情報処理学会論文誌　数理モデル化と応用　Vol.11　No.1　45–52 (Mar. 2018)

多群出現順位統計量に基づく時系列データの変換

山岸 祐已[1,a)]　岩﨑 清斗[2,b)]　斉藤 和巳[1,c)]

受付日 2017年8月28日，再受付日 2017年10月16日，
採録日 2017年11月23日

概要：データカテゴリの時系列的変化を明確に示し，それらを複数カテゴリ間で比較することを目的として，出現順位を用いた統計量によるデータ変換手法を提案する．ここでの変換手法は，勢力変化可視化法と有意な勢力変化をするカテゴリ検出法である．出現情報の時系列データにおける代表的な変換手法としてはKleinbergのバースト検知がよく知られているが，継続的な傾向分析や，複数カテゴリ間の比較には向いていない．よって我々は，出現傾向の指標として出現順位統計量を考え，多群を扱えるように拡張した手法を提案する．提案法は，出現情報を徐々に変化する傾向指標として変換するため，長期的な傾向変化をとらえやすく，また，各カテゴリの傾向指標は他のカテゴリすべてを基準としているため，任意の複数カテゴリ間の比較が容易である．評価実験では，人工データと現実データを用い提案法の有効性を検証する．

キーワード：順位和検定，時系列データ，傾向分析，バースト検知，one-against-all

Converting of Stream Data Based on Multi-category Appearance Order Statistics

Yuki Yamagishi[1,a)]　Kiyoto Iwasaki[2,b)]　Kazumi Saito[1,c)]

Received: August 28, 2017, Revised: October 16, 2017,
Accepted: November 23, 2017

Abstract: We propose a data conversion method by using appearance order statistics with the aim of clarify the temporal changing of data categories and compare them between multi-categories. Here, this conversion method is a visualization of a power struggle of categories and detection method of categories with a significant power change. Although Kleinberg's burst detection is well known as a representative conversion method in time series data of appearance information, this method is not suitable for continuous trend analysis or comparison between multi-categories. Therefore, we consider the appearance order statistics as a trend indicator and extend the statistics to be able to deal with multi-category. Since the proposed method converts appearance information as a trend indicator which changing gradually, it can easy to capture long-term trend changes. In addition, since the trend indicators of each category are based on all the other categories, it is easy to compare between arbitrary multi-categories. In the evaluation experiment, we verify the effectiveness of the proposed method using synthetic data and real data.

Keywords: rank sum test, time series data, trend analysis, burst detection, one-against-all

1. はじめに

本論文では，データカテゴリの出現傾向を明確に示し，それらを複数カテゴリ間で比較することを目的として，時間方向の順序を用いた多群順位統計量による傾向分析手法を提案する．時系列データの研究では，現時点の状況解析や将来予測に焦点を当てているものもあるが，本研究は，Kleinberg [1] や Swan ら [2] と同様に，回顧的（retrospec-

[1] 静岡県立大学
University of Shizuoka, Shizuoka 422–8526, Japan
[2] 静岡県工業技術研究所
Industrial Research Institute of Shizuoka Prefecture, Shizuoka 421–1221, Japan
[a)] yamagissy@gmail.com
[b)] kiyoto1_iwasaki@pref.shizuoka.lg.jp
[c)] k-saito@u-shizuoka-ken.ac.jp

tive）な枠組みによる時系列データからの情報抽出，すなわち，過去に何が起きどのような変化をしていたかということに焦点を当てている研究と類似している．

たとえば，Kleinberg の研究は，文書ストリーム内のトピックの出現をバーストとして表現し，その入れ子構造を推定することによって，ある期間におけるトピックのアクティビティを要約し，それらの分析を容易にしている．この Kleinberg の手法は，バーストが自然に状態遷移として現れる隠れマルコフモデルを使用しており，電子メールメッセージの階層構造を識別することができている．出現頻度が大きく変化する時系列データについては，既存のバースト検出技術 [1] とともに，ウィンドウに基づく手法 [3] や複数ストリームを対象とした手法 [4] なども適応可能であるが，出現頻度がほぼ一定，もしくは大きな変化がないものについては，これら既存手法の有効性は低いことが予想される．さらに，既存のバースト検出技術は，単一カテゴリのバーストを検出するものであり，複数カテゴリとその分布の変化に着目していないため，複数カテゴリの傾向変化を検出することには向いていない．

一方，Swan らの研究は，仮説検定に基づいた時間経過による特徴出現モデルを使用し，コーパス内の主要トピックに対応する情報をクラスタとして生成することに成功している．本研究も同様に，過去に起こった現象を理解するという目的を持っているが，あくまで出現傾向を指標化した時系列データへの変換を行うものであるため，このような研究のモチベーションとも離れている．

よって我々は，出現傾向の指標として時間方向の順位統計量を考え，多群を扱えるように拡張した手法を提案する．提案法は，新たに出現したオブジェクトとともに徐々に変化する指標を与えるため長期的な傾向変化をとらえやすく，また，各カテゴリの指標は他のカテゴリすべてを基準としているため，任意の複数カテゴリ間の比較が容易である．人工データを用いた実験では，ナイーブな手法とともに，時系列データのバースト検出の最先端技術として Kleinberg の手法 [1] を比較対象とし，提案法による定量的評価の妥当性を検証する．現実データを用いた実験でも，同様の比較手法を用いて提案法の性能と特性を評価する．

2. 提案法

2.1 問題設定

出現時刻で昇順ソートされたオブジェクト集合と，それらが有するカテゴリ集合をそれぞれ $\mathcal{K}$ と $\mathcal{J}$ とする．ここで，それぞれの要素数は $K = |\mathcal{K}|$ と $J = |\mathcal{J}|$ とし，各要素は整数と同一視されるとする．つまり，$\mathcal{K} = \{1, \cdots, k, \cdots, K\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ である．なお，オブジェクト k は最古のものが 1，最新のものが K となるよう，出現順に並んでいるものとする．このとき，オブジェクト k がカテゴリ j を有する場合は 1，それ以外の場合は 0 となっている J 行 K 列の行列を $Q(q_{j,k} \in \{0,1\})$ とすると，オブジェクト k が有するカテゴリ数は $t_k = \sum_{i=1}^{J} q_{i,k}$，オブジェクト k までのカテゴリ j の出現数は $I_{j,k} = \sum_{i=1}^{k} q_{j,i}$，オブジェクト k までの全カテゴリの総出現数は $I_k = \sum_{i=1}^{J} I_{i,k}$ のように表せる．

いま，オブジェクトに付随してカテゴリが出現するとし，以降では，オブジェクト出現からカテゴリ出現へと視点を変える．このとき，オブジェクト k が唯一のカテゴリのみ有する $t_k = 1$ の場合では，オブジェクト k に付随して出現したカテゴリ j の出現順位は $I_{k-1} + 1$ であるが，複数のカテゴリを有する $t_k > 1$ の場合では，平均順位を考えなければならないため，その出現順位は $r_k = I_{k-1} + (1 + t_k)/2$ となる．ここでの目的は，オブジェクトとカテゴリの集合が与えられたとき，出現順位の値が大きい（新しい），または逆に小さい（古い）オブジェクトが有意に多く含まれるカテゴリを定量的に評価する指標の構築である．以下には，Mann-Whitney の統計量 [5] に基づく自然な拡張法を示す．

2.2 多群出現順位統計量

Mann-Whitney の二群順位統計量 [5] を多群に拡張し，オブジェクトの出現順位に適用する方法について述べる．いま，カテゴリ j に着目すれば，このカテゴリに属するオブジェクト集合 $\{k \in \mathcal{K} : q_{j,k} = 1\}$ と，このカテゴリに属さないオブジェクト集合 $\{k \in \mathcal{K} : q_{j,k} = 0\}$ の二群に分割することができる．よって，Mann-Whitney の二群順位統計量に従い，次式により，カテゴリ j に対し出現順位統計量の z-score を求めることができる．

$$z_j = \frac{u_j - \mu_j}{\sigma_j}. \tag{1}$$

ここで，統計量 u_j，出現順位の平均 μ_j，および，その分散 σ_j^2 は次のように計算される．

$$u_j = \sum_{i=1}^{K} r_i q_{j,i} - \frac{I_{j,K}(I_{j,K}+1)}{2}, \tag{2}$$

$$\mu_j = \frac{I_{j,K}(I_K - I_{j,K})}{2}, \tag{3}$$

$$\sigma_j^2 = \frac{I_{j,K}(I_K - I_{j,K})}{12}\left((I_K+1) - \sum_{i=1}^{K}\frac{t_i^3 - t_i}{I_K(I_K-1)}\right). \tag{4}$$

すなわち，u_j は順位和に基づく統計量であり，その平均と分散が μ_j と σ_j^2 である．ただし，各オブジェクトが複数のカテゴリを有しえないケースでは，式 (4) の t_i を含む項，すなわち平均順位を扱うための補正値の計算は不要である．この多群順位統計量は，基本的には 2 クラス分類器の SVM（Support Vector Machine）[6] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる．

以上より，式 (1) で求まる z-scorez_j により，最新オブジェクト K までの各カテゴリ j が，出現順位の値が大きい（新しい），または逆に小さい（古い）オブジェクトを有意に多く含むかを定量的に評価することができる．よって，任意のオブジェクト k 出現時における同様の定量的評価ができるよう，上記の z-score を拡張する．任意のオブジェクト k に対応した次式により，オブジェクト k までのカテゴリ j に対し z-score$z_{j,k}$ を求めることができる．

$$z_{j,k} = \frac{u_{j,k} - \mu_{j,k}}{\sigma_{j,k}}. \tag{5}$$

ここで，統計量 $u_{j,k}$，出現順位の平均 $\mu_{j,k}$，および，その分散 $\sigma^2_{j,k}$ は次のように計算される．

$$u_{j,k} = \sum_{i=1}^{k} r_i q_{j,i} - \frac{I_{j,k}(I_{j,k}+1)}{2}, \tag{6}$$

$$\mu_{j,k} = \frac{I_{j,k}(I_k - I_{j,k})}{2}, \tag{7}$$

$$\sigma^2_{j,k} = \frac{I_{j,k}(I_k - I_{j,k})}{12}\left((I_k+1) - \sum_{i=1}^{k}\frac{t_i^3 - t_i}{I_k(I_k-1)}\right). \tag{8}$$

先ほどと同様，各オブジェクトが複数のカテゴリを有しえないケースでは，式 (8) の t_i を含む項，すなわち平均順位を扱うための補正値の計算は不要である．

以上より，式 (5) で求まる z-score$z_{j,k}$ により，オブジェクト k までの各カテゴリ j が，出現順位の値が大きい（新しい），または逆に小さい（古い）オブジェクトを有意に多く含むかを定量的に評価することができる．すなわち，この $z_{j,k}$ が正の方向に大きければ大きいほど，オブジェクト k の直近での出現が有意に多いということであり，カテゴリ j の勢力が伸びていることになる．逆に，$z_{j,k}$ が負の方向に大きいということは，過去に比べて勢力が衰えていることになる．また，異常検知を目的として有意水準を設定すれば，$z_{j,k}$ から求まる有意確率を使った仮説検定が可能である．

2.3　時系列データ変換アルゴリズム

まず，式 (5) で求まる z-score$z_{j,k}$ の計算量はすべてのオブジェクトとすべてのカテゴリについて算出した場合でも $O(KJ)$ と高速であり，オンライン処理においても新たに追加されたオブジェクトごとに $O(J)$ の計算量しかかからない．詳細には，まず，統計量 $I_{j,k}$ と I_k については，$I_{j,k+1} \leftarrow I_{j,k} + g_{j,k+1}$ と $I_{k+1} \leftarrow I_k + t_{k+1}$ で更新できる．一方，$M_{j,k}$ と M_k を次式で定義すれば，

$$M_{j,k} \leftarrow \sum_{i=1}^{k} r_i q_{j,i}, \tag{9}$$

$$M_k \leftarrow \sum_{i=1}^{k}\frac{t_i^3 - t_i}{I_k(I_k-1)}, \tag{10}$$

$M_{j,k+1}$ と M_{k+1} を次式で定義すれば，

$$M_{j,k+1} \leftarrow M_{j,k+1} + r_{k+1} q_{j,k+1}, \tag{11}$$

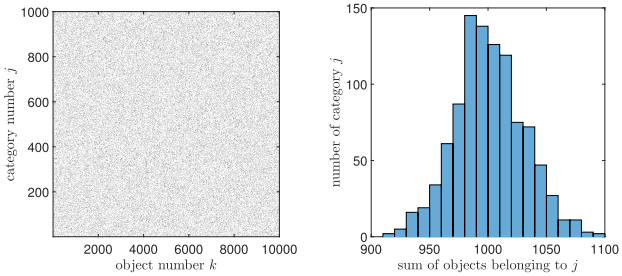
$$M_{k+1} \leftarrow \frac{I_k(I_k-1)M_k + t_{k+1}^3 - t_{k+1}}{I_{k+1}(I_{k+1}-1)}. \tag{12}$$

したがって，式 (5) から (8) に基づき，すべての j に対して，$u_{j,k+1}$，$\mu_{j,k+1}$，$\sigma^2_{j,k+1}$，および，$z_{j,k+1}$ を $O(J)$ で求めることができる．

最後に，求めた z-score$z_{j,k}$ によるデータ変換法として，勢力変化可視化法と，有意な勢力変化をするカテゴリ検出法を提案する．まず，勢力変化可視化法では，各カテゴリ $j \in \mathcal{J}$ に対し，オブジェクト番号 k を横軸に，z-score$z_{j,k}$ を縦軸にした曲線として可視化する．この可視化により，時間につれて，各カテゴリの勢力変化を視覚的に把握できることが期待できる．一方，カテゴリ検出法では，ユーザ指定のオブジェクト出現時刻 k と p 値に対し，有意な勢力変化を起こしているカテゴリを出力する．この検出により，与えられたデータにおいて注目すべきカテゴリを知ることができる．

3.　人工データによる実験

ここでは，異常検知の側面から，提案法の z-score がトレンド分析の定量的評価法として有効であるかどうかを検証する．検証に用いたデータは，基本パターンカテゴリ 1,000 個と異常パターンカテゴリ 1 個を有する 10,000 オブジェクトを出現確率に従ってランダムに生成したものである．ここで，オブジェクト k での各カテゴリの出現確率を α_k とし，基本パターンは固定確率 $\alpha_k = 0.1$ とする．基本パターンカテゴリ 1,000 個を有する 10,000 オブジェクトを生成した例を**図 1** (a) に示す．図の横軸はオブジェクトの時系列順序（昇順）k を，縦軸はカテゴリ j を，黒点はカテゴリの



(a) 基本パターンカテゴリの生成例（黒点 $q_{j,k} = 1$，白点 $q_{j,k} = 0$）
(a) Example of generated data of the basic pattern category. (black dots $q_{j,k} = 1$, white dots $q_{j,k} = 0$)

(b) 基本パターン生成例におけるカテゴリ出現数の度数分布
(b) Frequency distribution of categories in an example of the basic pattern.

図 1　人工データにおける基本パターンの生成例
Fig. 1　Example of generated data of the basic pattern in synthetic data.

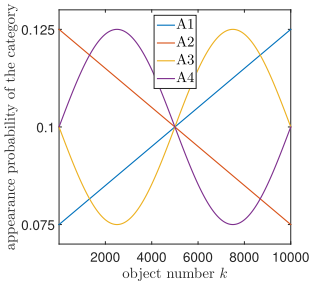
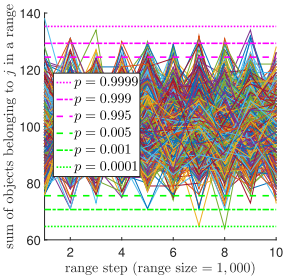


図 2　人工データにおける各異常パターンの確率変動

Fig. 2　Probability fluctuations of the anomaly patterns in synthetic data.

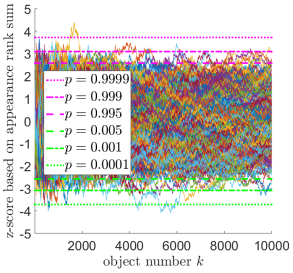
出現を示す $q_{j,k} = 1$ を，白点は $q_{j,k} = 0$ をそれぞれ表している．また，この基本パターンカテゴリの生成例におけるカテゴリ出現数の度数分布を図 1 (b) に示す．図の横軸はカテゴリ j の出現数，すなわちカテゴリ j に属するオブジェクト数を，縦軸はカテゴリ数をそれぞれ示す．両図より，基本パターン生成例の各カテゴリの出現数は，成功確率 0.1，試行回数 10,000 の二項分布の平均 $10000 \times 0.1 = 1000$ と標準偏差 $\sqrt{10000 \times 0.1 \times (1-0.1)} = 30$ におおよそ従っていることが分かる．実際，この生成例を用いたコルモゴロフ–スミルノフ検定（二項分布：試行回数 10,000，成功確率 0.1）において，有意確率は 0.1763 となったため，一般的な有意水準 0.05 を基準とすると，仮定した二項分布に従っているといえる．この基本パターンに対し，同様の出現数分布を想定しつつ，出現確率がわずかに変化するような異常パターンを考える．今回，異常パターンは 4 種類（A1，A2，A3，A4）とし，それぞれの異常パターンにおける出現確率は $\alpha_k = 0.075 + 0.05(k/10000)$，$\alpha_k = 0.125 - 0.05(k/10000)$，$\alpha_k = 0.1 + 0.025\sin(-2\pi k/10000)$，$\alpha_k = 0.1 + 0.025\sin(2\pi k/10000)$ とした．異常パターン A1 から A4 の確率変動のプロットを**図 2** に示す．図の横軸はオブジェクトの時系列順序 k を，縦軸はカテゴリ j の出現確率を表している．

まず，提案法の z-score が，基本パターンにおいて異常性を示さないことを有意確率に基づいて確認する．ここで，ナイーブな手法として，時系列順に見たときのオブジェクト W ごとのカテゴリ出現数を，バースト検知に関する手法として，Kleinberg の手法 [1] を比較手法として用いる．提案法における有意確率は提案 z-score に基づいたものであり，ナイーブな手法における有意確率は，成功確率 0.1，試行回数 W の二項分布の平均 $W \times 0.1$ と標準偏差 $\sqrt{W \times 0.1 \times (1-0.1)}$ に基づいたものである．今回，ナイーブな手法における出現数を数える範囲は $W = 1000$（全体で 10 ステップ），Kleinberg の手法のパラメータ設定は，異常パターンにおける最大出現確率 0.125 と基本パターンの出現確率 0.1 の比率を考慮して $s = 1.2$，$\gamma = 1.0$ とした．先ほどの基本パターン生成例にそれぞれの手法を適応したときの結果例と参考となる有意確率を**図 3** に示す．図 3 (a) より，ナイーブな手法において，基本パターンの各範囲ステップは参考となる有意確率を大きく外れていないことが分かる．この結果はランダムに生成しているため当然であるが，図 3 (b) より，提案法の z-score も，参考となる有意確率から大きく外れていないことが分かる．すなわち，基本パターンにおいて，提案法はナイーブな手法と同様に異常性を示さないといえる．また，図 3 (c) より，すべてのカテゴリにおいて，異常パターンを想定したバーストが検出されていないため，バースト検知の観点からも，異常性を示さないのは妥当であることが分かる．



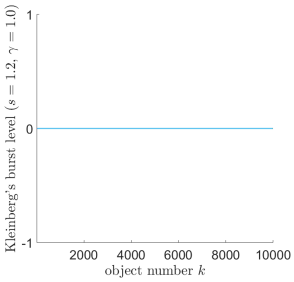
(a) 1,000 オブジェクトごとのカテゴリ出現数

(a) Frequencies of each category per 1,000 objects.



(b) 提案法

(b) Proposed method.



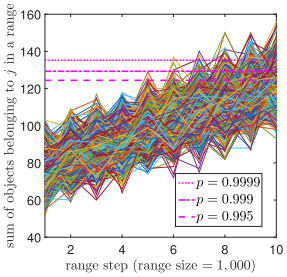
(c) Kleinberg の手法（パラメータ設定 $s = 1.2$，$\gamma = 1.0$）

(c) Kleinberg's method (parameter settings $s = 1.2$, $\gamma = 1.0$).

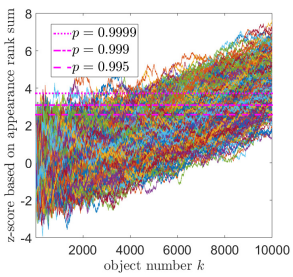
図 3　基本パターンの結果例

Fig. 3　Example results of the basic pattern.

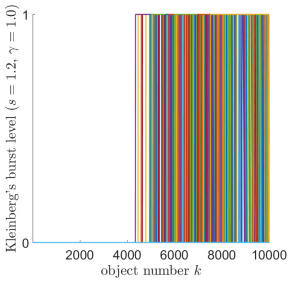
次に，基本パターンカテゴリ 1,000 個と異常パターンカテゴリ 1 個を有する 10,000 オブジェクトを，各異常パターンごとに 1,000 回ずつ独立に生成し，両手法に適応したときの結果と，参考となる有意確率を，A1 から A4 までそれぞれ**図 4**，**図 5**，**図 6**，**図 7** に示す．図 4 (a)，4 (b) より，出現確率が 0.075 から 0.125 に線形に増加する場合（A1）において，ナイーブな手法の最終ステップ値は参考となる有意確率を大きく超えることはほとんどないが，提案法の最終値 $z_{j,K}$ はほとんどのカテゴリが参考となる有意確率を大きく超えていることが分かる．また，図 4 (c) より，後半に出現したオブジェクトに対してバーストが検出されているため，バースト検知の観点からも，提案法の結果は妥当であることが分かる．図 5 (a)，5 (b) より，出現確率が 0.125 から 0.075 に線形に減少する場合（A2）において，ナイーブな手法の最終ステップ値は参考となる有意確率を大



(a) 1,000 オブジェクトごとのカテゴリ出現数
(a) Frequencies of each category per 1,000 objects.

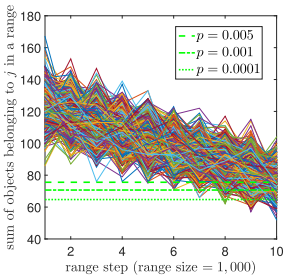


(b) 提案法
(b) Proposed method.

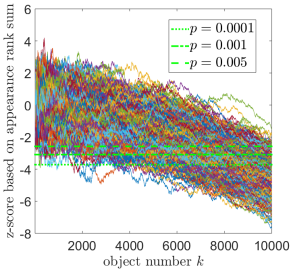


(c) Kleinberg の手法（パラメータ設定 $s = 1.2$，$\gamma = 1.0$）
(c) Kleinberg's method (parameter settings $s = 1.2$, $\gamma = 1.0$).

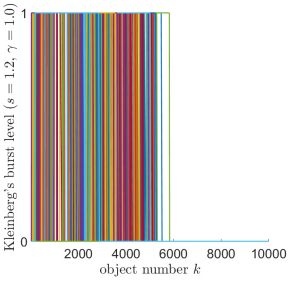
図 4　異常パターン A1 における結果
Fig. 4　Results of the anomaly pattern A1.



(a) 1,000 オブジェクトごとのカテゴリ出現数
(a) Frequencies of each category per 1,000 objects.

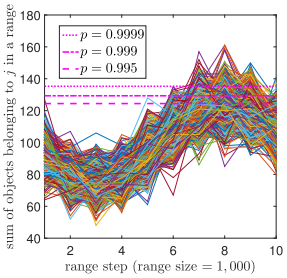


(b) 提案法
(b) Proposed method.

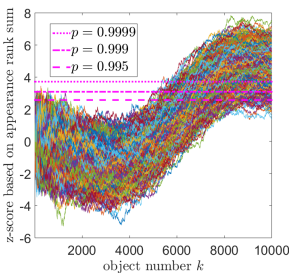


(c) Kleinberg の手法（パラメータ設定 $s = 1.2$，$\gamma = 1.0$）
(c) Kleinberg's method (parameter settings $s = 1.2$, $\gamma = 1.0$).

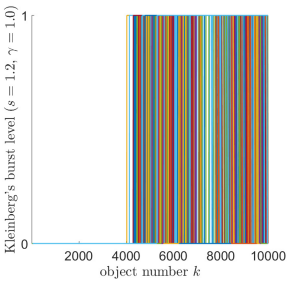
図 5　異常パターン A2 における結果
Fig. 5　Results of the anomaly pattern A2.



(a) 1,000 オブジェクトごとのカテゴリ出現数
(a) Frequencies of each category per 1,000 objects.

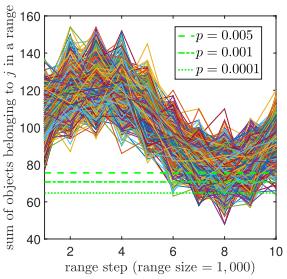


(b) 提案法
(b) Proposed method.

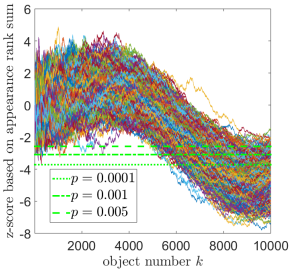


(c) Kleinberg の手法（パラメータ設定 $s = 1.2$，$\gamma = 1.0$）
(c) Kleinberg's method (parameter settings $s = 1.2$, $\gamma = 1.0$).

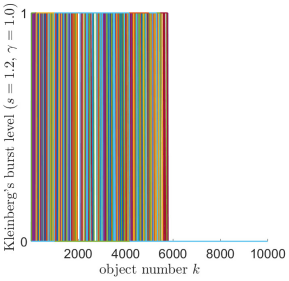
図 6　異常パターン A3 における結果
Fig. 6　Results of the anomaly pattern A3.



(a) 1,000 オブジェクトごとのカテゴリ出現数
(a) Frequencies of each category per 1,000 objects.



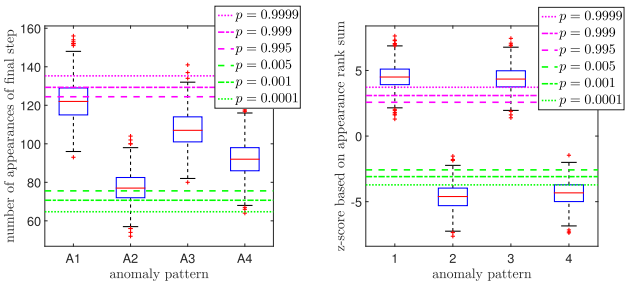
(b) 提案法
(b) Proposed method.



(c) Kleinberg の手法（パラメータ設定 $s = 1.2$，$\gamma = 1.0$）
(c) Kleinberg's method (parameter settings $s = 1.2$, $\gamma = 1.0$).

図 7　異常パターン A4 における結果
Fig. 7　Results of the anomaly pattern A4.

きく下回ることはほとんどないが，提案法の最終値 $z_{j,K}$ はほとんどのカテゴリが参考となる有意確率を大きく下回っていることが分かる．また，図 5 (c) より，前半に出現したオブジェクトに対してバーストが検出されているため，バースト検知の観点からも，提案法の結果は妥当であることが分かる．図 6 (a)，6 (b) より，出現確率がサインカー



(a) 1,000 オブジェクトごとのカテゴリ出現数（最終ステップ）

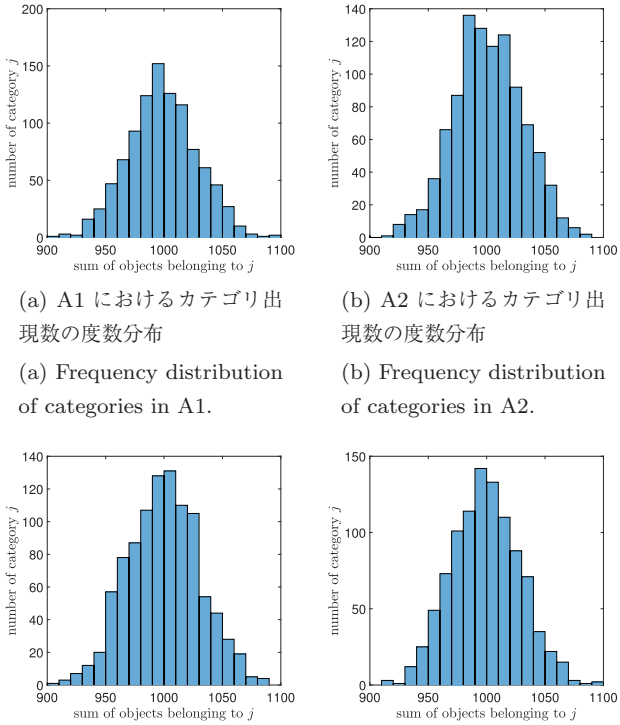
(a) Frequencies of each category per 1,000 objects (in the final step).

(b) 提案法（最終値 $z_{j,K}$）

(b) Proposed method (the final value $z_{j,K}$).

図 8　各異常パターンにおける最終値の分布

Fig. 8　Distributions of the final values in each anomaly pattern.



(a) A1 におけるカテゴリ出現数の度数分布

(a) Frequency distribution of categories in A1.

(b) A2 におけるカテゴリ出現数の度数分布

(b) Frequency distribution of categories in A2.

(c) A3 におけるカテゴリ出現数の度数分布

(c) Frequency distribution of categories in A3.

(d) A4 におけるカテゴリ出現数の度数分布

(d) Frequency distribution of categories in A4.

図 9　各異常パターンにおけるカテゴリ出現数の度数分布

Fig. 9　Frequency distribution of categories in each anomaly pattern.

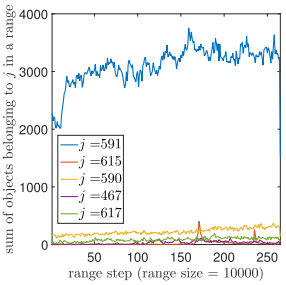
ブに従って減少したのち増加する場合（A3）において，ナイーブな手法の最終ステップ値と出現確率最大時の値は，ともに参考となる有意確率を大きく超えることはほとんどないが，提案法の z-score は出現確率最大時に急激に増加し，出現確率が 0.1 となる最終値においてもほとんどのカテゴリが参考となる有意確率を上回っていることが分かる．この異常パターン A3 は A1 と類似するパターンなので，A1 同様，参考となる有意確率を上回るのは妥当な結果といえる．また，図 6 (c) より，A1 同様後半に出現したオブジェクトに対してバーストが検出されているため，バースト検知の観点からも，提案法の結果は妥当であることが分かる．図 7 (a)，7 (b) より，出現確率がサインカーブに従って増加したのち減少する場合（A4）において，ナイーブな手法の最終ステップ値と出現確率最小時の値は，ともに参考となる有意確率を大きく下回ることはほとんどないが，提案法の z-score は出現確率最小時に急激に減少し，出現確率が 0.1 となる最終値においてもほとんどのカテゴリが参考となる有意確率を下回っていることが分かる．この異常パターン A4 は A2 と類似するパターンなので，A2 同様，参考となる有意確率を下回るのは妥当な結果といえる．また，図 7 (c) より，A2 同様前半に出現したオブジェクトに対してバーストが検出されているため，バースト検知の観点からも，提案法の結果は妥当であることが分かる．

これらの結果のまとめとして，生成された各異常パターンにおける最終値の分布を**図 8** に示す．図 8 より，提案法はナイーブ法に比べ，最終値において正確に異常パターンを検出できていることが分かる．さらに，**図 9** から見て取れるように，生成された各異常パターンのカテゴリ出現数の度数分布は基本パターン（図 1 (b)）とほぼ同様であるため，提案法はカテゴリ出現数の分布に差異がなくても異常性を定量的に表現できていることが分かる．実際，生成された各異常パターン A1，A2，A3，A4 のデータを用いたコルモゴロフ–スミルノフ検定（二項分布：試行回数 10,000，成功確率 0.1）において，有意確率はそれぞれ 0.0722，0.2978，0.3830，0.1923 となったため，一般的な有意水準 0.05 を基準とすると，異常パターンのカテゴリ出現数の分布は，基本パターンと同様に，仮定した二項分布に従っているといえる．以上のことから，提案法の z-score は，異常検知の側面から，トレンド分析の定量的評価法として有効であるといえる．また，今回の実験において，バースト検知の観点からも提案法の z-score による定量的評価が妥当であることが分かったが，あくまでバースト情報を値の変動によって表現することができているだけなので，提案法単体ではバーストの判定手法になっていないことに注意されたい．

4.　現実データによる実験

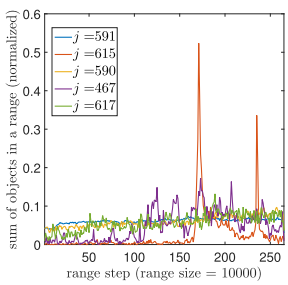
ここでは，現実データに提案法を適応し，ナイーブな手法と Kleinberg の手法 [1] との比較から，提案法の有用性を検証する．今回用いた現実データは，大規模レシピ投稿サイト “cookpad” [*1] から取得した，レシピ ID と各レシピのカテゴリ情報である．レシピ ID は昇順ソートし，先頭か

[*1] https://cookpad.com/



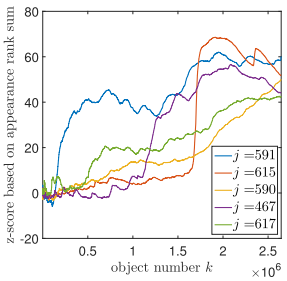
(a) 10,000 オブジェクトごとの出現数

(a) Frequencies of each category per 10,000 objects.



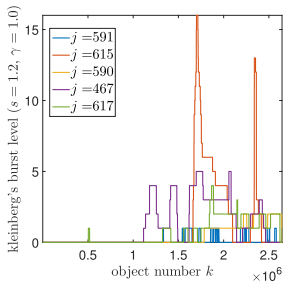
(b) 10,000 オブジェクトごとの出現数（正規化）

(b) Frequencies of each category per 10,000 objects (normalized).



(c) 提案法

(c) Proposed method.

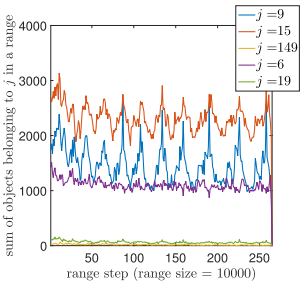


(d) Kleinberg の手法（パラメータ設定 $s = 1.2$, $\gamma = 1.0$）

(d) Kleinberg's method (parameter settings $s = 1.2, \gamma = 1.0$).

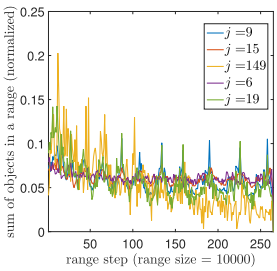
図 10　提案法における最終値上位 5 カテゴリの各手法の結果

Fig. 10　Results of each method of the top five categories in the final value of the proposed method.



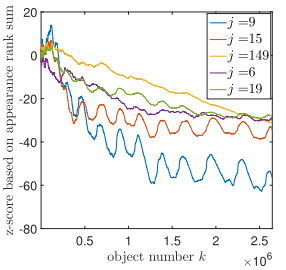
(a) 10,000 オブジェクトごとの出現数

(a) Frequencies of each category per 10,000 objects.



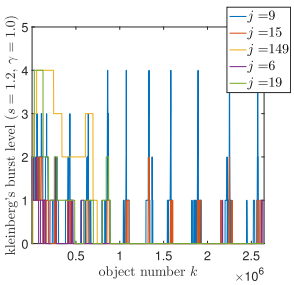
(b) 10,000 オブジェクトごとの出現数（正規化）

(b) Frequencies of each category per 10,000 objects (normalized).



(c) 提案法

(c) Proposed method.



(d) Kleinberg の手法（パラメータ設定 $s = 1.2$, $\gamma = 1.0$）

(d) Kleinberg's method (parameter settings $s = 1.2, \gamma = 1.0$).

図 11　提案法における最終値下位 5 カテゴリの各手法の結果

Fig. 11　Results of each method of the lower five categories in the final value of the proposed method.

ら $1, \cdots, k, \cdots, K$ としている．また，カテゴリは cookpad の分類に基づくものである．今回用いたデータセットは，レシピ数 $K = 2645326$，カテゴリ数 $J = 631$，カテゴリの総出現数 $I_K = 16996763$ である．

まず，提案法の z-score の最終値における上位 5 カテゴリ，すなわちカテゴリの勢力（出現確率）が増加している意味での異常性が最も高い 5 カテゴリについての検証結果を**図 10** に示す．なお，今回の上位 5 カテゴリは最上位から順に $j = 591$ の「かんたん」，$j = 615$ の「塩レモン」，$j = 590$ の「おすすめ」，$j = 467$ の「スムージー」，$j = 617$ の「離乳食」である．10,000 オブジェクトごとの出現数（図 10 (a)）は，「かんたん」とそれ以外ではスケールに差があるため，参考として正規化した出現数（図 10 (b)）を用いる．両図より，カテゴリの出現確率はどれも増加傾向を示していることは見て取れるが，スケールはもとより，長期的に徐々に増加しているものと，バーストとともに増加しているものが混ざっているため，カテゴリ間の勢力変化の比較を直接的に行うことは難しい．それに対し，提案法の結果（図 10 (c)）では，長期的な増加傾向も，バースト的な増加傾向も，すべて出現確率の増加として定量的に表現することができているため，カテゴリ間の勢力変化の比較を直接的に行うことが容易である．Kleinberg の手法による結果（図 10 (d)）を見ると，各カテゴリのバーストは後半に集中しているため，バースト検出の側面からも提案法の定量的評価の妥当性が証明されているといえる．また，Kleinberg の手法においても，パラメータを固定した状態では各カテゴリでバーストのスケールに差が出てしまうため，カテゴリ間の勢力変化の比較を直接的に行うことは難しい．

次に，提案法の z-score の最終値における下位 5 カテゴリ，すなわちカテゴリの勢力（出現確率）が減少している意味での異常性が最も高い 5 カテゴリについての検証結果を**図 11** に示す．なお，今回の下位 5 カテゴリは最下位から順に $j = 9$ の「お菓子」，$j = 15$ の「たまご」，$j = 149$ の「ベーグル」，$j = 6$ の「魚介のおかず」，$j = 19$ の「チーズケーキ」である．10,000 オブジェクトごとの出現数とその正規化値を図 11 (a)，11 (b) にそれぞれ示す．両図より，正規化した出現数においては，$j = 149$ の「ベーグル」と

$j = 19$ の「チーズケーキ」の減少が著しいことは分かるが，それら 2 カテゴリはスケールが小さいため，明確に勢力が変化したかどうかは分かりにくい．さらに，他の 3 カテゴリはスケールは大きいが減少傾向が小さいため，カテゴリ間の勢力変化の比較を直接的に行うことは難しい．それに対し，提案法の結果（図 11 (c)）では，正規化した出現数において減少が著しい 2 カテゴリも，スケールが大きく減少傾向が小さい他の 3 カテゴリも，勢力が明確に減少していることを表すことができているため，カテゴリ間の勢力変化の比較を直接的に行うことが容易である．Kleinberg の手法による結果（図 10 (d)）を見ると，すべてのカテゴリにおいてバーストが発生しているのは前半部分であるため，バースト検出の側面からも提案法の定量的評価の妥当性が証明されているといえる．しかし，Kleinberg の手法は，周期性が強い $j = 9$ の「お菓子」と $j = 15$ の「たまご」のバーストを後半部分でも検出し続けているため，バースト検出だけでは長期的な勢力減少に気づきにくいことが示唆される．

5. おわりに

時間方向の順序を用いた多群順位統計量によって，カテゴリの傾向変化や勢力関係を定量的に評価する手法を提案した．提案法の z-score は，カテゴリの異常な確率変動によって大きく変動することが可能であるため，人工データを用いた異常検知の側面から，トレンド分析の定量的評価法として有効であることを示した．また，現実データを用いた実験においては，スケールの大小やバースト・周期性の有無に左右されることなく，各カテゴリの長期的な傾向変化と，カテゴリ間の勢力関係の変化を定量的に示すことができた．

謝辞　本研究は，科研費基盤研究（c）15K00429 の支援を受けて行ったものである．

参考文献

[1] Kleinberg, J.: Bursty and hierarchical structure in streams, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD-2002*), pp.91–101 (2002).
[2] Swan, R. and Allan, J.: Automatic generation of overview timelines, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2000*), pp.49–56 (2000).
[3] Zhu, Y. and Shasha, D.: Efficient Elastic Burst Detection in Data Streams, *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD-2003*), pp.336–345 (2003).
[4] Sun, A., Zeng, D. and Chen, H.: Burst Detection from Multiple Data Streams: A Network-based Approach, *IEEE Trans. Systems, Man, & Cybernetics Society, Part C*, Vol.40, pp.258–267 (2010).
[5] Mann, H.B. and Whitney, D.R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Ann. Math. Statist.*, Vol.18, No.1, pp.50–60 (1947).
[6] Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., New York, NY, USA (1995).

山岸 祐己　（正会員）

静岡県立大学経営情報学部客員共同研究員．日本学術振興会特別研究員（PD）．2017 年静岡県立大学大学院経営情報イノベーション研究科博士後期課程修了．データマイニングの研究に従事．日本データベース学会会員．

岩﨑 清斗

静岡県工業技術研究所電子科研究員．2011 年法政大学工学部システム制御工学科卒業．センサネットワークの研究に従事．

斉藤 和巳　（正会員）

静岡県立大学経営情報学部教授．1985 年慶応義塾大学理工学部数理科学科卒業，1998 年東京大学博士（工学）．複雑ネットワークの研究に従事．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員．