Vol.2017-MPS-115 No.10
2017/9/26

多群出現順位統計量に基づく時系列データの変換

山岸 祐己[1,a)]　岩﨑 清斗[2]　斉藤 和巳[1,b)]

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．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捉えやすく，また，各カテゴリの傾向指標は他のカテゴリ全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．

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]　Kazumi Saito[1,b)]

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. 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.

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 と Allan [2] と同様に，回顧的 (retrospective) 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．

例えば，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

[1] 静岡県立大学
University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422–8526, Japan
[2] 静岡県工業技術研究所
Industrial Research Institute of Shizuoka Prefecture, 2078 Makigaya, Aoi-ku, Shizuoka 421-1221, Japan
[a)] yamagissy@gmail.com
[b)] k-saito@u-shizuoka-ken.ac.jp

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 と Allan 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共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捉えやすく，また，各カテゴリの指標は他のカテゴリ全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は，ナイーブな手法と共に，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，提案法の性能と特性を評価する．

2. 提案法

2.1 問題設定

時系列データの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．オブジェクト k によって出現したカテゴリ j の出現順位は基本的に $I_{k-1} + 1$ であるが，$t_k > 1$ の場合，すなわちオブジェクト k が複数のカテゴリを有する場合は平均順位を考えなければならないため，オブジェクト k における出現順位は $r_k = (I_{k-1} + 1 + I_k)/2$ 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 時間方向多群順位統計量

Mann-Whitney の二群順位統計量 [5] を多群に拡張し，時間方向に適用する方法について述べる．Mann-Whitney の二群順位統計量に従い，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．

$$z_{j,k} = \frac{u_{j,k} - \mu_{j,k}}{\sigma_{j,k}}. \tag{1}$$

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．

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

$$\mu_{j,k} = \frac{I_{j,k}(I_k - I_{j,k})}{2}, \tag{3}$$

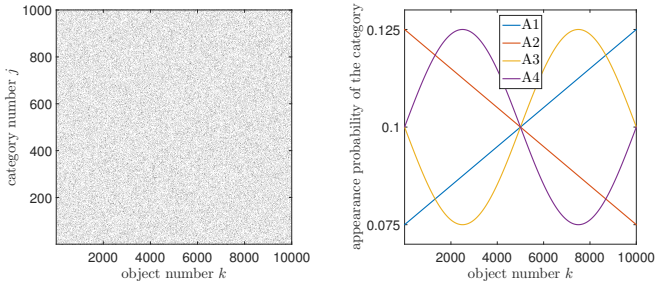
$$\sigma^2_{j,k} = \frac{I_{j,k}(I_k - I_{j,k})}{12}\left((I_k+1) - \sum_{i=1}^{k}\frac{t_i^3 - t_i}{I_k(I_k-1)}\right). \tag{4}$$

ただし，各オブジェクトが複数のカテゴリを有し得ないケースでは，式 (4) の t_k を含む項の計算は不要である．よって，式 (1) 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．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．

3. 人工データによる実験

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 オブジェクトを人工的に生成し，提案法が単純な異常パターンを検出できるかどうかを検証する．ここで，オブジェクト k でのカテゴリーの出現確率を α_k とし，基本パターンは固定確率 $\alpha_k = 0.1$ とする．これに対し，異常パターン A1 から A4 はそれぞれ $\alpha_k = 0.075 + 0.05(k/10000)$，$\alpha_k = 0.125 - 0.05(k/10000)$，$\alpha_k = 0.1 + 0.025\sin(-2\pi k/10000)$，$\alpha_k = 0.1 + 0.025\sin(2\pi k/10000)$ のように設定した．基本

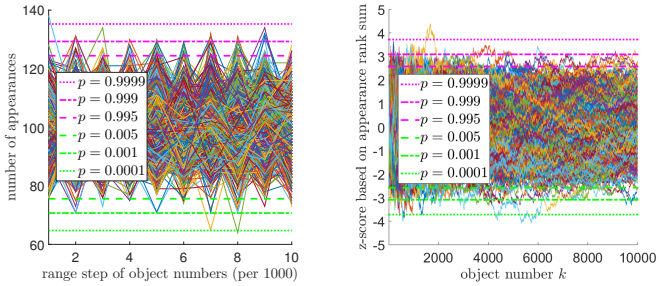
パターンカテゴリ 1,000 個を有する 10,000 オブジェクトを生成した例を図 1(a) に，異常パターン A1 から A4 の確率変動のプロットを図 1(b) にそれぞれ示す．



(a) 基本パターンカテゴリの生成例（黒点 $q_{j,k} = 1$）　(b) 各異常パターンの確率変動

図 1　人工データにおける基本パターンと異常パターン

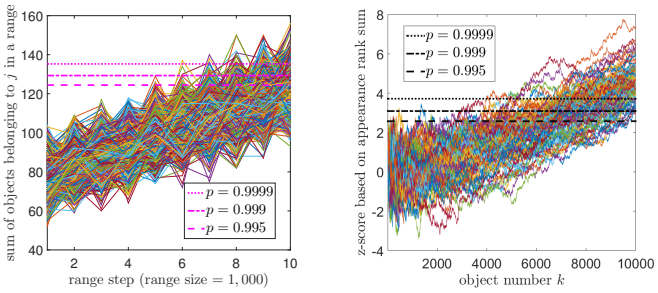
まず，提案手法が，基本パターンにおいて異常性を示さないことを有意確率に基いて確認する．ここで，比較対象となるナイーブな手法として，R オブジェクトごとのカテゴリ出現数を用いる．つまり，提案手法における有意確率は提案 z-score に基づいたものであり，ナイーブな手法における有意確率は平均 $R \times 0.1$，標準偏差 $\sqrt{R \times 0.1 \times (1-0.1)}$ に基づいたものである．基本パターンにそれぞれの手法を適応したときの有意確率を図 2 に示す．図より，両手法の値とも参考となる有意確率から大きく外れないことがわかる．これらのまとめとして，最終値の分布を図 7 に示す．以上のことから，提案法は，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，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．



(a) R オブジェクトごとのカテゴリ出現数　(b) 提案手法

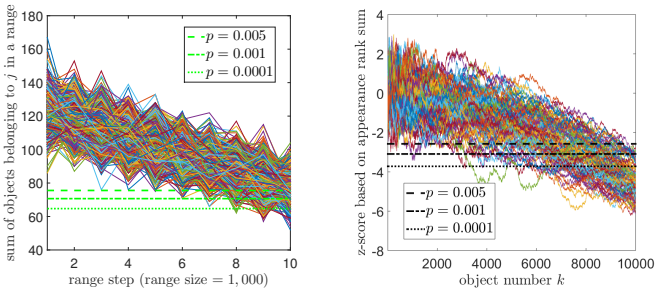
図 2　基本パターンにおける有意確率

次に，異常パターンにそれぞれの手法を適応したときの有意確率を A1 から A4 までそれぞれ図 3 から図 6 に示す．なお，ここでの結果は基本パターンと異常パターンを 100 回独立に生成し，その都度異常パターンにのみ両手法を適応したものを示している．図より，ナイーブな手法の値では，全ての異常パターンにおいて参考となる有意確率を大幅に超えることは殆ど無いことがわかる．それに対し提案法は，確率変動によって値が大きく動き，最終値では参考となる有意確率を大幅に超えていることがわかる．



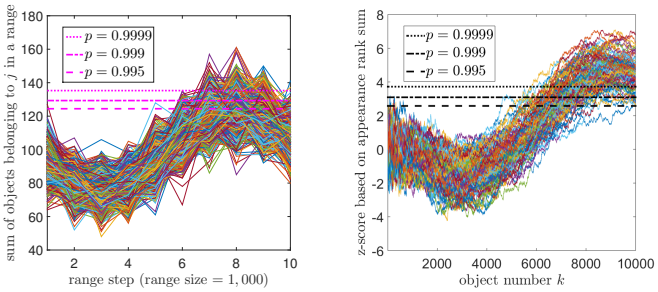
(a) R オブジェクトごとのカテゴリ出現数　(b) 提案手法

図 3　異常パターン A1 における有意確率



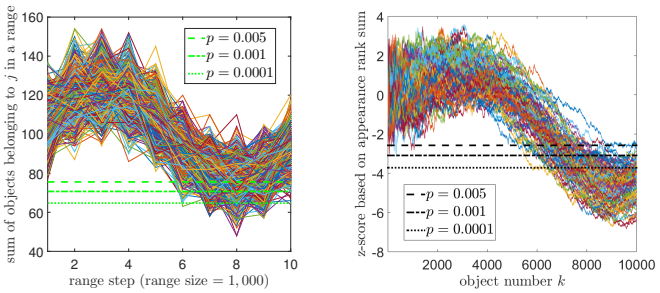
(a) R オブジェクトごとのカテゴリ出現数　(b) 提案手法

図 4　異常パターン A2 における有意確率



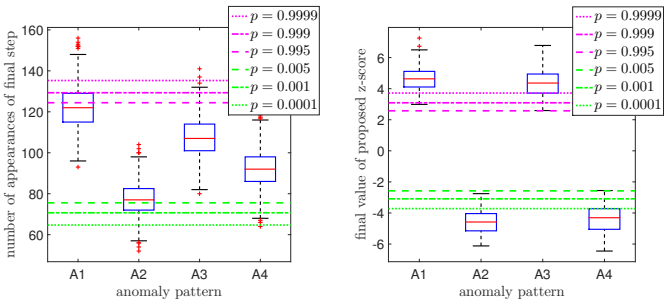
(a) R オブジェクトごとのカテゴリ出現数　(b) 提案手法

図 5　異常パターン A3 における有意確率



(a) R オブジェクトごとのカテゴリ出現数　　(b) 提案手法

図 6　異常パターン A4 における有意確率



(a) R オブジェクトごとのカテゴリ出現数（最終ステップ）　　(b) 提案手法（最終値）

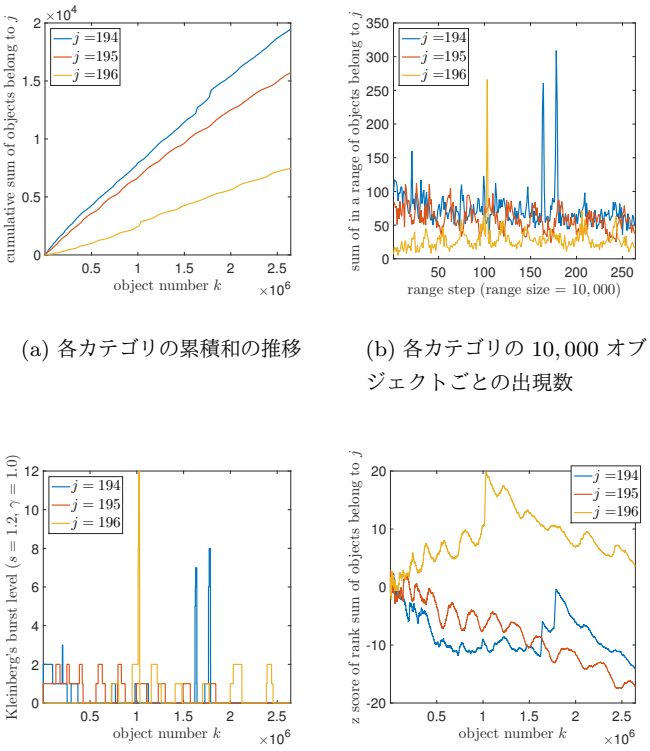
図 7　最終値を用いた異常検知の精度

4. 現実データによる実験

現実データとして，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を用いる．レシピ ID は昇順ソートし，先頭から $1, \cdots, k, \cdots, K$ 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 $K = 2,645,326$，カテゴリ数 $J = 631$，カテゴリの総出現数 $I_K = 16,996,763$ である．

実験結果例として，パスタソースのサブカテゴリである “トマトソース（$j = 194$）”，“ホワイトソース（$j = 195$）”，“ジェノベーゼソース（$j = 196$）” の 3 カテゴリの比較について述べる．まず，各カテゴリの累積和の推移（図 8(a)）を見ると，3 カテゴリの傾向変化や勢力関係の変化は殆どないように思える．次に，各カテゴリの 10,000 オブジェクトごとの出現数（図 8(b)）を見ると，急激に出現数が増えた時期や，ノイズが混じった周期性の存在が分かるようになっている．しかし，この図からも 3 カテゴリの傾向変化や勢力関係の変化は明確には分からない．図 8(b) で見られるような出現数の急激な増加や周期性などは，図 8(c) の Kleinberg の手法による結果で簡潔に表されている．短期間の傾向変化について知りたいのであれば，図 8(c) の結果だけでも十分有用であるが，長期的な傾向変化や，3 カテゴリの勢力関係の変化について知りたい場合は情報として不十分であるように思える．これらの結果に対し，提案法による結果（図 8(d)）では，周期性，短期的・長期的な傾向変化，カテゴリ間の勢力関係といった情報が明確に示されている．

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



(a) 各カテゴリの累積和の推移　　(b) 各カテゴリの 10,000 オブジェクトごとの出現数

(c) Kleinberg の手法による結果（パラメータ設定 $s = 1.2$, $\gamma = 1.0$）　　(d) 提案法による結果

図 8　各カテゴリの統計情報と手法の比較

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，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人工データによる実験で示した．また，現実データを用いた実験においては，基本的な統計情報や時系列分析の代表的な手法から得られる情報に加えて，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の情報を示すことができた．

謝辞

本研究は，JSPS 特別研究員奨励費 16J11909 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, *Proceedings of the 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, *Proceedings of the 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, *Proceedings of the 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 Transactions on 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).