_3$ ($\mathcal{D}_3$)	6/8	1/8	1/8
$\boldsymbol{p}_4$ ($\mathcal{D}_4$)	1/3	1/3	1/3

構築する．もし，パラメータベクトルに一切の変化がない，すなわち $C_0 = \emptyset$ とするならば，先程の対数尤度は $\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{1,j}$ となる．ここで，$\hat{p}_{1,j} = \sum_{n \in \mathcal{N}} s_{n,j} / |\mathcal{N}|$ である．よって，K 個の分割点を持つ場合と分割点を一切持たない場合との対数尤度比は $\mathcal{LR}(C_K) = \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0)$ のように与えられる．最終的に，この問題は $\mathcal{LR}(C_K)$ を最大化する C_K の探索問題に帰着できる．

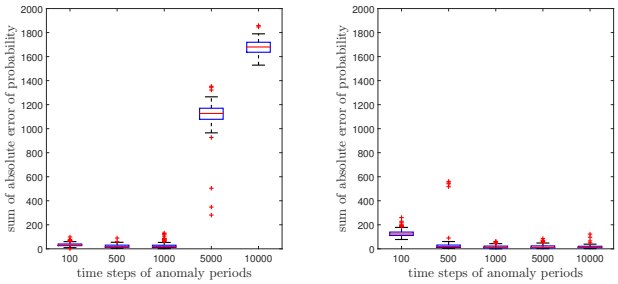
この問題は，貪欲法等を使えばある程度効率的に解けることが期待されるが，今回の実験では我々が既に提案した貪欲法と局所改善を交互に行うアルゴリズム [6] を使用する．

3 評価実験とまとめ

今回の提案手法と Kleinberg の手法 [1] との比較実験を行う．実験で用いるのは 3-カテゴリーをランダムに発生させた人工ストリームデータで，出現確率を階段関数の形で変化させたものである．真の分割点数は $K = 3$ であり，異常（バースト）区間は $\mathcal{D}_2$ と $\mathcal{D}_3$ とした．異常区間の両端部分の観測ステップ数を $|\mathcal{D}_1| = 10000$, $|\mathcal{D}_4| = 10000$ で固定したまま，異常区間の観測ステップ数は 5 パターン (100, 500, 1000, 5000, 10000) に変化させ，それぞれ 100 サンプルずつ生成した．Kleinberg の手法のパラメータ（今回は $s = 1.5, \gamma = 1.0$）に則って，出現確率の変化は表 1 のように設定した．なお，提案手法のアルゴリズムにおける χ^2 検定の有意水準は今回 $p = 0.0001$ である．

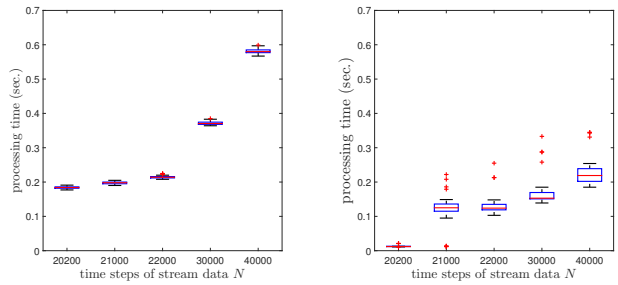
各手法による異常検出結果を基にした推定出現確率と，真の出現確率との絶対誤差の比較を図 1 に示す．図より，提案手法は極端に短い異常区間 $|\mathcal{D}_2| = |\mathcal{D}_3| = 100$ においては上手く機能していないが，それ以外の場合においての性能は良好である．逆に，Kleinberg の手法は，異常区間が長くなってくると上手く機能しないという特徴が見られる．更に，各手法に要した計算時間の比較を図 2 に示す．図より，提案手法は個々のサンプルの問題に依存して計算時間がブレてしまっているが，全体的な計算時間については，全ての場合において提案手法の方が短いことがわかる．

謝辞 本研究は， JSPS 特別研究員奨励費 15K00311 の支援を受けて行ったものである．



(a) Kleinberg の手法　(b) 提案手法

図 1: 真の出現確率との絶対誤差の比較



(a) Kleinberg の手法　(b) 提案手法

図 2: 計算時間の比較 (Intel(R) Xeon(R) X5690 @3.47GHz)

参考文献

[1] Jon M. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 91–101, 2002.

[2] Yunyue Zhu and Dennis E. Shasha. Efficient elastic burst detection in data streams. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 336–345, 2003.

[3] Aaron Sun, Daniel Dajun Zeng, and Hsinchun Chen. Burst detection from multiple data streams: A network-based approach. *IEEE Trans. Systems, Man, and Cybernetics, Part C*, Vol. 40, No. 3, pp. 258–267, 2010.

[4] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pp. 291–297, 2014.

[5] Russell C. Swan and James Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–56, 2000.

[6] Yuki Yamagishi, Seiya Okubo, Kazumi Saito, Kouzou Ohara, Masahiro Kimura, and Hiroshi Motoda. A method to divide stream data of scores over review sites. In *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, pp. 913–919, 2014.

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.