D-021

カテゴリの出現順序に基づくレビューデータの異常検出

増田 大純 † 佐野 歓基 † 山岸 祐己 † 和泉 舞 ‡ 高林 貴仁 ‡
† 静岡理工科大学 情報学部 ‡ 株式会社良品計画

1 はじめに

レビューデータは，評点をはじめとして，投稿したユーザの性別や年代といった属性など，様々なカテゴリを仮定することができる．本研究では，それらのカテゴリの分布にかかわらず適応可能な，ノンパラメトリック検定に基づく異常検出手法を提案する．提案手法は Mann-Whitney の U 検定を多群に拡張し，時系列データに適応するためにオンライン処理ができるよう変形したものである．現実のレビューデータを用いた実験では，提案手法による異常値を z-score と p 値で可視化することによって，各カテゴリの出現頻度の変化とその異常性が説明できることを示す．

2 提案手法

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 までのカテゴリ j の出現数は $I_{j,n} = \sum_{i=1}^{n} q_{j,i}$ 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 U 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 U 統計量に従い，次式により，カテゴリ j に対し U 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．

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

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

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

すなわち，u_j は順位和に基づくカテゴリ j の U 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 U 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，その分散 $\sigma_{j,n}^2$ は次のように計算される．

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

Anomaly Detection in Review Data Based on the Order of Appearance of Categories
†Taijun MASUDA †Kanki SANO †Yuki YAMAGISHI ‡Mai IZUMI ‡Takahito TAKABAYASHI
†Shizuoka Institute of Science and Technology
‡Ryohin Keikaku Co., Ltd.

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

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

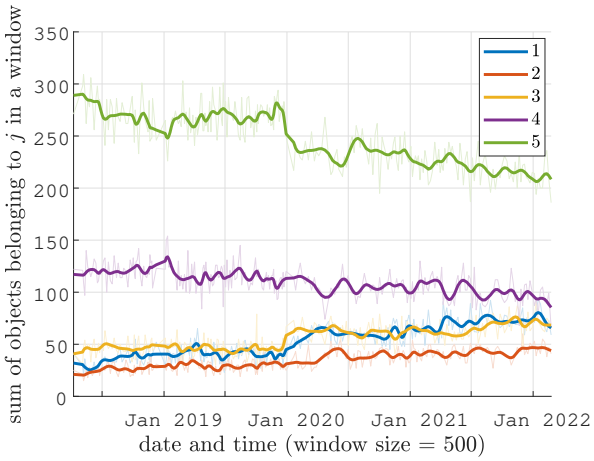
$$\sigma^2_{j,n} = \frac{I_{j,n}(I_n - I_{j,n})}{12}\left((I_n + 1) - \sum_{i=1}^{n} \frac{t_i^3 - t_i}{I_n(I_n - 1)}\right). \tag{8}$$

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．

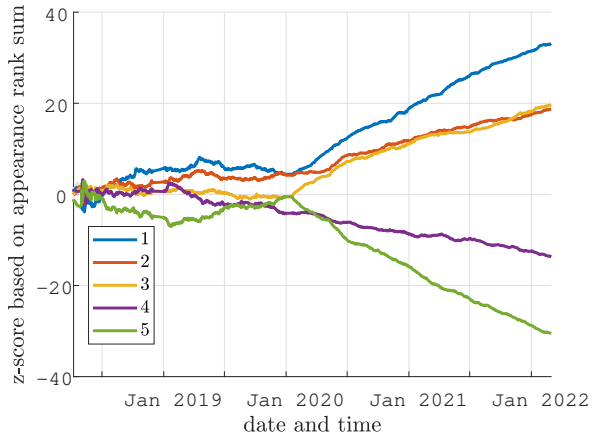
以上より，式 (5) で求まる z-score $z_{j,n}$ により，オブジェクト n までの各カテゴリ j が，出現順位の値が大きい（新しい），または逆に小さい（古い）オブジェクトを有意に多く含むかを定量的に評価することができる．すなわち，この $z_{j,n}$ が正の方向に大きければ大きいほど，タイムステップ n の直近での出現が有意に多いということであり，カテゴリ j の勢力が伸びていることになる．逆に，$z_{j,n}$ が負の方向に大きいということは，過去に比べて勢力が衰えていることになる．また，式 (5) で求まる z-score $z_{j,n}$ の計算量は全てのオブジ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．この多群 U 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．

3　評価実験とまとめ

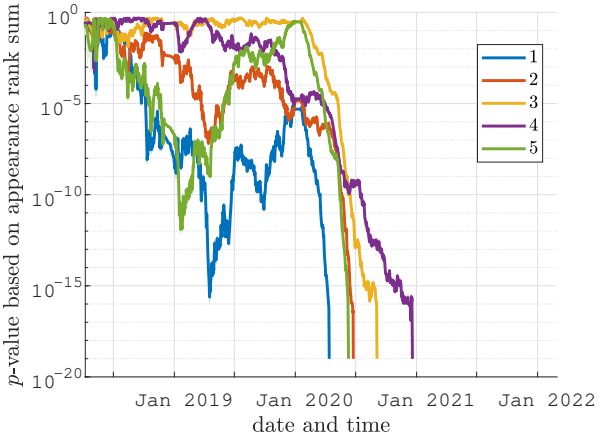
実験では無印良品 (https://www.muji.com/jp/ja/store) のレビューデータセットを用いる．データセットに含まれるレビュー数は 136,916 であり，期間は 2018 年 4 月から 2022 年 2 月である．今回は 1 点から 5 点のレビュー点数をカテゴリとした ($J = 5$)．図 1 より，単純な出現数の増減（図 1a）に対し，多群 U 統計量の z-score $z_{j,n}$（図 1b）は，各点数 j の長期的な出現傾向の変化を示すことができていることが見て取れる．特に，2020 年を過ぎたあたりからそれらの変化が激しくなり，そこから変化が継続していることも分かる．さらに，図 1c のように $z_{j,n}$ を p 値（ここでは両側検定）に変換すれば，各点数 j の出現傾向の変化の異常性が直感的に分かりやすくなり，どの時期にどの順番でアンダーフローしたかなども分かるようになる．



(a) ウィンドウサイズ 500 におけるカテゴリ j の出現数



(b) 多群 U 統計量によるカテゴリ j の z-score



(c) 多群 U 統計量によるカテゴリ j の p 値

図 1: 評価実験結果

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.

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.