4B-05

カテゴリ変数セットの距離を用いた ファッション・コーディネートのコミュニティ分析

大槻　明†　　川村雅義†

日本大学†

1. はじめに

カテゴリカルデータを対象としたクラスタリング手法自体はこれまでに提案されてきたが[1-5]，ファッション・コーディネートデータのように，ファッションアイテムのセットで表されるカテゴリ変数に対してクラスタリングを行う先行研究は殆どない．ゆえに，本研究では，コーディネート内アイテムの組合せパターンの類似性に着目し，この組合せパターンから一致度を計算するためのモデル開発した．そして，この一致度を反転して距離に変換し，この距離に基づいたクラスタリング手法を提案する．評価実験では，Wear サイト[6]に掲載されているコーディネートデータを用いて 2 通りの検証を行った．

2. 提案手法

2.1 コーディネートデータの定義

1 コーディネートデータは，1～n のアイテムから構成される．ここでいうアイテムとは，式 1 に示すように 3 つの属性のセットである．1 つ目の属性はアイテムタイプ（例：シャツ，スカート，靴など），2 つ目はアイテムのブランド，3 つ目はアイテムの色である．

$$1Coordinate = \{(itemType1, brand1, color1), (itemType2, brand2, color2), \dots\} \quad (1)$$

2.2 コーディネートデータの一致度計算

コーディネートデータ i と j の一致度を計算するモデル CPM_{ij}(Coordination Patterns Muched)を式 2 のとおり定義する．

$$CPM_{ij} = M_{ij} + \frac{S_{ij}}{2*M_{ij}} \times 0.99 \quad (2)$$

M_{ij} は，コーディネートデータ ij 間で一致したアイテムタイプの数を表し，S_{ij} は，一致したアイテムタイプにおけるブランド or カラーの一致数を表す．なお，ブランド及びカラーが共に一致した場合を想定して，分母の M_{ij} に 2 を掛けている．最後に，末尾の 0.99 は，アイテムタイプだけが一致する場合との重複を避けるための調整用の値である．

2.3 クラスタリングモデル

前節で計算したコーディネート間の一致度を一致度行列（Similarity Matrix，SM）として表すと式 3 のようなイメージとなる（数値はダミー）．

$$SM = \begin{bmatrix} 0.00 & 0.00 & 1.00 & 1.99 \\ 0.00 & 0.00 & 2.74 & 2.74 \\ 1.00 & 2.74 & 0.00 & 1.99 \\ 1.99 & 2.74 & 1.99 & 0.00 \end{bmatrix} \quad (3)$$

そして，式 3 を頂点の集合 V と辺の集合 E の対として表現すると，コーディネートデータ無向グラフ（CoG）は式 4 のとおり表わされる．$\{\}^n$ の n は CPM_{ij} を表す．また，式 4 をグラフネットワークとして表したものが図 1 である．

$$CoG = V, E$$
$$V = \{c1, c2, c3, c4\},$$
$$E = \{c1, c3\}^{1.00}, \{c1, c4\}^{1.99}, \{c2, c3\}^{2.74}, \{c2, c4\}^{2.74}, \{c3, c4\}^{1.99} \quad (4)$$

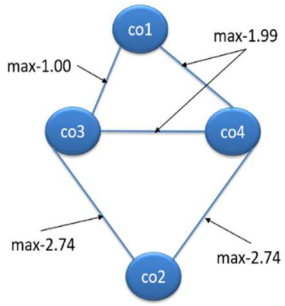

図 1　グラフネットワークイメージ

（c_n はコーディネートデータ，数値は CPMij の最大値=max から各 CPMij を引いた値）

3. 評価実験

3.1 評価実験の対象データ

図 2 は，Wear サイトに 2013 年 1 月から 2014 年 12 月の間に掲載された約 15 万件のコーディネートデータ数を四半期ごとに分けて表示したものである．本実験では，2014 年 1-3 月期に合わせる形で，2014 年の四半期ごとに 1600 件ずつサンプリング抽出したうえで実験した．

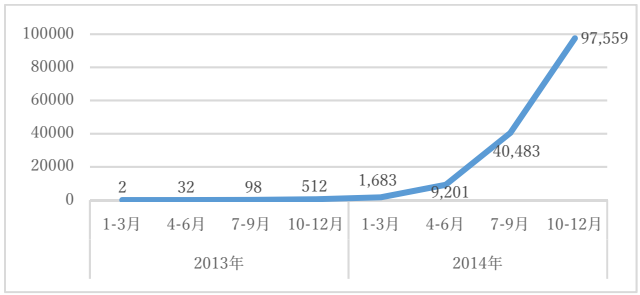

図 2　取得したコーディネートデータ数

3.2 シルエット分析による検証

クラスタ内の面が適切に分割されているかどうかを検証するためにシルエット（silhouette）値[7]を用いて検証した．シルエット値は，他のクラスタの点と比べてその点が自身のクラスタ内の他の点にどれくらい相似しているかを示す尺度であり，i 番目の点のシルエット値 S_i は，式 5 のように求められる．

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (5)$$

なお，2.2 節で示した本手法（一致度計算手法）

に良好なクラスタリング手法についても調べるために，本手法は次に示すクラスタリングで試行した．

- 辺媒介性に基づく手法（edge betweenness）[8]
- ランダムウォークに基づく手法（walktrap）[9]
- infomap 法に基づく手法（infomap）[10]

図 3 は，図 2 の 2014 年度 1-3 月期の 1600 件のデータを用いて，上述した各手法でシルエット値を求めた結果であるが，k-medoids，k-modes 及び rock よりも，Edge betweenness，Walktrap 及び Infomap の方がシルエット値（平均値）が高かった．

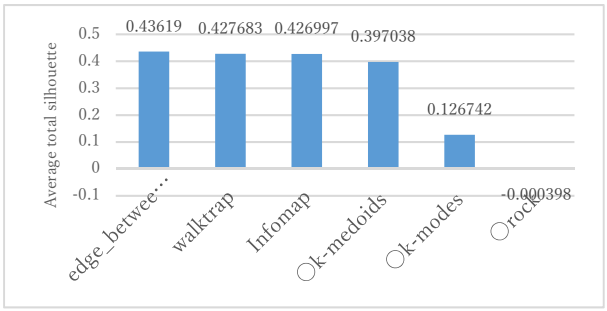

図 3　Average total silhouette

3.3　クラスタの特徴分析によるアイテム単体のトレンドアイテムの調査

本手法＋edge betweenness を用いてクラスタリングを行った結果を図 4 に示す．図 4 の四半期ごとにメンバ数上位 5 位までのクラスタを対象にトレンドアイテムタイプを調べた．

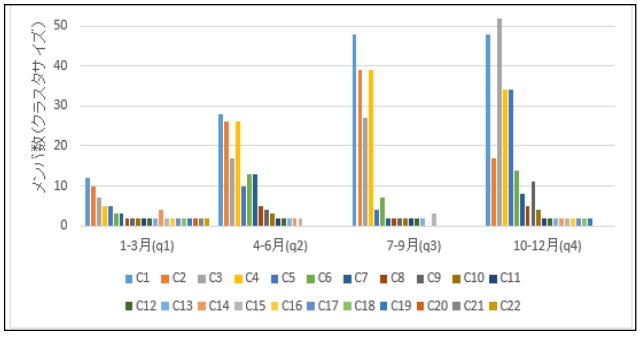

図 4　2014 年度四半期の各クラスタのメンバ数

（C_n の C はクラスタ，n はクラスタ番号を表す）

本研究では，式 6 に示すとおりクラスタ内全アイテムタイプ数の 1/2 以上頻出しているアイテムタイプをトレンドアイテムタイプ（TI）と定義した．

$$TI = IF \geq \frac{IA}{2} \qquad (6)$$

IF はアイテムタイプの頻出数を，IA はクラスタ内の全アイテムタイプ数をそれぞれ表す．図 5～8 に四半期ごとの各クラスタの特徴（どのようなトレンドアイテムタイプが登場していたか）を示す．四半期を通して多くのクラスタに登場していたトレンドアイテムタイプは，「シャツ・ブラウス」，「T シャツ・カットソー」，「パンツ」，「スニーカー」であった．

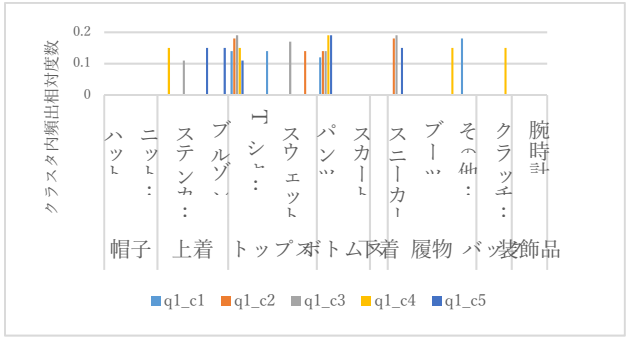

図 5　2014 年度 1-3 月期の各クラスタの特徴

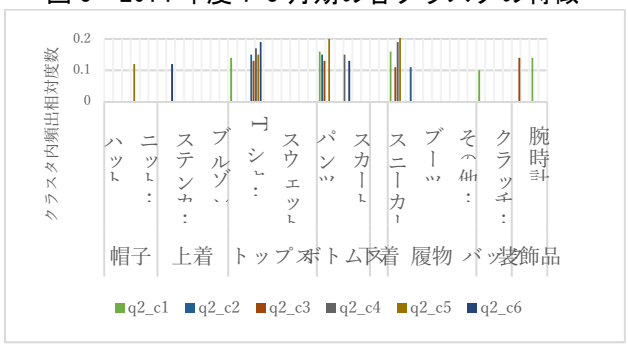

図 6　2014 年度 4-6 月期の各クラスタの特徴

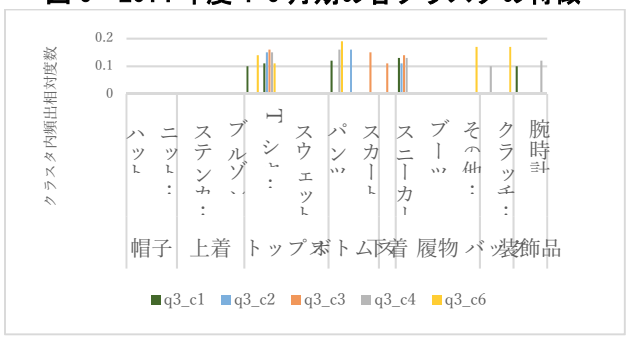

図 7　2014 年度 7-9 月期の各クラスタの特徴

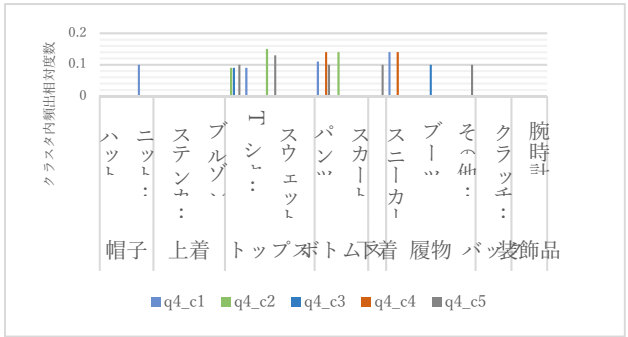

図 8　2014 年度 10-12 月期の各クラスタの特徴

参考文献

[1] Huang, ZX., "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, Vol.2, No.3, pp.283-304, 1998.
[2] Huang, ZX, Ng, MK., "A fuzzy k-modes algorithm for clustering categorical data", IEEE Transactions on Fuzzy Systems, Vol.7, No.4, pp.446-452, 1999.
[3] Ahmad, Amir, "Dey, Lipika: A k-mean clustering algorithm for mixed numeric and categorical data", Data & Knowledge Engineering, Vol.63, No.2, pp.503-527, 2007.
[4] Guha, S., Rastogi, R. and Shim, K., "ROCK: A Robust Clustering Algorithm for Categorical Attributes", ICDE, 1999.
[5] Kaufman L., Rousseeuw P.J., "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons, 1990.
[6] Wear サイト, "http://wear.jp/
[7] Rousseeuw, P.J., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". J. Comput. Appl. Math., 20, pp.53-65, 1987.
[8] Girvan, M. & Newman, M. E. J., " Community structure in social and biological networks", Proc. Natl. Acad. Sci. USA 99, 7821-7826, 2002.
[9] Pascal Pons, Matthieu Latapy, "Computing Communities in Large Networks Using Random Walks", Journal of Graph Algorithms and Applications, Vol. 10, no. 2, pp. 191-218, 2006.
[10] Martin Rosvall,Carl T. Bergstrom, "Maps of random walks on complex networks reveal community structure", PNAS,vol.105, no.4, pp.1118-1123, 2008.