DEIM Forum 2022 F21-5

レビューサイトにおける評点分布の時系列変化に基づいた異常ユーザの検出

鈴木麻由里† 古山 陽菜† 山岸 祐己†

† 静岡理工科大学情報学部コンピュータシステム学科 〒 437–0032 静岡県袋井市豊沢 2200–2
E-mail: †{1818087.sm,1818132.fh,yamagishi.yuki}@sist.ac.jp

あらまし 本論文では，レビューの評点データを順序カテゴリカルデータとして扱い，ユーザが投稿した評点データおよびアイテムに投稿された評点データそれぞれに対して評点分布の時系列異常検知を行う．さらに，その検出結果における各ユーザの異常性を定量的に評価する手法を提案する．一般に，ユーザのレビュー評点行動の異常性を定量的に評価する方法として，投稿レビュー数を考慮した評点平均値の z-score などが考えられるが，このような手法では極端に高い（低い）評点を頻繁に投稿するようなユーザの検出に限られる可能性が高い．それに対し，提案手法は，ユーザが投稿した評点の分布およびアイテムに投稿された評点の分布を考慮することによって，ユーザの多様な異常性を定量的に評価することを可能としている．

キーワード 多項分布モデル，順序統計量，レジームスイッチング，異常検出

1 はじめに

レビューサイトでは，ユーザがアイテムに対して評点をつける行動が日々行われているが，この評点行動が他のユーザと比較したうえでの相対的な評価として行われているか，またはユーザ独自の基準による評価として行われているかなどは不明瞭である．例えば，あるユーザが投稿したレビュー評点 $X_1,\cdots,X_n$ の平均 $\bar{X}=\sum_{i=1}^{n}X_i/n$ とレビュー投稿数 n を用いれば，ユーザ全体から求めた平均評点の期待値 $E[\bar{X}]$ と標準偏差 $\sigma(X)$ によって $(\bar{X}-E[\bar{X}])/(\sigma(X)/\sqrt{n})$ のように標準化することができ，評点行動の異常性の定量的評価として考えることができる．しかし，この値は，極端に高い（低い）評点を多く投稿するようなユーザのみが極めて大きく（小さく）なることが予想できるため，例えば，各アイテムにおいて少数派の評点ばかり投稿するようなユーザや，ある時期から投稿している評点の分布が大きく変化するユーザなどの検出も考慮すると，それぞれ別のアプローチが必要となる．よって，本論文では，回顧的 (retrospective) な枠組みによる時系列データからの構造抽出 [1], [2] と同様の考え方で，ユーザの評点行動の異常性を定量的に評価するための分析手法を提案する．

2 多項分布レジームスイッチング

時系列データを $\mathcal{D}=\{(s_1,t_1),\cdots,(s_N,t_N)\}$ とする．ここで，s_n と t_n は，J カテゴリの状態と n 番目の観測時刻をそれぞれ表す．$|\mathcal{D}|=N$ を観測数とすると，$t_1\leqq\cdots\leqq t_n\leqq\cdots\leqq t_N$ となる．n はタイムステップとし，$\mathcal{N}=\{1,2,\cdots,N\}$ をタイムステップ集合とする．また，k 番目のレジームの開始時刻を $T_k\in\mathcal{N}$，$\mathcal{T}_K=\{T_0,\cdots,T_k,\cdots,T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上 $T_0=1$，$T_{K+1}=N+1$ とする．すなわち，$T_1,\cdots,T_K$ は推定される個々のスイッチングタイムステップであり，$T_k<T_{k+1}$ を満たすとする．そして，$\mathcal{N}_k$ を k 番目のレジーム内のタイムステップ集合とし，各 $k\in\{0,\cdots,K\}$ に対して $\mathcal{N}_k=\{n\in\mathcal{N};T_k\leqq n<T_{K+1}\}$ のように定義する．なお，$\mathcal{N}=\mathcal{N}_0\cup\cdots\cup\mathcal{N}_K$ である．

いま，各レジームの状態分布が J カテゴリの多項分布に従うと仮定する，$\boldsymbol{p}_k$ を k 番目のレジームにおける多項分布の確率ベクトルとし，$\mathcal{P}_K$ はそれら確率ベクトルの集合，つまり $\mathcal{P}_K=\{\boldsymbol{p}_0,\cdots,\boldsymbol{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$L(\mathcal{D};\mathcal{P}_K,\mathcal{T}_K)=\sum_{k=0}^{K}\sum_{n\in\mathcal{N}_k}\sum_{j=1}^{J}s_{n,j}\log p_{k,j}. \tag{1}$$

ここで，$s_{n,j}$ は $s_n\in\{1,\cdots,J\}$ を

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

のように変換したダミー変数である．各レジーム $k=0,\cdots,K$ と各状態 $j=1,\cdots,J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j}=\sum_{n\in\mathcal{N}_k}s_{n,j}/|\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると以下の式が導ける．

$$L(\mathcal{D};\hat{\mathcal{P}}_K,\mathcal{T}_K)=\sum_{k=0}^{K}\sum_{n\in\mathcal{N}_k}\sum_{j=1}^{J}s_{n,j}\log\hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

しかし，式 (3) だけでは $\mathcal{T}_K$ の導入によってどれだけ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，レジームスイッチングを考慮しないときの尤度からの改善度合いを評価することは重要であるため，尤度比最大化問題として目的関数を構築し直す．もし，レジームスイッチングのような変化が存在しない，すなわち $\mathcal{T}_0=\emptyset$ と仮定する

と，式 (3) は

$$L(\mathcal{D};\hat{\mathcal{P}}_0,\mathcal{T}_0)=\sum_{n\in\mathcal{N}}\sum_{j=1}^{J}s_{n,j}\log\hat{p}_{0,j}, \tag{4}$$

となる．ここで，$\hat{p}_{0,j}=\sum_{n\in\mathcal{N}}s_{n,j}/N$ である．よって，K 個のスイッチングを持つ場合と，スイッチングを持たない場合の対数尤度比は

$$LR(\mathcal{T}_K)=L(\mathcal{D};\hat{\mathcal{P}}_K,\mathcal{T}_K)-L(\mathcal{D};\hat{\mathcal{P}}_0,\mathcal{T}_0). \tag{5}$$

のように与えられる．最終的に，この問題は上記の $LR(\mathcal{T}_K)$ を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

式 (5) 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，任意の K について解くために，貪欲法と局所探索法を組み合わせた方法 [3] を用いる．なお，本実験では貪欲法アルゴリズムの終了条件として最小記述長原理 (MDL) [4] を採用する．

3　多群順序統計量

多項分布レジームスイッチングの問題設定と同様に，時系列データの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．つまり，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．すなわち，$\mathcal{N}=\{1,\cdots,n,\cdots,N\}$ 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 $d_n=\sum_{i=1}^{J}q_{i,n}$，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 $d_n=1$ 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 $d_n>1$ 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 $r_n=I_{n-1}+(1+d_n)/2$ 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 [5] 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．

$$z_j\ =\ \frac{u_j-\mu_j}{\sigma_j}. \tag{6}$$

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．

$$u_j\ =\ \sum_{i=1}^{N}r_iq_{j,i}-\frac{I_{j,N}(I_{j,N}+1)}{2}, \tag{7}$$

$$\mu_j\ =\ \frac{I_{j,N}(I_K-I_{j,N})}{2}, \tag{8}$$

$$\sigma_j^2\ =\ \frac{I_{j,N}(I_N-I_{j,N})}{12}\left((I_N+1)-\sum_{i=1}^{N}\frac{d_i^3-d_i}{I_N(I_N-1)}\right). \tag{9}$$

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 (9) の d_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 (6) 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．

$$z_{j,n}\ =\ \frac{u_{j,n}-\mu_{j,n}}{\sigma_{j,n}}. \tag{10}$$

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．

$$u_{j,n}\ =\ \sum_{i=1}^{n}nq_{j,i}-\frac{I_{j,n}(I_{j,n}+1)}{2}, \tag{11}$$

$$\mu_{j,n}\ =\ \frac{I_{j,n}(n-I_{j,n})}{2}, \tag{12}$$

$$\sigma_{j,n}^2\ =\ \frac{I_{j,n}(I_n-I_{j,n})}{12}\left((I_n+1)-\sum_{i=1}^{n}\frac{d_i^3-d_i}{I_n(I_n-1)}\right). \tag{13}$$

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 (13) の d_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 (10) 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 (10) 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．

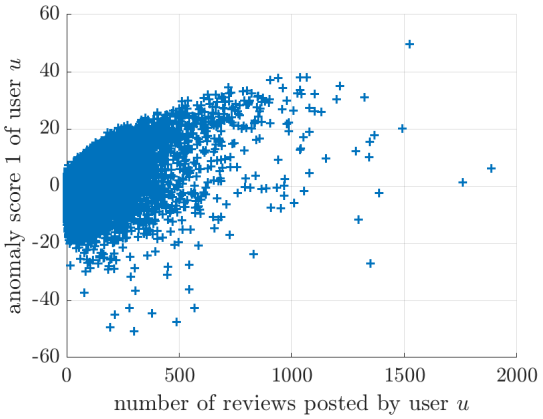
4　評価実験とまとめ

コスメレビューサイトの @cosme [1] 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 レビューを評価実験の対象とした．

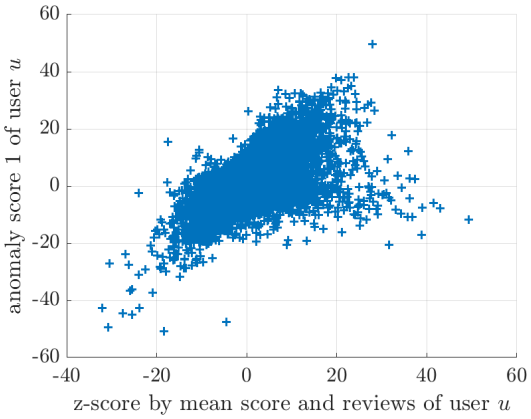
まず，各レビューの評点 0 点から 7 点をカテゴリ $J=8$ として，各アイテムで提案レジームスイッチングを行い，$\mathcal{T}_K$ の近似解 $\hat{\mathcal{T}}_K$ を求めた．ここで，アイテム $i \in \mathcal{I} = \{1,2,\cdots,I\}$, $I = 14526$ の $\mathcal{D}$, $\hat{\mathcal{T}}_K$ を，それぞれ $\mathcal{D}_i$, $\hat{\mathcal{T}}_{i,K}$ とし，そのときの $\hat{\mathcal{P}}_K$ を $\hat{\mathcal{P}}_{i,K} = \{\hat{\boldsymbol{p}}_{i,0},\cdots,\hat{\boldsymbol{p}}_{i,K}\}$ とする．このとき，ユーザ $u \in \mathcal{U} = \{1,2,\cdots,U\}$, $U = 701854$ がアイテム i のレジーム k で投稿したレビューの評点カテゴリを $s_{u,i,k} \in \{1,\cdots,J\}$ とし，$s_{n,j}$ と同様に $s_{u,i,k,j}$ のようなダミー変数とすれば，ユーザ u が投稿したレビュー評点 $s_{u,i,k,j}$ の対数尤度は $\sum_{j=1}^{J} s_{u,i,k,j} \log \hat{p}_{i,k,j}$ となる．この対数尤度の平均と投稿レビュー数による z-score をユーザ u の提案異常値 1 とした．同様に，ユーザ u の多群順序統計量による各評点の z-score $z_{j,n}$ を $z_{u,j,n}$ として求め，その最大値と最小値のそれぞれの絶対値の和 $|\max_{j\in\mathcal{J},n\in\mathcal{N}} z_{u,j,n}| + |\min_{j\in\mathcal{J},n\in\mathcal{N}} z_{u,j,n}|$ をユーザ u の提案異常値 2 とした．

図 1a より，各ユーザの投稿レビュー数が多ければ多いほど，提案異常値 1 の正値は増加する傾向が見て取れるが，提案異常値 1 の負値ではそのような傾向は無いように見える．また，図 1b より，各ユーザの平均評点と投稿レビュー数による z-score と提案異常値 1 は，概ね正の相関を示しているが，平均評点の z-score が高い一部のユーザは，提案異常値 1 では負値となっていることがわかる．両図より，提案異常値 1 の正値に関しては，おおむね平均評点の z-score と同等となることが予想されるが，提案異常値 1 の負値に関しては，平均評点の z-score において正値のユーザも多く含まれることが予想される．この符号が反転しているユーザについては，平均評点が低いアイテムに対し，常に高い評点を投稿するという評点行動が影響していると考えられる．

図 2a より，提案異常値 2 も，各ユーザの投稿レビュー数が多ければ多いほど増加する傾向があるが，投稿レビュー数が少ないユーザでも提案異常値 2 が高くなり得ることが見て取れる．また，図 2b より，提案異常値 2 は，各ユーザの平均評点と投稿レビュー数による z-score が高いユーザだけではなく，低いユーザに対してもある程度高い値を付与していることがわかる．両図より，特に平均評点の z-score が正値のユーザほど提案異常値 2 が高くなる傾向があるため，投稿する評点の分布が時間とともに変化しているユーザは，比較的高い評点を投稿する傾向があると考えられる．



(a) 各ユーザの投稿レビュー数と提案異常値 1



(b) 各ユーザの平均評点と投稿レビュー数による z-score と提案異常値 1 の比較

図 1: 提案異常値 1 の評価

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．

1：https://www.cosme.net/

文　献

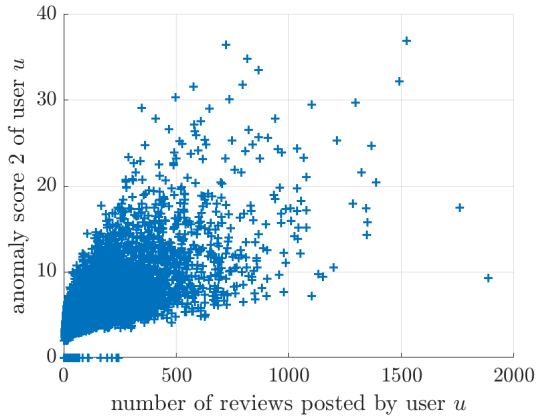
[1] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pp. 91–101, 2002.

[2] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pp. 49–56, 2000.

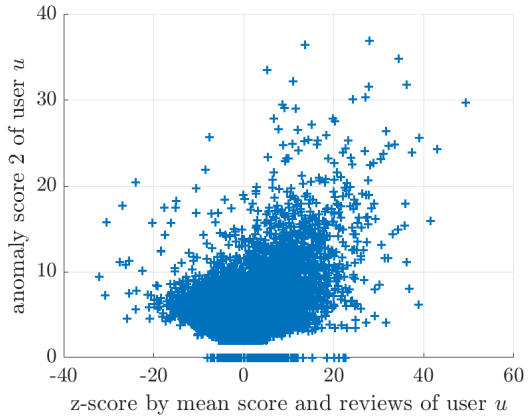
[3] Yuki Yamagishi and Kazumi Saito. Visualizing switching regimes based on multinomial distribution in buzz marketing sites. In *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017*, Vol. 10352 of *Lecture Notes in Computer Science*, pp. 385–395. Springer, 2017.

[4] J. Rissanen. Modeling by shortest data description. *Automatica*, Vol. 14, No. 5, pp. 465–471, September 1978.

[5] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other.



(a) 各ユーザの投稿レビュー数と提案異常値 2



(b) 各ユーザの平均評点と投稿レビュー数による z-score と提案異常値 2 の比較

図 2: 提案異常値 2 の評価

Ann. Math. Statist., Vol. 18, No. 1, pp. 50–60, 03 1947.

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