6Q-07

価格順データにおけるアイテムの重要属性の抽出

塚本 慎司† 仲村 聡眞† 山岸 祐己† 祝田 龍一‡ 斉藤 和巳§
†静岡理工科大学 情報学部 ‡株式会社ジャストサービス・ネット §神奈川大学 理学部

1 はじめに

価格情報とカテゴリを有するデータにおいて，アイテムの価格に大きく関わるカテゴリを抽出することは，価格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，価格順にソートされたカテゴリカルデータに適応し，価格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，各カテゴリがどの価格帯で大きく関与しているかも推定する．評価実験では，中古車価格情報を用いて，各通称名の価格決定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

価格順（昇順）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，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．n は観測ステップとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ 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 $T_k \in \mathcal{N}$，$\mathcal{T}_K = \{T_0, \cdots, T_k, \cdots, T_{K+1}\}$ 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，$T_1, \cdots, T_K$ 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 $k \in \{0, \cdots, K\}$ 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，$\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ 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．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．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

Extraction of Important Attributes of Items in Data Sorted by Price
†Shinji TSUKAMOTO †Soma NAKAMURA †Yuki YAMAGISHI
‡Ryuichi HODA §Kazumi SAITO
†Shizuoka Institute of Science and Technology
‡Just Service Net Inc.
§Kanagawa University

られたとき，出現順位の値が大きい（価格が高い），または逆に小さい（価格が安い）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．

$$z_{j,n} = \frac{u_{j,n} - \mu_{j,n}}{\sigma_{j,n}}. \tag{6}$$

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 $\sigma^2_{j,n}$ は次のように計算される．

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

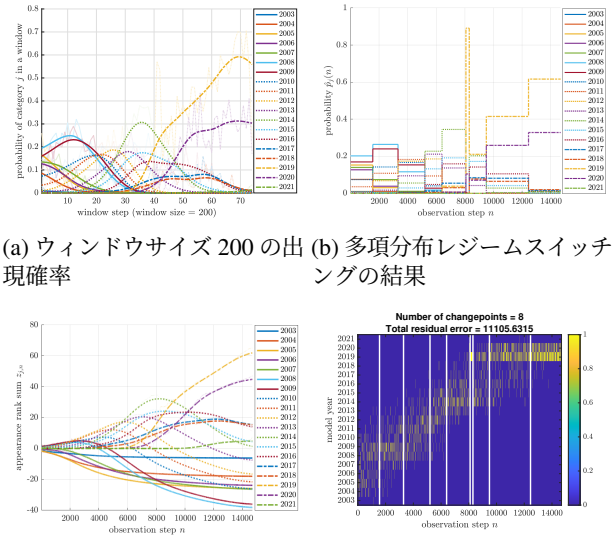
$$\mu_{j,n} = \frac{I_{j,n}(n - I_{j,n})}{2}, \tag{8}$$

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

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（価格が高い），または逆に小さい（価格が安い）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

中古車情報サイトの carsensor から 2021 年 1 月に取得したデータのうち，最も台数が多かった通称名に対して評価実験（図 1）を行った．ここで，データは価格順で昇順ソートし，対象カテゴリとして年式（$J = 19$）を扱った．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 findchangepts [4] が 1.12 秒（検出変化点数 8）だったのに対し，多項分布レジームスイッチングは 13.05 秒（検出スイッチング観測ステップ $K = 8$）と，10 倍近い時間がかかったが，提案手法も実用時間で十分実行可能であることがわかった．ここで，findchangepts で設定した変化点数は，提案手法が自動で検出したスイッチング観測ステップ数であるため，両手法の結果がほぼ同様となったことから，提案手法の有意性の高さがうかがえる．さらに，多群順位統計量は，平均計算時間が 0.0046 秒と十分高速であるとともに，価格の推移に連動して出現確率が大きく変化する重要カテゴリを，定量的に評価できていることが見て取れる．



(a) ウィンドウサイズ 200 の出現確率　(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値　(d) MATLAB R2020b findchangepts 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.