クラスタリングの概念と意思決定支援への応用

本多　克宏

ヒトは何かを認識する際に，ある法則に従ってまとまりをつくる「群化」を行うことが心理学で指摘されている．データ分析においても初歩的アプローチとして，似たもの同士からなるクラスターへの教師なし分類が有力な手法として確立されており，「クラスタリング」と総称される．本稿では，いくつかのクラスタリング手法の概念やアルゴリズムを紹介するとともに，意思決定支援への展開が期待される共クラスタリング手法を概説し，応用事例としてインターネット上で「くちコミ」を疑似的に実装する協調フィルタリングへの適用について述べる．

キーワード：クラスタリング，共クラスタリング，協調フィルタリング

1. はじめに

人間は複雑な事象を説明しようとする際に，ある種の単純化を通した要約を試みる．例えば，多様な植物からなる森を目にした場合，そこに生える木々をコンピュータのごとく１本ずつすべてリストアップするのでは，まさに「木を見て森を見ず」となり人間らしい理解とは言えない．むしろ人間は同種の植物を大まかな一くくりとみなして，森の全体を概観するであろう．

人間が持つ自然な知覚特性として，ある法則に基づいて事象を要約する群化（体制化）が心理学でも指摘されており [1]，ゲシュタルト心理学と呼ばれる学派が構成された．関連性の強い物体形状を一まとまりとしてとらえることで，意味のある文字や記号を知覚する行動などが代表例である．

クラスタリング（あるいは，クラスター分析とも呼ばれる）は，情報の群化を数理モデルにおいて実現する技術ととらえられるものである．多くの対象（標本，個体など）を互いに似たもので構成されるグループ（クラスターと呼ばれる）にまとめることを目標とする．判別分析が教師情報（クラス情報）を元にして未知の対象を既存のいずれかのクラスへ分類するのに対して，クラスタリングは教師情報を持たずに未知のクラスへ分類することから，教師なし分類とも呼ばれる．情報の単純化を通した情報要約や層別のために，多変量データ解析の前処理に用いられることが多いが，多変量データからのデータマイニング手法としても展開が広がっている．本稿では，いくつかのクラスタリング手法の概念やアルゴリズムを紹介するとともに，意思決定支援への展開が期待される共クラスタリング手法を概説し，応用事例としてインターネット上で「くちコミ」を疑似的に実装する協調フィルタリングへの適用について述べる．

ほんだ　かつひろ
大阪府立大学大学院工学研究科
〒 599–8531　大阪府堺市中区学園町 1–1

2. 主なクラスタリング手法の概念

2.1 階層的な手法と非階層的な手法

クラスタリング手法は，大別すると階層的な手法と非階層的な手法に分けられる [2]．階層的な手法では，個々の対象を類似度の大きいものから順に結合していく（もしくは全体集合を順に分割していく）ことで，データの構造が階層構造（樹形図，デンドログラム）として表現される．例えば，未分類の動物の集合から系統樹を構築しようとする問題に相当する．目的に応じて大分類から小分類まで任意の数のクラスターに分類できる利点もあるが，分類対象が多い場合には対象間の距離の計算が膨大となり，実用的でない．

一方，非階層的な手法では，クラスター数があらかじめ指定されてバッチ処理により一括で抽出するモデルや，逐次処理により個体の割り当てが行われるモデル，ある程度のまとまりを逐次的に抽出するモデルなどが用いられる．階層的な手法に比べて計算効率がよく，大規模なデータの処理に適していることから，さまざまな研究が進んでいる．

2.2 *k*-Means から Fuzzy *c*-Means へ

n 個体についての m 次元観測値（変量，項目などの値）$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})^T$，$i = 1, \ldots, n$ が与えられた際に，n 個体をあらかじめ定められた個数（C 個とする）のクラスターに分類する問題を考える．

プロトタイプに基づく非階層的な手法の代表例に k-

Means [3] がある．各クラスターを代表するプロトタイプとして平均ベクトル $\boldsymbol{b}_c$ を用い，クラスター内でのプロトタイプと個体間の 2 乗距離の総和が最小となるように，個体の割り当てと平均ベクトルの算出が繰り返される．

$$L_{km} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \left\| \boldsymbol{x}_i - \boldsymbol{b}_c \right\|^2 \tag{1}$$

u_{ci} は個体 i のクラスター c への帰属度を表すメンバシップであり，帰属すれば 1，それ以外は 0 となり，c についての和が 1 となる制約が課される．

メンバシップを $u_{ci} \in [0,\ 1]$ の実数値に拡張することで，クラスターへの帰属をより柔軟に表現するアプローチにファジィクラスタリングがある．Bezdek らは k-Means の目的関数が u_{ci} について非線形関数となるように，べき乗重みを付加した Fuzzy c-Means（FCM）[4] の目的関数を提案した．

$$L_{fcm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta} \left\| \boldsymbol{x}_i - \boldsymbol{b}_c \right\|^2 \tag{2}$$

θ が分割のファジィ度を調整し，1 のときは k-Means に帰着され，大きくなるほど分割があいまいとなる．k-Means のクリスプな分割に比べて，ノイズの影響が軽減され，大域的最適解の頻度が増加する効果がある．

また，k-Means の目的関数へ u_{ci} についての非線形項を付加してファジィ化する手法も提案されている [5, 6]．

2.3 確率モデルに基づくクラスタリング

多峰性の確率密度関数を推定する混合分布モデル [7] は，要素分布への帰属確率をファジィメンバシップと同一視すれば，クラスタリングモデルともみなされる．特に，ガウス混合分布（正規混合分布）は，要素分布の分散共分散行列を対角行列とし，分散要素を 0 に漸近すれば，k-Means を近似したモデルとなる．

EM アルゴリズムを用いたガウス混合分布モデル推定 [8] や決定論的アニーリングによるデータ分類 [9] などは，確率モデルに基づくソフトな分類手法の代表例である．

2.4 その他の展開

非線形な境界を持つクラスターの抽出のために，高次元特徴空間への写像を考慮したカーネル法の適用も試みられている [5]．また，クラスターのプロトタイプを直線や線形多様体，2 次曲面などに拡張するなど，種々の形状を持つクラスターの抽出に関する提案も多い [10]．

回帰モデルをプロトタイプとする Fuzzy c-Regression Models（FCRM）[11] や主成分モデルとの融合である線形ファジィクラスタリング [6] などは，局所的な多変量データ解析の手法としても有効で，データマイニングへの利用価値が高い．

3. 共クラスタリングへの展開

3.1 k-Means における変量選択と共起関係行列のクラスタリング

k-Means において，個体の分類のみにとどまらず，個体を特徴づける変量についても重要度を推定するアプローチが提案された．Huang らによる W-k-Means [12] では，個体については k-Means と同様にクリスプな分類を行いながら，FCM の概念に基づいて変量の重要度 w_j，$j = 1, \ldots, m$ を同時に推定している．

$$L_{wkm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} w_j^{\theta} (x_{ij} - b_{cj})^2 \tag{3}$$

w_j は変量間の相対的な重要度を表し，j に関する和が 1 となる制約が用いられる．クラスター構造を強調する（クラスター内誤差の小さい）変量が重要視される．

類似した概念として，個体と項目の共起関係行列から関連の強い個体と項目を同時に抽出する共クラスタリングが考えられた．個体 i と項目 j の関連性の度合いを r_{ij} とし，値が大きいほど関連性が高いものとする．例えば，個体を文書，項目をキーワードとし，文書中でのキーワードの使用頻度を r_{ij} とおく．文書クラスタリングでは，類似した文書群をクラスターとして抽出するのと同時に，クラスター内で使用頻度の高いキーワードを見つけることが目的となる．

クラスター c での項目 j の重要度を w_{cj} とおいたとき，呉らはクラスター内での凝集度を基準とした目的関数を用いる Fuzzy Clustering for Categorical Multivariate Data（FCCM）[13] を提案した．

$$\begin{aligned} L_{fccm} = & \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} w_{cj} r_{ij} + \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log u_{ci} \\ & + \lambda_w \sum_{c=1}^{C}\sum_{j=1}^{m} w_{cj} \log w_{cj} \end{aligned} \tag{4}$$

エントロピー項はファジィ化のために付加された非線形項 [5] である．また，FCCM を大規模データに適応可能に拡張した Fuzzy CoDoK [14] なども提案されている．

これらの方法では，個体については c についての和

が 1 となる制約から排他的な割り当てがされるのに対して，項目については j についての和が 1 となる制約が用いられるために，クラスター内での相対的な重要度のみが推定され，クラスターへの項目の絶対的な帰属度とはなっていない．そのため，複数のクラスターで w_{cj} が大きな値を持つ項目や，逆にいずれにも小さな値しか持たない項目などが存在しうる．

3.2 主成分分析としての k-Means と共クラスタリングへの拡張

データ分類を目的とする k-Means を，多次元データの次元縮約を目的とする主成分分析と同一視する指摘が行われた [15]．Oja ら [16] はニューロ学習の観点から共クラスタリングモデルを構築した．$n \times m$ データ行列 $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ を行列 P により縮約する場合，$X \approx PP^TX$ による近似が行要素（個体）のクラスタリングに相当し，$X \approx XPP^T$ による近似が列要素（項目）のクラスタリングに相当する．特に，X の要素がすべて非負の場合は，Non-Negative Matrix Factorization により P の行または列が非負要素のみで互いにほぼ直交するメンバシップベクトルとなる．

上記の概念を組み合わせて，個体と項目を同時にクラスタリングする近似モデル $X \approx LL^TXRR^T$ が提案された．L および R がそれぞれ行要素（個体）および列要素（項目）のメンバシップ指標となる．

Non-Negative Matrix Factorization によるアプローチでは，メンバシップベクトルがほぼ直交することから，前節の FCCM や Fuzzy CoDok と違って，個体と項目のいずれも単一のクラスターに帰属する排他的な分割となる．

3.3 逐次的な共クラスター抽出

筆者らは，逐次的な共クラスター抽出のモデルを提案した [17]．選択的に排他的な制約を課すことができるなど，上述の二つのアプローチとは異なる観点からの提案である．

個体や項目といった区別なく，単に n 個のノード間の関係性からなるネットワークグラフにおいて，グラフ構造の均衡がとれているシステムでは互いに正の関係性を持つノードからなる二つのグループに分割されることが示されており [18]，関係性行列の最大固有値に対応する固有ベクトルから分類指標を得るアプローチが提案されている [19]．このモデルは，Oja らの主成分学習モデルを 2 クラスター問題で議論したものに相当し，固有値問題に帰着している点が異なる．

一方，標本間の非負の類似度を要素に持つ $n \times n$ 関係性行列からファジィクラスターを逐次的に抽出するアプローチも提案されている [20]．均衡化手法を多クラスター問題に拡張したモデルともとらえられ，クラスターに帰属済みの個体を排除する制約を付加しながら，逐次的に固有値問題を繰り返し解いてメンバシップ指標を推定する．

筆者ら [17] は，購買履歴情報のように個体と項目の関連が $\{0, 1\}$ で表される $n \times m$ データ行列 X を考え，上記のグラフ構造均衡化と逐次的なファジィクラスター抽出の手法を拡張することで，逐次的な共クラスター抽出法を提案した．まず，$n \times m$ 長方形状行列を正方形状に修正するために，対角ブロックに O 行列を持つ $(n+m) \times (n+m)$ ブロック行列を考える．

$$S = \begin{pmatrix} O & X \\ X^T & O \end{pmatrix} \tag{5}$$

未知の対角ブロックに O 行列をおくアイデアは，グラフ構造均衡化で未知要素を 0 とおくことに倣っている．つぎに，行列 $S = \{s_{ij}\}$ を非負の類似度行列とみなして逐次的にクラスターメンバシップベクトル $\boldsymbol{w}_k = (w_{k1}, \ldots, w_{k,n+m})^T$ を抽出するために，以下の目的関数の最大化を考える．

$$L_{sfce} = \sum_{i=1}^{n+m} \sum_{\substack{j=1 \\ j \neq i}}^{n+m} s_{ij} w_{ki} w_{kj} - \frac{1}{k-1} \sum_{t=1}^{k-1} \beta_t \, \mathrm{dup}(\boldsymbol{w}_k, \boldsymbol{w}_t) \tag{6}$$

第 1 項は呉ら [13] のクラスター内凝集度に相当する．第 2 項は抽出済みのクラスターを排除するペナルティ（内積の値など）であり，重み β_t によりクラスターの重複を防ぐ．

クラスター指標 $\boldsymbol{w}_k$ は，行列 S の最大固有値に対応する固有ベクトルとなる．ただし，第 2 クラスター以降の抽出では，ペナルティを考慮して対角要素を以下に置き換える．

$$s_{ii} = -\frac{1}{k-1} \sum_{t=1}^{k-1} \beta_t w_{ti}^2 \tag{7}$$

関係性行列のスパース性を利用すれば，固有値問題を $n \times n$ または $m \times m$ の行列に縮約できるため，計算効率を改善することもできる [21]．

個々の個体や項目にペナルティを負荷するこのアプローチは，すべての個体と項目にペナルティを負荷する場合には Oja ら [16] と同様に個体・項目ともに排他的なクラスター割り当てとなる一方，項目についてペナルティを負荷しない場合には呉ら [13] と同様にクラ

スター内での相対的な項目の重要度を算出できる．また，項目について選択的に排他的制約を負荷することもできる．

4. 意思決定支援のための協調フィルタリング

4.1 協調フィルタリングの問題設定

意思決定支援への応用事例として，協調フィルタリング [22] を取り上げる．ユーザ間の履歴を比較することで未選択・未購入のアイテムから特定のユーザに特化した推薦を提示するシステムであり，ネット通販の最大手の Amazon.com [23] などで実用化されていることで有名である．近傍に基づくアルゴリズム [22] では，ユーザ間の相関係数から類似ユーザを探索し，類似ユーザの評価値の平均値から未評価値が予測されており，さながら「ネットワーク上での疑似的なくちコミ」とみなされる．

4.2 逐次的なユーザ・アイテム共クラスター抽出による協調フィルタリング

筆者ら [17] は，逐次的な共クラスター抽出を協調フィルタリングに応用し，購買履歴を用いた実験において，近傍に基づく手法に勝る推薦性能を確認した．嗜好の類似したユーザをクラスターにまとめながら，同時に，当該ユーザ群に嗜好されるアイテムを帰属させる．図 1 にユーザ・アイテム共クラスターのイメージを示す．「O」が購入済みアイテムを示し，クラスター内で未購入のアイテムを推薦する．

多数のユーザの履歴情報をクラスター構造に縮約するアプローチは，推薦アイテムの探索効率を向上させたり，情報を保管するメモリ領域を節約したりするほか，個々のユーザの履歴を匿名化する効果もある．購買履歴はセンシティブな個人情報であり，慎重な取り扱いが求められるが，クラスター構造に情報縮約するモデルでは個人特定が不可能となっており，軽量性・匿名性を兼ね備えていることから，クラウドサーバに代表される外部サーバや簡易な端末への実装が可能になるなどの利点がある．

また，協調フィルタリングへの適用では，逐次的な共クラスター抽出での柔軟な制約条件設定が優位性を持つ．ユーザについては帰属クラスターを特定するために排他的な割り当てが求められるが，アイテムについてはその限りではない．例えば，図 1 の場合，アイテム 6 のような人気のあるアイテムは複数のクラスターで共有され，大きなメンバシップを持つべきであり，Oja らのモデルのように本質的にメンバシップが排他的となるのは好ましくない．一方で，特定のユーザ群でしか嗜好されないアイテムの場合は，排他的な制約を負荷した方が他のクラスターへの悪影響を軽減できると期待される．

4.3 アイテムに関する制約についての検証

本稿では，選択的にアイテムを排他的に分類する効果について，検証する．日本経済新聞社の NEEDS-SCAN/PANEL より，2000 年の購買調査の対象となった 996 世帯が 18 種類の製品を所有しているか否かのデータ（996 × 18 データ行列）を用いた．予測の対象製品としては，所有率が 35%程度のピアノ，VD，オーブンの 3 種類を用いた．検証では，半数のユーザをテストユーザとし，予測対象アイテムの値を購入の有無にかかわらず 0（未購入）とした際の予測値から，実際の購入の有無を推定した．逐次的な共クラスター抽出により，ユーザ・アイテムを 10 個の共クラスターに分類した．ユーザはメンバシップ最大のクラスターに割り振り，当該クラスターでのアイテムのメンバシップをもとに推薦アイテムを選別した．

図 2 に，(1) 全アイテムに排他的な制約を負荷しない場合，(2) 全アイテムに排他的な制約を負荷する場合，(3) 選択的に排他的な制約を負荷する場合，の 3 種類での推薦性能を比較する．

診断性能の指標である ROC 曲線 [24] で推薦性能を評価し，ROC 曲線の下の面積（大きいほど性能が高

	アイテム1	アイテム2	アイテム3	アイテム4	アイテム5	アイテム6
ユーザ1	O		O			O
ユーザ2	O	O				O
ユーザ3		O	O			
ユーザ4				O		O
ユーザ5				O	O	O
ユーザ6					O	O

図 1 ユーザ・アイテム共クラスターの例

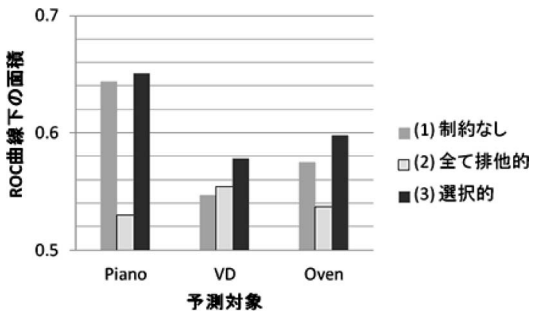

図 2 制約条件による推薦性能の比較

い）を指標値とした．また，(3) で選択的に制約を負荷するアイテムとしては，ユーザ一人当たりのメンバシップ指標値が小さく，所有ユーザ数に比してメンバシップ指標が小さいコーヒーメーカーを選んだ．

図から，(2) のように全アイテムに排他的な制約を用いると推薦性能が劣ることがわかる．一方で，(1) のように全アイテムに共有を許すのではなく，(3) のように特有性の高いアイテムには選択的に制約を負荷することで，推薦性能が向上した．したがって，アイテムの特性を考慮した柔軟な制約条件の設定が必要であるといえる．

5. おわりに

本稿では，いくつかのクラスタリング手法の概念を解説した．特に意思決定支援と関連の強い共クラスタリングに関して，代表的な手法の間の相違を議論した．また，協調フィルタリングへの応用事例を紹介し，分類における制約条件の違いが情報選別に与える影響を示した．実世界の顧客情報などへの応用ではデータの大規模化が大きな問題となる．大規模データの取り扱いは今後の重要な研究課題となる．

参考文献

[1] 菊池正（編），『感覚知覚心理学』，朝倉書店 2008.

[2] M. R. Anderberg（著），西田 英郎（訳），『クラスター分析とその応用』，内田老鶴圃, 1988.

[3] J. B. MacQueen, “Some Methods of Classification and Analysis of Multivariate Observations,” in *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, 281–297, 1967.

[4] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[5] S. Miyamoto, H. Ichihashi and K. Honda, *Algorithms for Fuzzy Clustering*, Springer-Verlag, 2008.

[6] K. Honda and H. Ichihashi, “Regularized Linear Fuzzy Clustering and Probabilistic PCA Mixture Models,” *IEEE Trans. Fuzzy Syst.*, **13**-4 (2005), 508–516.

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[8] 赤穂昭太郎，「EM アルゴリズム—クラスタリングへの適用と最近の発展—」，『日本ファジィ学会誌』，**12**–5 (2000), 594–602.

[9] K. Rose, E. Gurewitz and G. Fox, “A Deterministic Annealing Approach to Clustering,” *Pattern Recogn. Lett.*, **11** (1990), 589–594.

[10] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, John Wiley & Sons, 1999.

[11] R. J. Hathaway and J. C. Bezdek, “Switching Regression Models and Fuzzy Clustering,” *IEEE Trans. on Fuzzy Systems*, **1**-3 (1993), 195–204.

[12] J. Z. Huang, M. K. Ng, H. Rong and Z. Li, “Automated Variable Weighting in *k*-means Type Clustering,” *IEEE Trans. Pattern Anal. Machine Intell.*, **27**-5 (2005), 657–668.

[13] C.-H. Oh, K. Honda and H. Ichihashi, “Fuzzy Clustering for Categorical Multivariate Data,” in *Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, 2154–2159, 2001.

[14] K. Kummamuru, A. Dhawale and R. Krishnapuram, “Fuzzy Co-clustering of Documents and Keywords,” in *Proc. of 12th IEEE international conference on fuzzy systems*, 772–777, 2003.

[15] C. Ding and X. He, “K-means Clustering via Principal Component Analysis,” in *Proc. Int. Conf. Mach. Learning*, 225–232, 2004.

[16] E. Oja, A. Ilin, J. Luttinen and Z. Yang, “Linear Expansions with Nonlinear Cost Functions: Modeling, Representation, and Partitioning,” in *2010 IEEE World Congress on Computational Intelligence, Plenary and Invited Lectures*, 105–123, 2010.

[17] 本多克宏，市橋秀友，野津亮，「逐次学習を伴う線形ファジィクラスタリングによる適応的な協調フィルタリング」，『システム制御情報学会論文誌』，**20**-7 (2007), 283–291.

[18] D. Cartwright and F. Harary, “Structural Balance, A Generalization of Heider’s Theory,” *Psychological Review*, **63** (1956), 167–293.

[19] O. Katai and S. Iwai, “On the Characterization of Balancing Processes of Social Systems and the Derivation of the Minimal Balancing Processes,” *IEEE Trans. Systems, Man, and Cybernetics*, **SMC-8**, 5 (1978), 337–348.

[20] K. Tsuda, M. Minoh and K. Ikeda, “Extracting Straight Lines by Sequential Fuzzy Clustering,” *Pattern Recognition Letters*, **17** (1996), 643–649.

[21] 本多克宏，市橋秀友，野津亮，「逐次的クラスタ抽出に基づく協調フィルタリングの改良型アルゴリズム」，『システム制御情報学会論文誌』，**23**-12 (2010), 288–290.

[22] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gardon and J. Riedl, “Grouplens: Applying Collaborative Filtering to Usenet News,” *Communications of the ACM*, **40**-3 (1997), 77–87.

[23] G. Linden, B. Smith and J. York, “Amazon.com Recommendations: Item-to-Item Collaborative Filtering,” *IEEE Internet Computing*, **Jan-Feb** (2003), 76–80.

[24] J. A. Swets, “Measuring the Accuracy of Diagnostic Systems,” *Science*, **240**-4857 (1988), 1285–1289.

