O-031

駅を基準とした近隣地域における賃貸情報の類似度分析

佐藤 亘† 能代 哲太† 齊藤 光† 山岸 祐己†‡ 青木 成樹§ 橋本 正洋¶
† 静岡理工科大学 ‡ 理化学研究所 § マリンオープンイノベーション機構 ¶ 法政大学

1 はじめに

一般に，賃貸情報サイトの検索結果や地価公示価格を参照することによって，ある程度各地域の家賃相場が推定できる．しかし，その地域での住みやすさを推定するためには，モデル地域となる別の地域との類似度を測定するアプローチ等が必要となる．よって，本研究では，賃貸情報を利用した地域間類似度として，順位和検定を利用した新たな距離を提案する．より具体的には，間取りのような複数カテゴリと，主要駅からの距離や家賃といった系列に基づく順位を用いることによって，各地域を特徴づけることを試みる．現実データにおける実験では，同じ路線内の各駅の近隣地域間の類似度を求めるとともに，より大きい地域との類似度を求めることで，モデル地域の推定も行う．

2 提案手法

複数カテゴリの時系列的変化を比較するため，出現順位を用いた統計量によるデータ変換を行う．この手法は，Mann-Whitney の U 検定 [1] を基盤とし，多群を扱えるよう拡張したものであり，データの出現頻度の傾向変化を z-score として表現する．静的な分析手法による指標を，動的な視点で捉えられるよう可視化するため，基準となる系列においての変化の指標化が期待できる．また，各カテゴリの z-score の最終値を特徴量とすれば，同じ次元（カテゴリ）数における類似度分析等が可能となる．

賃貸物件データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，s_n と t_n は，J カテゴリの属性と，基準となる系列における n 番目の数値をそれぞれ表す．$|\mathcal{D}| = N$ を物件数とすると，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．n は基準系列のソート結果におけるインデックスとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ をインデックス集合とする．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．このとき，インデックス 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，インデックス 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，インデックスとカテゴリの集合が与えられたとき，系列の値が大きい，または逆に小さいインデックス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インデックス集合 $\{n \in \mathcal{N} : q_{j,n} = 1\}$ と，このカテゴリに属さないインデックス集合 $\{n \in \mathcal{N} : q_{j,n} = 0\}$ 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．

$$z_j = \frac{u_j - \mu_j}{\sigma_j}. \tag{1}$$

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．

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

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

$$\sigma_j^2 = \frac{I_{j,N}(I_N - I_{j,N})}{12} \tag{4}$$

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，各インデックスが複数のカテゴリを有するケースでは，式 (4) において同順位補正が必要となる．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) [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，最新インデックス 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）インデックスを有意に多く含むかを定量的に評価することができる．また，z-score z_j の計算量は全てのインデックス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インデックスごとに $O(J)$ の計算量しかかからない．なお，対象が数値属性であっても，適当な J カテゴリに離散化すれば，本手法が適応可能となる．また，異常検知を目的として有意水準を設定すれば，z_j から求まる有意確率を使った仮説検定が可能である．

Similarity Analysis of Rental Information in Neighborhoods Based on Railway Stations
†Wataru SATO †Tetta NOSHIRO †Hikaru SAITO †‡Yuki YAMAGISHI §Shigeki AOKI ¶Masahiro HASHIMOTO
†Shizuoka Institute of Science and Technology
‡RIKEN
§MaOI
¶Hosei University

3　評価実験とまとめ

国内最大級の賃貸情報サイトである SUUMO * における，静岡鉄道静岡清水線（静岡市内を東西に走る路線）の各駅から半径 1 km 圏内の物件情報と，東海道本線の藤枝駅，浜松駅，静岡駅の各駅から半径 3 km 圏内の物件情報を用いて，提案手法を適応する．今回は基準系列を駅からの距離 (km の昇順) とし，カテゴリを間取り（代表的な 7 種類の間取り $J = 7$）とした．提案手法の z-score z_j と，比較として各間取りの物件数および平均距離 (km) を各カテゴリの特徴量とし，コサイン類似度における階層クラスタリング（群平均法 [3]，閾値は最大連結距離の 30 %）の結果を用いて検証を行う．なお，頻度が 0 となる駅と間取りの組み合わせ（今回は 1 パターンのみ）の平均距離については，当該駅の他カテゴリの指標の平均値を用いることとする．

図 1 より，物件数を特徴量としたときは，かつて清水市の中心地だった新清水が他の駅と乖離していることが分かる．また，図 2 より，平均距離を特徴量としたときは，古庄・長沼・日吉町といった子育て世代に人気のエリアが乖離している．両図より，基本的な特徴量の利用によって得られる知見はあるものの，住みやすさの指標化等を考慮すると，さらに明確な対極的構造が求められる．

一方，提案手法による図 3 は，静岡市の中心地である新静岡・日吉町を乖離させ，その対極に県立美術館前・草薙といった大学生が居住するエリアを位置付けている．さらに，東海道本線の静岡県内主要駅の特徴量を追加した提案手法の結果（図 4）では，多くの駅が静岡駅の特徴の縮図であることを示唆するとともに，浜松駅や藤枝駅といった工業都市の特徴が，静岡市全体や静岡市の中心街の特徴とは異なることも示唆している．

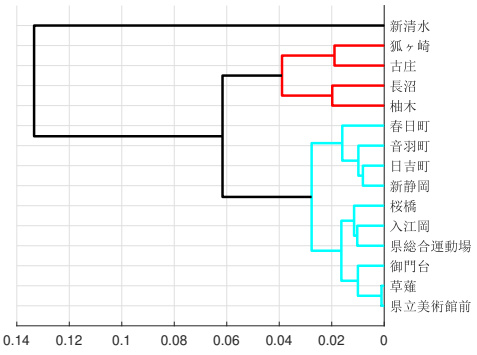

図 1: 各間取りの物件数によるクラスタリング結果

*https://suumo.jp/

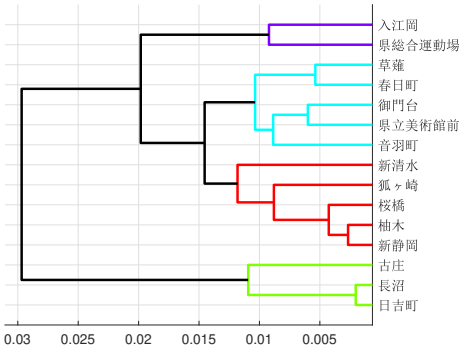

図 2: 各間取りの平均距離によるクラスタリング結果

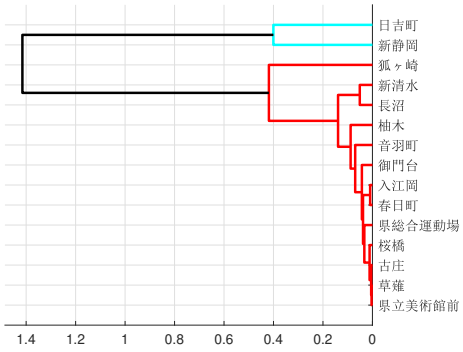

図 3: 各間取りの z_j によるクラスタリング結果

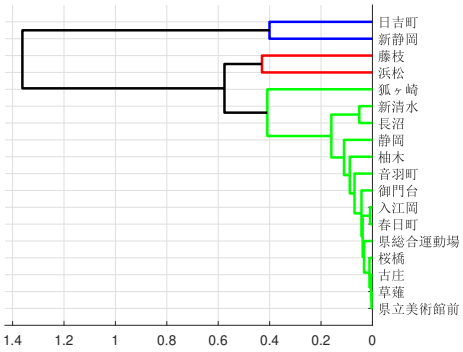

図 4: 各間取りの z_j によるクラスタリング結果（東海道本線の静岡県内主要駅を追加）

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.

[3] R.R. Sokal, C.D. Michener, and University of Kansas. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas, 1958.