7M-01

順序カテゴリカルデータにおける出現頻度の異常検出

佐野 歓基† 佐藤 道大† 池田 朋樹† 山岸 祐己† 斉藤 和巳‡
† 静岡理工科大学 情報学部 ‡ 神奈川大学 理学部

1 はじめに

本論文では，順序カテゴリカルデータの時系列的変化の検出を行い，その異常性を定量的に評価する手法を提案する．一般に，順序尺度の性質を持つ離散値で表現された時系列データを分析または可視化するためには，事前に移動平均値などに変換する必要があるが，提案手法では，そのような離散値をカテゴリカルデータとして扱うことによって，事前にパラメータを一切設定することなく分析または可視化するとともに，各離散値の出現頻度の時系列的変化を定量的に評価する．評価実験では，レビュー時系列データを用いて提案手法の有効性を検証する．また，実験において，提案手法は高速計算機を使わずとも実用的な時間で動作することも示す．

2 多群順序統計量

データセットにおけるタイムステップ集合と，それらが有するカテゴリ集合をそれぞれ $\mathcal{N}$ と $\mathcal{J}$ とする．ここで，それぞれの要素数は $N = |\mathcal{N}|$ 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，$\mathcal{N} = \{1, \cdots, n, \cdots, N\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ である．なお，オブジェクト n は最古のものが 1，最新のものが N となるよう，出現順に並んでいるものとする．このとき，タイムステップ n 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 N 列の行列を Q ($q_{j,n} \in \{0, 1\}$) とすると，オブジェクト n が有するカテゴリ数は $t_n = \sum_{i=1}^{J} q_{i,n}$，タイムステップ n までの全カテゴリの総出現数は $I_n = \sum_{i=1}^{J} I_{i,n}$ 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 n が唯一のカテゴリのみ有する $t_n = 1$ の場合では，オブジェクト n に付随して出現したカテゴリ j の出現順位は $r_n = I_{n-1} + 1$ であるが，複数のカテゴリを有する $t_n > 1$ 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 $r_n = I_{n-1} + (1 + t_n)/2$ となる．ここでの目的は，タイムステップ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）タイムステップ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．

Mann-Whitney の二群順位統計量 [1] を多群に拡張し，カテゴリ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タイムステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 1\}$ と，このカテゴリに属さないタイムステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 0\}$ 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．

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

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

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

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得ないケースでは，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) [2] 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-score z_j により，最新オブジェクト N 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 n 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 n に対応した次式により，タイムステップ n までのカテゴリ j に対し z-score $z_{j,n}$ を求めることができる．

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

ここで，統計量 $u_{j,n}$, 出現順位の平均 $\mu_{j,n}$，および，そ

Anomaly Detection of Frequency of Appearance in Ordinal Categorical Data
†Kanki SANO †Michihiro SATO †Tomoki IKEDA †Yuki YAMAGISHI ‡Kazumi SAITO
†Shizuoka Institute of Science and Technology
‡Kanagawa University

の分散 $\sigma^2_{j,n}$ は次のように計算される．

$$u_{j,n} = \sum_{i=1}^{n} nq_{j,i} - \frac{I_{j,n}(I_{j,n}+1)}{2}, \quad (6)$$

$$\mu_{j,n} = \frac{I_{j,n}(n-I_{j,n})}{2}, \quad (7)$$

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

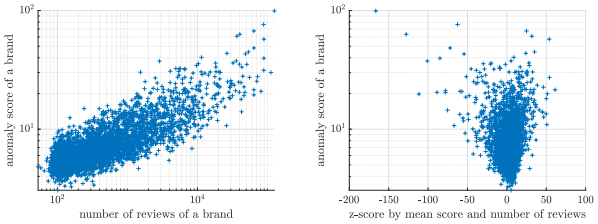
先程と同様，各オブジェクトが複数のカテゴリを有し得ないケースでは，式 (8) の t_i を含む項，すなわち平均順位を扱うための補正値の計算は不要である．

以上より，式 (5) で求まる z-score $z_{j,n}$ 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 $z_{j,n}$ が正の方向に大きければ大きいほど，タイムステップ n 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，$z_{j,n}$ 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，式 (5) で求まる z-score $z_{j,n}$ の計算量は全てのオブジェクトと全てのカテゴリについて算出した場合でも $O(NJ)$ 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．また，異常検知を目的として有意水準を設定すれば，$z_{j,n}$ から求まる有意確率を使った仮説検定が可能である．

3 評価実験とまとめ

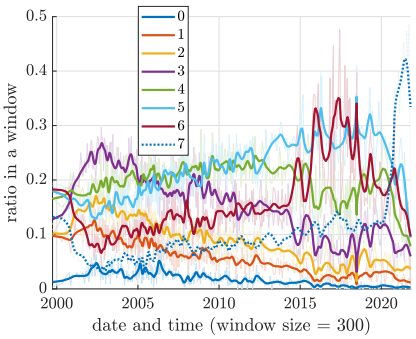
コスメレビューサイトの @cosme [*] における，被レビュー数が 100 以上の 14526 アイテムのレビューのうち，評点とユーザの紐づけがされている 6512843 レビューを対象とし，アイテムに紐づけられている 3527 ブランドごとで評価実験を行った．各レビューの評点 0 点から 7 点をカテゴリ（$J=8$）とし，全ブランドにおける多群順序統計量による各評点の z-score を求めたところ，MATLAB R2021b (Intel(R) Core(TM) i7-10710U CPU @ 1.10GHz) での実行で 10 回の平均計算時間は 29.91 秒だった．

今回は，多群順序統計量による各評点の z-score $z_{j,n}$ の最大値と最小値のそれぞれの絶対値の和 $|\max_{j\in\mathcal{J},n\in\mathcal{N}} z_{j,n}| + |\min_{j\in\mathcal{J},n\in\mathcal{N}} z_{j,n}|$ を各ブランドの異常値とした．図 1a より，各ブランドの被レビュー数が多ければ多いほど，提案異常値が高くなる傾向が見て取れる．また，図 1b より，平均評点と被レビュー数による z-score と比較して，提案異常値は平均評点の z-score が 0 となるようなブランドに対しても高い異常値を示していることがわかる．図 2a は被レビュー数が最多のブランドの各評点の割合を可視化したもので，図 2b は同ブランドの多群順序統計量による各評点の z-score $z_{j,n}$ を可視化したものである．両図より，多群順序統計量は，評点の長期的な出現頻度の変化を定量的に評価できていることがわかる．

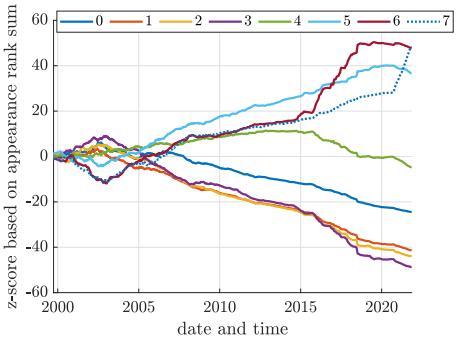


(a) 各ブランドの被レビュー数と提案異常値　(b) 各ブランドの平均評点の z-score と提案異常値の比較

図 1: 多群順序統計量を用いた提案異常値の評価



(a) ウィンドウサイズ 300 における各評点の割合



(b) 多群順序統計量による各評点の z-score $z_{j,n}$

図 2: 被レビュー数最多のブランドにおける実験結果

謝辞 本研究は科研費基盤研究 (C) 18K11441 の支援を受けて行ったものである．

参考文献

[1] H. B. Mann and D. R. Whitney. 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, 03 1947.

[2] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[*] https://www.cosme.net/