

多項分布型レジームスイッチングと多群順位統計量によるレビュー分析

古山 陽菜† 小野田 颯† 中村 章吾† 山岸 祐己† 斉藤 和巳‡

† 静岡理科大学 情報学部 ‡ 神奈川大学 理学部

1 はじめに

一般に、レビュー点数の時系列データを可視化するためには、単位時間やウィンドウサイズを設定し、移動平均値などに変換する必要があるが、本論文では、レビュー点数をカテゴリカルデータとして扱うことによって、パラメータを一切設定することなく可視化するとともに、レビュー点数の出現確率の変動を定量的に評価する手法を提案する。評価実験では、大規模なレビューデータセットを用いて、多様なサイズのレビュー時系列データに対する提案手法の有効性を検証する。また、実験において、提案する二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 提案手法

2.1 多項分布レジームスイッチング

時系列データを $\mathcal{D} = \{(s_1, t_1), \dots, (s_N, t_N)\}$ 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、 $t_1 \leq \dots \leq t_n \leq \dots \leq t_N$ となる。 n はタイムステップとし、 $N = \{1, 2, \dots, N\}$ 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 $T_k \in N$ 、 $\mathcal{T}_K = \{T_0, \dots, T_k, \dots, T_{K+1}\}$ 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, \dots, 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、 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 $k \in \{0, \dots, K\}$ に対して $N_k = \{n \in N; T_k \leq n < T_{k+1}\}$ のように定義する。なお、 $N = N_0 \cup \dots \cup N_K$ 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、 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 $\mathcal{P}_K = \{p_0, \dots, p_K\}$ 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。

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

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

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

のように変換したダミー変数である。各レジーム $k = 0, \dots, K$ と各状態 $j = 1, \dots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in N_k} s_{n,j} / |N_k|$ 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。

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

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

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

となる。ここで、 $\hat{p}_{0,j} = \sum_{n \in N} s_{n,j} / N$ 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

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

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。なお、式 (5) の最大化については既存研究のアルゴリズム [1] を採用し、最小記述長 (Minimum Description Length) 原理に基づいて終了させることでパラメータフリーな手法とする。

Review Analysis With Switching Multinomial Distribution Type Regime Switching Detection and Multi-Category Appearance Order Statistics

†Haruna FURUYAMA †Hayate ONODA †Shogo NAKAMURA

†Yuki YAMAGISHI ‡Kazumi SAITO

†Shizuoka Institute of Science and Technology

‡Kanagawa University

2.2 多群順位統計量

前述の問題設定同様，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 N と \mathcal{J} 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，タイムステップ n まで
 のカテゴリ j の出現数は $I_{j,n} = \sum_{i=1}^n q_{j,i}$ のように表せ
 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 [2] 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 $\{n \in N : q_{j,n} = 1\}$ と，このカテゴリ
 に属さないタイムステップ集合 $\{n \in N : q_{j,n} = 0\}$ 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，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．

$$z_{j,n} = \frac{u_{j,n} - \mu_{j,n}}{\sigma_{j,n}} \quad (6)$$

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．

$$u_{j,n} = \sum_{i=1}^n nq_{j,i} - \frac{I_{j,n}(I_{j,n} + 1)}{2}, \quad (7)$$

$$\mu_{j,n} = \frac{I_{j,n}(n - I_{j,n})}{2}, \quad (8)$$

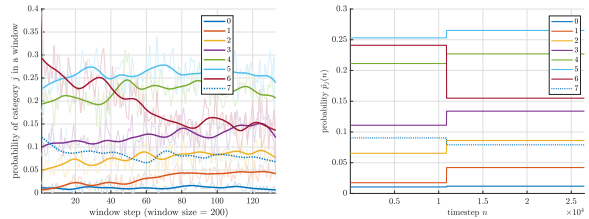
$$\sigma_{j,n}^2 = \frac{I_{j,n}(n - I_{j,n})(n + 1)}{12}. \quad (9)$$

以上より，式 (6) 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 (6) 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．この多群順位統計量は，基本的には 2 クラス分類
 器の SVM (Support Vector Machine) [3] を多クラス分類

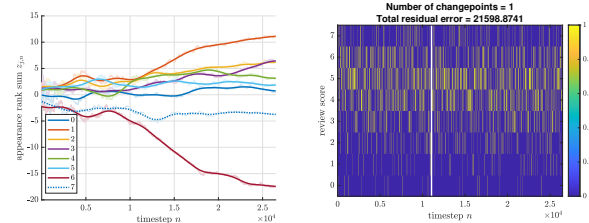
器に拡張するとき利用される one-against-all と類似
 した考え方となる．

3 評価実験とまとめ

コスメレビューサイトの @cosme から 2017 年 1 月に
 取得したデータのうち，最もレビュー数が多かったアイ
 テムのレビュー点数 ($J = 8$) に対して評価実験 (図 1) を
 行った．MATLAB R2020b において，Intel(R) Core(TM)
 i7-10710U CPU @ 1.10GHz を用いて 10 回の平均計算
 時間を計測した結果，MATLAB に実装されている find-
 changepts [4] が 0.77 秒 (検出位置 $n = 11056$) だったの
 に対し，多項分布レジームスイッチングは 0.39 秒 (検
 出位置 $n = 10898$) と，提案手法が十分高速であること
 がわかった．また，多群順位統計量は，平均計算時
 間が 0.0041 秒と十分高速であるとともに，他の二手法
 同様， $n \approx 11000$ あたりから大きく値が変化している
 ことが見て取れた．



(a) ウィンドウサイズ 200 の出 (b) 多項分布レジームスイッ
 現確率 チングの結果



(c) 多群順位統計量による変換 (d) MATLAB R2020b find-
 変換 changepts の結果

図 1: 評価実験結果

参考文献

- [1] 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.
- [2] 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.
- [3] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [4] Rebecca Killick, Paul Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, Vol. 107, pp. 1590–1598, 12 2012.