

数値属性の順序に基づく重要カテゴリの抽出手法

Nakazaki Flores Luis Alberto Chikao † 塚本 慎司 † 仲村 聡眞 † 芹澤 幸翼 †

山岸 祐己 † 齊藤 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

数値属性とカテゴリを有するデータにおいて、数値属性に大きく関わるカテゴリを抽出することは、数値属性の推定モデルの構築において重要であるといえる。よって、本論文では、時系列のカテゴリカルデータに対する分析手法を、数値属性に基づいてソートされたカテゴリカルデータに適応し、数値属性の推移とともに出現確率が大きく変動するようなカテゴリを抽出することを目的として、その変動量を定量的に評価する。また、それらの可視化結果によって、どのカテゴリが数値属性に大きく関与しているかも推定する。評価実験では、動画の再生数情報を用いて、再生数の増減に関与するカテゴリが抽出できるか検証する。

2 提案手法

2.1 多項分布レジームスイッチング

ある数値属性に基づいて昇順にソートされたデータを $\mathcal{D} = \{(s_1, t_1), \dots, (s_N, t_N)\}$ とする。ここで、 s_n と t_n は、 J カテゴリの状態と n 番目の観測ステップをそれぞれ表す。 $|\mathcal{D}| = N$ を観測数とすると、 $t_1 \leq \dots \leq t_n \leq \dots \leq t_N$ となる。 n は観測ステップとし、 $\mathcal{N} = \{1, 2, \dots, N\}$ を観測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 $T_k \in \mathcal{N}$ 、 $\mathcal{T}_K = \{T_0, \dots, T_k, \dots, T_{K+1}\}$ をスイッチング観測ステップ集合とし、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観測ステップであり、 $T_k < T_{k+1}$ を満たすとする。そして、 \mathcal{N}_k を k 番目のレジーム内の観測ステップ集合とし、各 $k \in \{0, \dots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leq n < T_{k+1}\}$ のように定義する。なお、 $\mathcal{N} = \mathcal{N}_0 \cup \dots \cup \mathcal{N}_K$ 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、 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。

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

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) の最適推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ 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。

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

したがって、スイッチング観測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

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

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

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

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。

2.2 多群順位統計量

前述の問題設定同様、観測ステップ集合と、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

A Method for Extracting Important Categories Based on the Order of Numerical Attributes

†Luis Alberto Chikao NAKAZAKI FLORES †Shinji TSUKAMOTO

†Soma NAKAMURA †Kosuke SERIZAWA †Yuki YAMAGISHI

‡Kazumi SAITO

†Shizuoka Institute of Science and Technology

‡Kanagawa University

のとき、観測ステップ 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、観測ステップ 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、観測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 (数値属性の値が大きい)、または逆に小さい (数値属性の値が小さい) 観測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観測ステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 1\}$ と、このカテゴリに属さない観測ステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 0\}$ 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、観測ステップ 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。

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

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

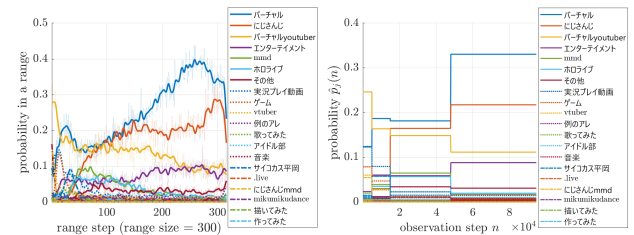
$$\sigma_{j,n}^2 = \frac{I_{j,n}(n - I_{j,n})(n + 1)}{12}. \quad (9)$$

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、観測ステップ 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。

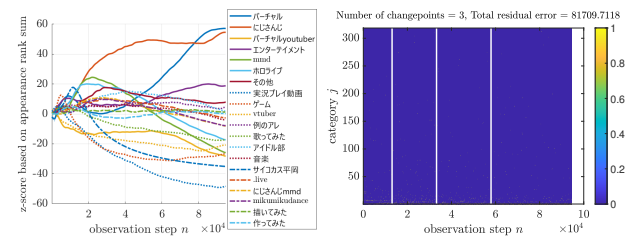
3 評価実験とまとめ

動画共有サイトの niconico に投稿された「VTuber」に関するタグを有する 98335 動画においてタグ情報を

集計し、先頭のタグとして 10 回以上出現している 318 種類のタグ ($J = 318$) と、それらを先頭のタグとして有する 94670 動画 ($N = 94670$) を対象として評価実験 (図 1) を行った。多項レジームスイッチングの結果は、各タグの出現確率推移を要約した形になっていることが見て取れる。また、提案手法が自動で検出したレジーム数 ($K = 4$) に基づき、MATLAB R2021a に実装されている findchangepts [4] の変化点数を 3 に設定すると、同様の結果が得られたため、提案手法の有意性の高さがうかがえる。さらに、多群順位統計量による変換値は、各タグを先頭のタグとして有する動画の再生数の偏りを定量的に評価できていることが確認できた。



(a) ウィンドウサイズ 300 の出現確率推移 (b) 多項分布レジームスイッチングの結果



(c) 多群順位統計量による変換 (d) MATLAB R2021a findchangepts の結果

図 1: 評価実験結果 (凡例は出現数上位 20 タグ)

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] 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 change points with a linear computational cost. *Journal of the American Statistical Association*, Vol. 107, pp. 1590–1598, 12 2012.