尤度に基づくストリームデータの異常検出手法の感度に関する実験

山岸　祐己†　　斉藤　和巳†

† 静岡県立大学経営情報イノベーション研究科　〒 422–8526 静岡県静岡市駿河区谷田 52–1
E-mail: †yamagissy@gmail.com, ††k-saito@u-shizuoka-ken.ac.jp

あらまし　ストリームデータにおける異常検出において，様々なバースト検知手法が考案されているが，一般的には 1-カテゴリー情報を扱うことを前提としているため，複数カテゴリー情報の分析には向いていない．特に代表的な手法は Kleinberg のバースト検知であるが，バースト検知そのものについても，想定されるバースト期間に応じて感度のパラメータを慎重に設定する必要がある．よって我々は，複数カテゴリー情報を持ったストリームデータにおける，多様な区間長の異常検出手法を提案する．提案手法は，複数カテゴリー情報の出現確率分布の変化に関する尤度最大化と尤度比検定に基づいて設計されており，ラフなパラメータ設定においても多様な区間長の異常（バースト）検出を可能としている．

キーワード　異常検出，バースト検知，ストリームデータ，時系列データ

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）の数理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，バースト状態を検知する感度のパラメータ (γ) を決定した時点で想定されるバースト（異常）区間の長さもある程度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察することが難しい．例えば，ある 2-カテゴリーがほぼ同時期にバーストを示したとしても，そのバーストの度合いは各カテゴリーによって相対的に決められているため，カテゴリー間の直接的な比較をすることはできない．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．Kleinberg の手法におけるパラメータは，バーストのスケーリングとバースト検知の感度の 2 つを要するのに対し，提案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 1 つであるため，汎用性が高い設計となっている．更に，感度のパラメータを固定した状態で，提案手法の方が様々な区間長の異常に対応できることを人工データによる比較実験で示す．

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

$$\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}, \tag{1}$$

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

$$s_{n,j} = \begin{cases} 1 \ \text{if} \ s_n = j; \\ 0 \ \text{otherwise.} \end{cases} \tag{2}$$

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

$$\mathcal{L}(\mathcal{D}; \boldsymbol{p}) = \sum_{n=1}^{N} \sum_{j=1}^{J} s_{n,j} \log p_j, \tag{3}$$

のように計算できる．式 (3) の最尤推定量は

$$\hat{p_j} = \frac{\sum_{n=1}^{N} s_{n,j}}{N}, \tag{4}$$

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

$$\mathcal{D}_k = \{n; T_{k-1} < t_n \leqq T_k\}, \tag{5}$$

すなわち

$$\mathcal{N} = \{1, 2, \cdots, N\} = \{1\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{K+1}, \tag{6}$$

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

$$\mathcal{L}(\mathcal{D}; \boldsymbol{P}_{K+1}, C_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}_k; \boldsymbol{p}_k), \tag{7}$$

のように計算できる．つまり，式 (7) の k と j についての最尤推定量は

$$\hat{p}_{k,j} = \frac{\sum_{n \in \mathcal{D}_k} s_{n,j}}{|\mathcal{D}_k|}, \tag{8}$$

となる．これらを式 (7) に代入すると，

$$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) = \sum_{k=1}^{K+1} \sum_{n \in \mathcal{D}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}, \tag{9}$$

が導ける．従って，この分割点検出問題は，式 (9) を最大化する C_K の探索問題に帰着できる．しかし，式 (9) 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，式 (9) は

$$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{1,j}, \tag{10}$$

となる．ここで，

$$\hat{p}_{1,j} = \frac{\sum_{n \in \mathcal{N}} s_{n,j}}{|\mathcal{N}|}, \tag{11}$$

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

$$\mathcal{LR}(C_K) = \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0), \tag{12}$$

のように与えられる．最終的に，この分割点検出問題は $\mathcal{LR}(C_K)$ を最大化する C_K の探索問題に帰着できる．

3. 分割点検出法

式 (12) を網羅的に解くと最適解が保証されるが，計算量が $O(N^K)$ となってしまうため，ある程度大きい N に対して $K \geqq 3$ となってしまうと，実用的な計算時間で解くことができない．したがって，我々は任意の K について解くための高速な解法を提案する．以下では，まず貪欲法 (A1) と局所探索法 (A2) を説明し，更にそれらを組み合わせた提案解法について説明する．

3.1 貪 欲 法

まず，貪欲法 (A1) の手順について説明する．このアルゴリズムは，バックトラッキングをしないデータの 2 分割の繰り返しである．つまり，既に選択された $(k-1)$ 個の分割点 C_{k-1} を固定したまま k 番目の分割点 T_k を C_{k-1} に新たに追加することを繰り返す．一般的に，$2(\mathcal{LR}(C_k) - \mathcal{LR}(C_{k-1}))$ は，N が十分に大きいとき χ^2 分布に従うことが知られているため，我々はこのアルゴリズムの終了条件として χ^2 検定を採用する．この χ^2 検定の危険率は事前に設定する必要がある．貪欲法アルゴリズムの手順は以下となる．

A1-1. $k = 1, C_0 = \emptyset$ のように初期化する．
A1-2. $T_k = \arg\max_{t_n \in \mathcal{T}} \{\mathcal{LR}(C_{k-1} \cup \{t_n\})\}$ を探索する．
A1-3. $C_k = C_{k-1} \cup \{T_k\}$ のように更新する．
A1-4. もし $2(\mathcal{LR}(C_k) - \mathcal{LR}(C_{k-1}))$ が，設定された危険率と自由度 $J-1$ における χ^2 の棄却限界値よりも小さければ，C_K を出力して終了する．
A1-5. $k = k + 1$ とし，A1-2 に戻る．

ここで，A1-3 での C_k の各分割点は，$T_{i-1} < T_i$ for $i = 2, \cdots, k$ を満たすように再インデックスする．明らかに，このアルゴリズムの計算量は $O(NK)$ と高速であるため，大規模な N にたいしても実用的な計算時間で結果を得ることが可能である．しかし，先程も説明したように，このアルゴリズムはバックトラッキングを行わないため，プアーな局所解に陥ってしまうことが危惧される．

3.2 局所探索法

次に，局所探索法 (A2) について説明する．このアルゴリズムは，A1 で得られた解 C_K から始まり，分割点の改善を 1 つずつ試みるものである．つまり，分割点 T_k を一度取り去り，残った $C_K \setminus \{T_k\}$ を固定して，よりよい尤度を得られる T'_k を探索することを $k = 1$ から K まで繰り返す．ここで，$\cdot \setminus \cdot$ は集合差を表す．もし，全ての k $(k = 1, \cdots, K)$ に対して分割点の置換が行われない，すなわち，全ての k に対して $T'_k = T_k$ ならば，これ以上の改善は望めないとして処理を終了する．局所探索法のアルゴリズムは以下となる．

A2-1. $k = 1, h = 0$ のように初期化する．
A2-2. $T'_k = \arg\max_{t_n \in \mathcal{T}} \{\mathcal{LR}(C_K \setminus \{T_k\} \cup \{t_n\})\}$ を探索する．
A2-3. もし $T'_k = T_k$ ならば $h = h + 1$ とし，さもなければ $h = 0$ として $C_K = C_K \setminus \{T_k\} \cup \{T'_k\}$ のように更新する．
A2-4. もし $h = K$ ならば C_K を出力して終了する．
A2-5. もし $k = K$ ならば $k = 1$, さもなければ $k = k + 1$ とし，A2-2 に戻る．

明らかに，このアルゴリズムの計算量は改善が終わらない限り増え続けてしまうが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍程度で終了することを我々は既に実験によって示している [6].

3.3 提 案 解 法

もし，計算量を最低限に抑えることを目的として，単純に貪

欲法アルゴリズムと局所探索法アルゴリズムを組み合わせると，

C1. A1 で C_K を得る．
C2. A2 で C_K を改善する．

となる．確かに，これだけでも十分な近似解が期待できるが，分割点数 K が貪欲法アルゴリズムによって決定されてしまうため，問題に対して不適切な分割点数のまま局所改善を行ってしまう恐れが大いにある．したがって我々は，不必要な分割点は極力追加せず，且つ必要な分割点は極力追加することを目的とした，アルゴリズムの反復的な組み合わせを提案する．提案解法の手順は以下となる．

P1. A1-1 から処理を開始する．
P2. A1-4 の処理後に $k \geqq 2$ ならば，C_k を C_K として出力する．
P3. C_K を A2 で改善し，改善した C_K を C_k として出力する．
P4. A1-5 から処理を再開させ，ステップ I2 へ戻る．

この手順では，分割点が追加される度に局所探索法アルゴリズムを行うため，更なる計算量の増加が予想されるが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍から十数倍程度で終了することを我々は既に実験によって示している [6]．結局のところ，提案解法において事前に設定が必要となるのは，貪欲法アルゴリズムにおける χ^2 検定の危険率のみであり，これが分割点数を大きく左右するため，本手法における感度のパラメータということになる．

4. 実　　験

4.1 人工データによる比較実験

我々は，ストリームデータの異常検出において，尤度の最大化が有効な方法であると仮定し，その最適化問題についての効率的な解法を提案した．ここでは，隠れマルコフモデルを用いた Kleinberg([1]) の手法を技術水準とし，提案手法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 の手法のスケーリングパラメータ（今回は $s = 1.5$ ）に則って，出現確率の変化は表 1 のように設定した．なお，Kleinberg の手法の感度のパラメータは $\gamma = 1.0$, 提案手法のアルゴリズムにおける χ^2 検定の危険率（感度のパラメータ）は $p = 0.0001$ で固定した．

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 から 4 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

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

上手く機能しないという特徴が見られる．感度のパラメータは固定しているため，パラメータを変化させてさらに検証する余地はあるが，異常区間の長さが予測できない状況を考慮すると，提案手法のほうが多様な区間長の異常に適応できる可能性が高いと言える．更に，各手法に要した計算時間の比較 (Intel(R) Xeon(R) X5690 @3.47GHz) を図 5,6 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．

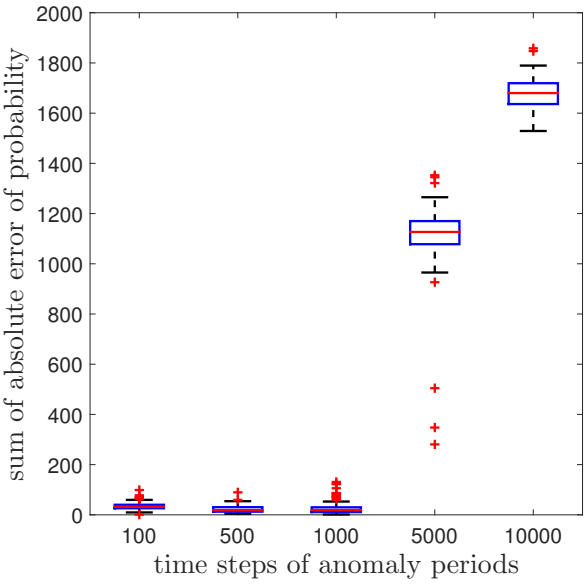

図 1　真の出現確率との絶対誤差の比較 (Kleinberg の手法)

参考までに，各異常区間における両手法の代表的な結果を図 7 から 16 に示す

4.2 現実データによる実験

我々は，国内の大規模化粧品レビューサイト “@cosme” (注1) から，各アイテムにつけられた今までのレビューの点数情報を取得し，最も多くのレビューを有する 2 アイテム (“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” と “Conditioner Essential (Albion)”) に対して提案手法を適応した．人工データによる実験のときと同様，χ^2 の危険率は $p = 0.0001$ だが，@cosme のレビュー点数の範囲は 0 から 7 ($j = 1, \cdots, 8$ として考える) であるため，自由度は $J - 1 = 7$ となる．図 17 と 18 に 2 アイテムの得点確率分布の分割結果を示す．両図より，提案手法は適切な分割数と様々な分割区間長によって，レビュー得点の確率分

(注1)：http://www.cosme.net/

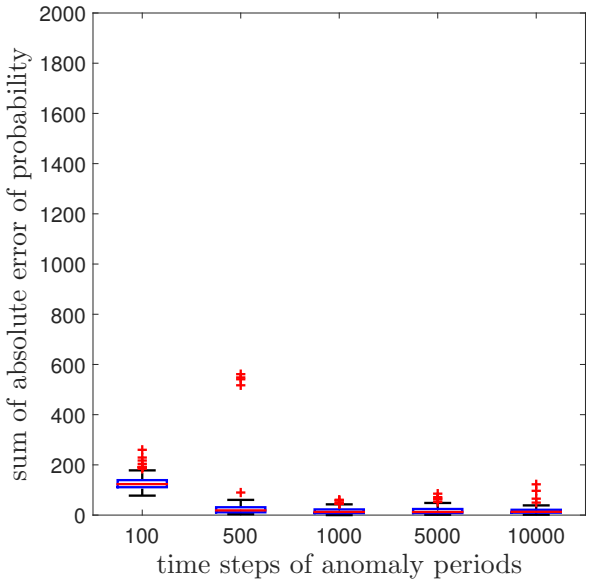

図 2　真の出現確率との絶対誤差の比較 (提案手法)

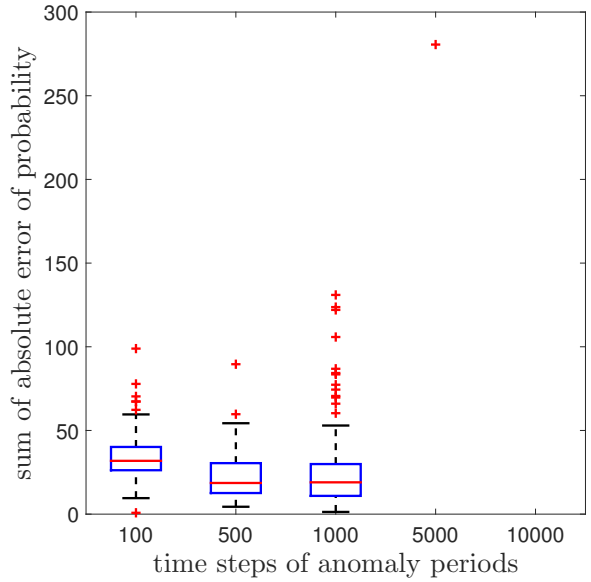

図 3　真の出現確率との絶対誤差の比較の詳細 (Kleinberg の手法)

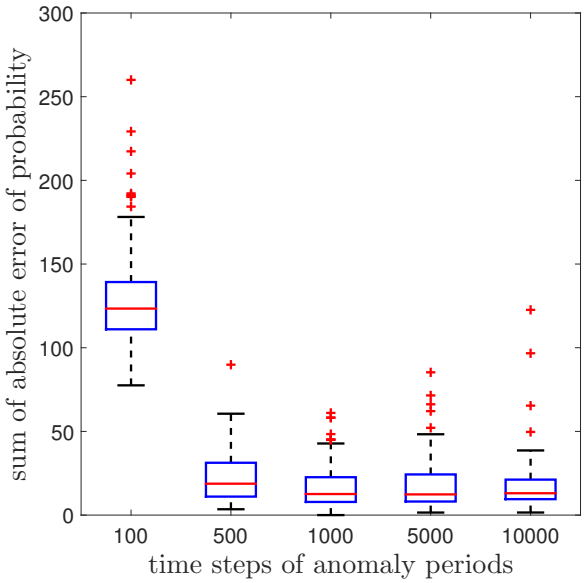

図 4　真の出現確率との絶対誤差の比較の詳細 (提案手法)

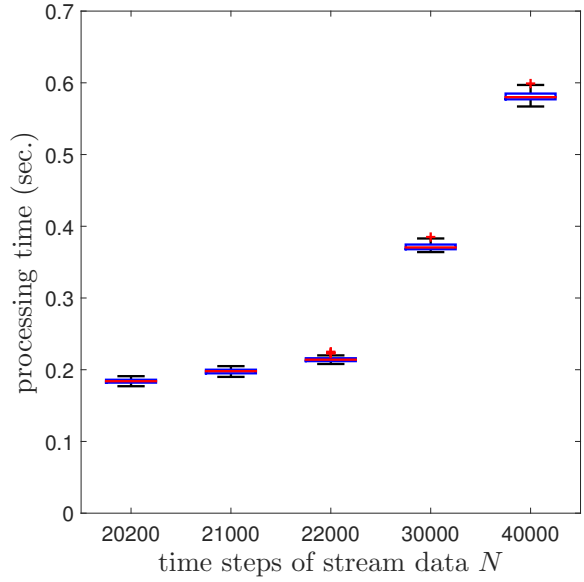

図 5　計算時間の比較 (Kleinberg の手法)

布における変化を表現できているように見える．更に，分割点が集中している期間や短い区間は得点分布の信頼性が有意に低く，長い区間は得点分布の信頼性が有意に高いということも自然に考えられる．

5. おわりに

代表的な異常（バースト）検出手法特有の問題を補うため，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した．我々は，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た問題設定を行い，その問題を高速且つ高精度に解くための解法を提案した．人工データを使った Kleinberg の手法との比較実験では，両手法の異常検出に関する感度のパラメータを固定した状態で，提案手法の方が様々な区間長の異常に対応できることを示した．今回の比較実験では，計算時間においても提案手法が僅かに優れている結果となった．現実データを使った実験では，レビュー点数の分布変化の視覚化に成功し，得点分布の信頼性指標における有用性を示した．

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．

文　献

[1] Jon M. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference*

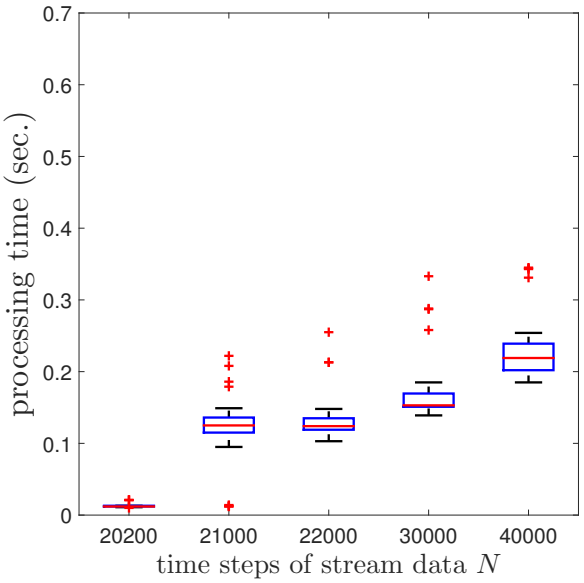

図 6　計算時間の比較 (提案手法)

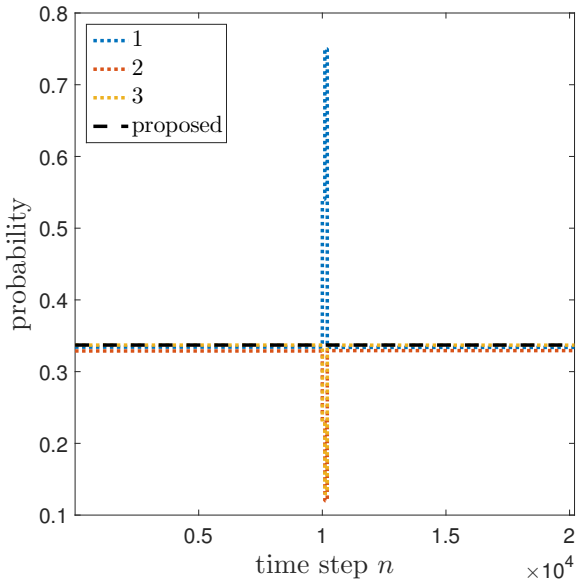

図 8　異常区間の観測ステップ = 100 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

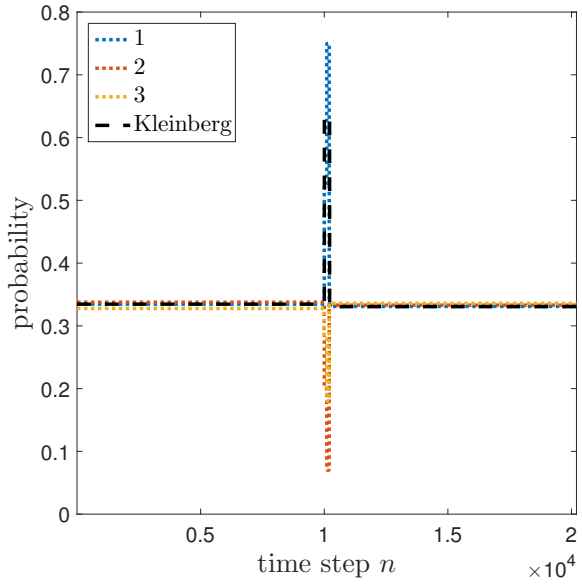

図 7　異常区間の観測ステップ = 100 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

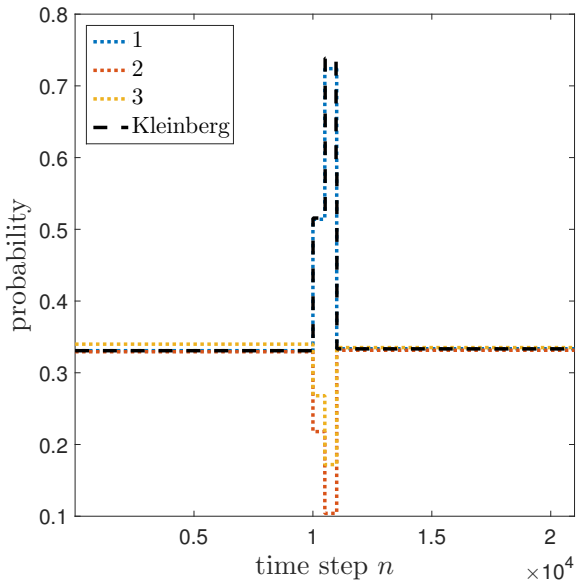

図 9　異常区間の観測ステップ = 500 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

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.

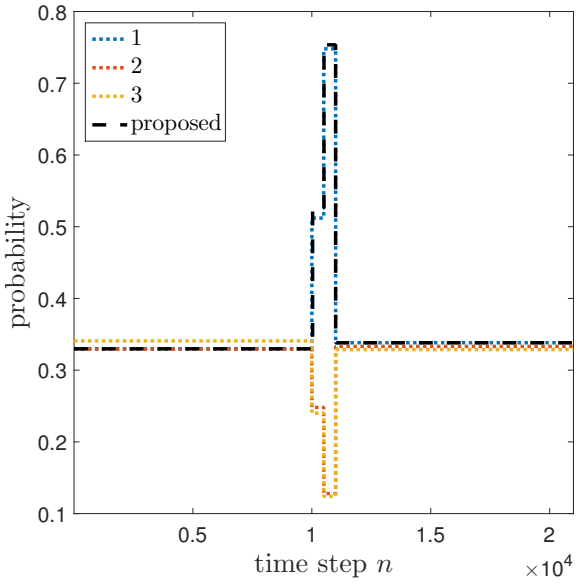

図 10　異常区間の観測ステップ = 500 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

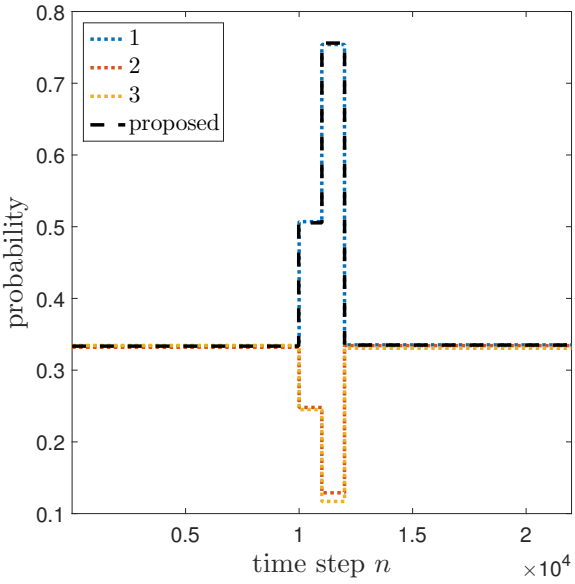

図 12　異常区間の観測ステップ = 1000 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

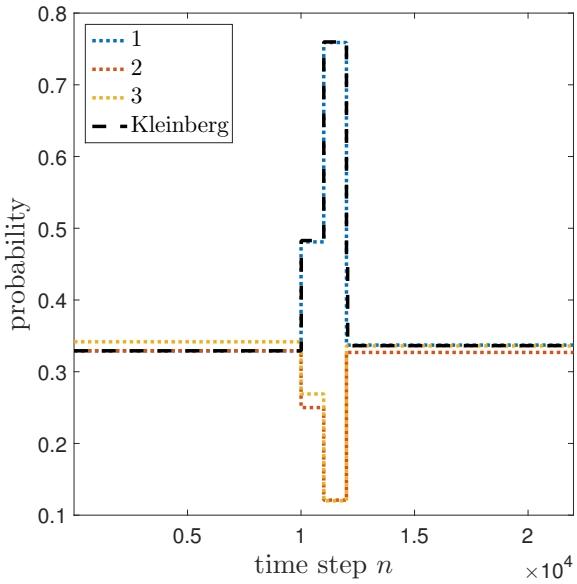

図 11　異常区間の観測ステップ = 1000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

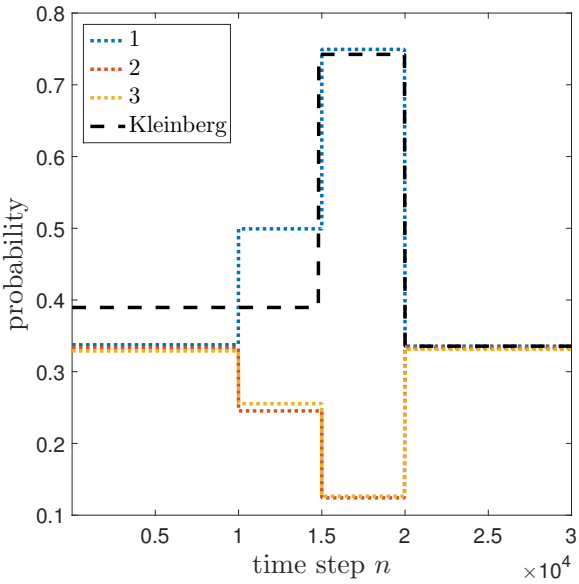

図 13　異常区間の観測ステップ = 5000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

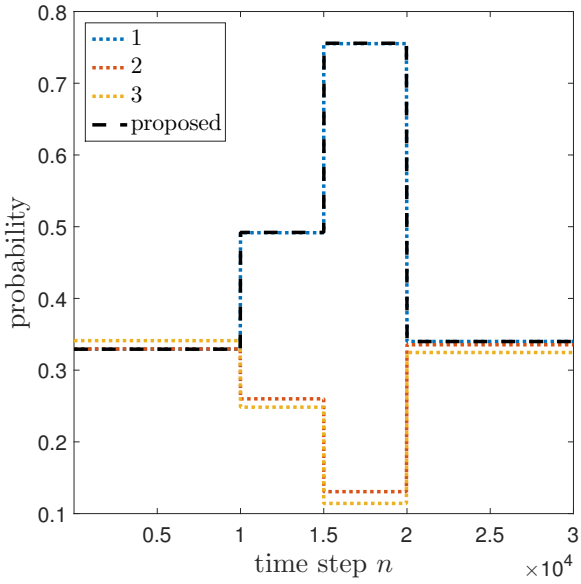

図 14　異常区間の観測ステップ = 5000 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

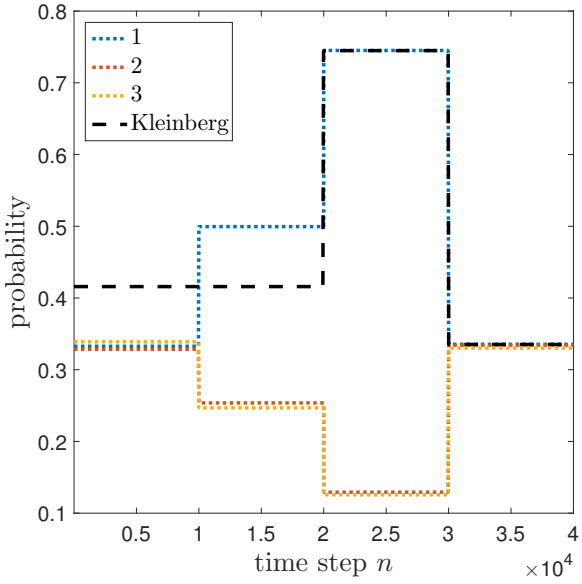

図 15　異常区間の観測ステップ = 10000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

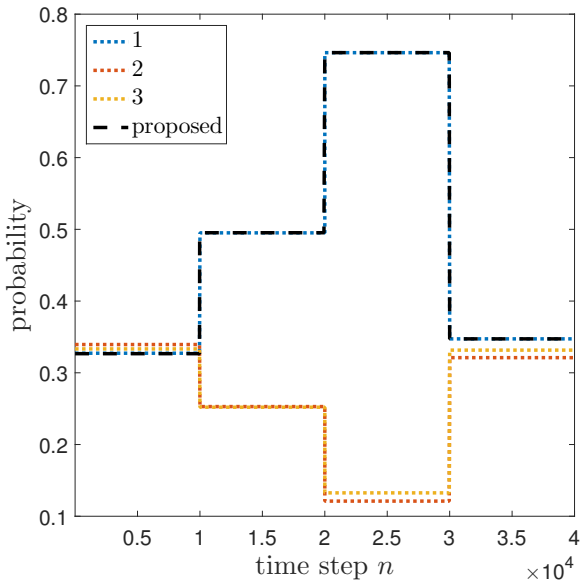

図 16　異常区間の観測ステップ = 10000 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

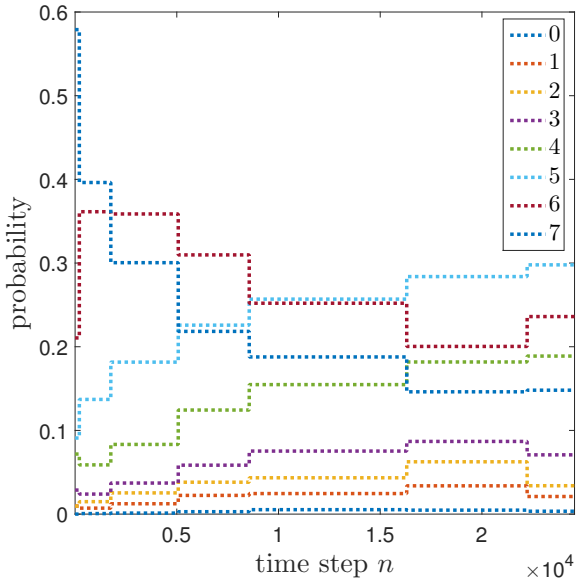

図 17　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” の結果

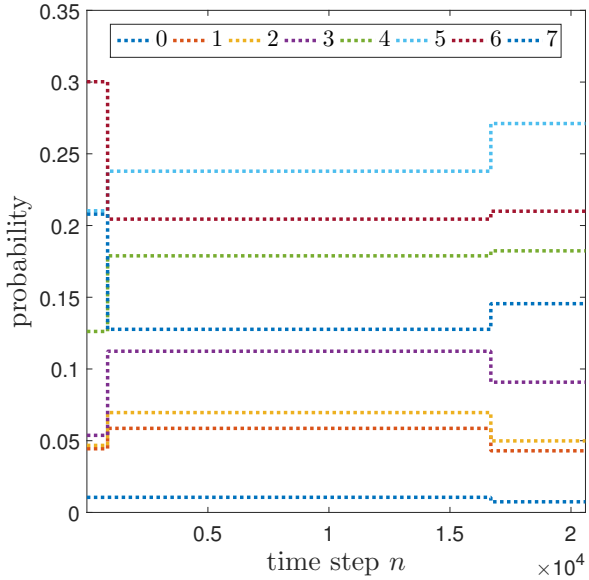

図 18　“Conditioner Essential (Albion)” の結果