論文解説：貪欲かつ楽観的なクラスタリングと天文データへの応用

奥野彰文 *[1,2,3], 服部公平 [4,3,5]

[1] 統計数理研究所 統計基盤数理研究系　[2] 理化学研究所 AIP センター
[3] 統計数理研究所 統計思考院　[4] 国立天文台 研究力強化戦略室　[5] ミシガン大学 天文学部

要旨

本稿は The Astrophysical Journal および Annals of the Institute of Statistical Mathematics に採択された我々の原著論文: Hattori et al. (2023) と Okuno and Hattori (2025) に関する解説です．解説の平易さを優先するため，より厳密な記述については原著論文をご参照ください．

キーワード: クラスタリング，不確実性，楽観的

1　研究背景

1.1　クラスタリング

ある $n \in \mathbb{R}$ 個のオブジェクト $i = 1, 2, \ldots, n$ があり，その属性を記述した $d \in \mathbb{R}$ 次元の変量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ があるとしましょう．このとき，観測された変量からオブジェクトの排反なグループ (クラスター) $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K \subset \{1, 2, \ldots, n\}$ を定める手続きをクラスタリングと呼びます．クラスタリングの方法は色々提案されていますが，基本的には変量に関する距離 $D(x_i, x_j)$ を何か適当に与えて，同一クラスタ内に属するオブジェクトの変量は近く，別のクラスタに属するオブジェクトの変量は遠くなるように割り当てを決定します．例えば K-means クラスタリングでは

$$\sum_{i=1}^{n} \|x_i - \mu_{c_i}\|_2^2$$

を最小化するようにオブジェクト i のクラスタへの割り当て $c_i \in \{1, 2, \ldots, K\}$ を定めます．ただし

$$\mu_k = \arg\min_{\mu \in \mathbb{R}^d} \sum_{i=1}^{N} \mathbb{1}(i \in \mathcal{C}_k) \|x_i - \mu\|_2^2$$

はクラスタ k の中心点を表しています．

1.2　変量の不確実性

本研究では，オブジェクト i が一点 $x_i \in \mathbb{R}^d$ に値を取るのではなく，特定の領域 $\mathcal{X}_i \subset \mathbb{R}^d$ のいずれかに値を取るであろうことが期待できる場合のクラスタリングを考えます．ここで $\mathcal{X}_i$ を不確実性集合 (uncertainty set) と呼びます．通常のクラスタリングは $x_1, x_2, \ldots, x_n$ を利用しますが，本研究では代わりに $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ を利用している点で異なります．

例 1 (天体の起源)．$i \in \{1, 2, \ldots, n\}$ がそれぞれ観測された天体として，どの天体がどの銀河に起源を持つのか，クラスタリングによって分類する問題を考えましょう．ここで天体 i の変量 $z_i = (z_{i1}, z_{i2}, z_{i3}, z_{i4}, z_{i5}, z_{i6})$ のうち，$(z_{i1}, z_{i2}, z_{i3}) \in \mathbb{R}^3, (z_{i4}, z_{i5}, z_{i6}) \in \mathbb{R}^3$ がそれぞれ天体 i の 3 次元空間内での現在の位置と速度ベクトルを表すとします．z_i は時間変化しますが，ある非線形変換により 3 次元に圧縮した変量 $x_i = f(z_i)$ は時間不変な物理量として知られており，時間の影響を無視するために x_i を用いてクラスタリングを行います．これらの天体は地球から観測されているので，太陽系から遠ければ遠いほど観測の不確実性は増し，また観測回数に応じて不確実性は減ります．したがって各天体 i ごとに観測の不確実性集合 $\mathcal{X}_i$ が異なります．

例 1 にもあるように，応用的な場面では観測値そのもの (つまり z_i) ではなく何らかの変換 f により調整された変量 $x_i = f(z_i)$ を用いてクラスタリングする場合があります．したがって観測値 z_i に正規性が仮定できる場合であっても x_i に正規性は仮定できず，不確実性集合 $\mathcal{X}_i$ は楕円状に限らず様々な形状を取ります．例 1 における，実際の数値シミュレーションでの不確実性集合の例[a]を図 1 に示します．

2　提案法

各天体 $i = 1, 2, \ldots, n$ の変量はそれぞれ，不確実性集合 $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ のいずれかに値を取ることが期待されています．ではこの不確実性集合をどのように扱えば良いのでしょうか?

2.1　貪欲かつ楽観的なクラスタリング

本研究では，不確実性を楽観的に扱うことでクラスタリングを行います．より具体的には，“天体の変量

* 責任著者，okuno@ism.ac.jp

[a] 正確には不確実集合を離散点で示したものを示しています．細長い線の一つ一つが $\mathcal{X}_i$ を表しています．

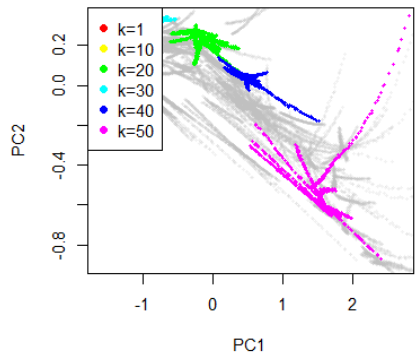

図 1: 不確実性集合 $\mathcal{X}_i$ (の離散点の主成分表示) の例

$(x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n =: \mathcal{A}$ が一番都合よく，凝縮したクラスタが得られる位置にあったと仮定してクラスタリングをするとどうなるか?" を考えます．提案法は貪欲に割り当てを更新することから，特に Greedy and Optimistic Clustering (GOC) と呼びます．

提案法 GOC

1. 初期値 $\Xi = (x_1, x_2, \ldots, x_n) \in \mathcal{A}$ を決める．例えば各 x_i を $\mathcal{X}_i$ の平均とする．
2. 利用したいクラスタリング法 (例えば K-means) を Ξ に適用し，クラスタの割り当て $c_1, c_2, \ldots, c_n \in \{1, 2, \ldots, K\}$ を決める．
3. 各クラスタの中心 μ_k を計算し，$\min_k \|x_i - \mu_k\|$ が最も小さくなるよう $x_i \in \mathcal{X}_i$ を更新する．
4. 上述の 2 と 3 を収束するまで繰り返す．

2.2 シミュレーションデータでの実験結果

ではここで，提案法を実験してみましょう．物理シミュレーションを介して天体の位置・速度ベクトルを計算し，そこから不確実性を伴って観測されたデータを用いて天体のクラスタリングをします．比較のため各 $\mathcal{X}_i$ の平均値に対して既存の K-means を適用した結果が図 2 に，提案法を適用した結果が図 3 にあります．図中の丸は各クラスタの大きさを表しています．図 2–3 を見れば明らかなように，提案法では各クラスタがより凝縮し，各クラスタの大きさが小さくなることが分かります．

前節での実験はシミュレーションデータを利用していますから，実際にどの天体が同じ銀河に起源をもつのか，背後にある真のクラスタの情報が利用できます．そこで Normalized mutual information や F-measure などといった指標を用いて提案法を評価してみると，いずれのハイパーパラメータの設定においても既存法よりスコアが高くなることが分かりました．より詳細な実験結果については Okuno and Hattori (2025) をご覧ください[b]．

[b] 実験に利用したデータやスクリプトが https://github.com/oknakfm/GOC に公開されています．

2.3 補足: 頻出の質問について

提案法 GOC に関する頻出の質問をまとめます．

Q1: **提案法は収束するのか?**

提案法は貪欲な最適化を行いますので，理論的には収束が示せません．一方で，$n = 275$ 天体程度で実験をしてみる 2–3 回程度の反復でほぼ収束して，10 回程度の反復で完全に収束します．

Q2: **計算量はどのくらいなのか?**

提案法のステップ 3 は並列計算可能ですので，ステップ 2 の計算量が支配的です．したがって (利用するクラスタリング法の計算量)×(反復回数) が計算量となります．天体の数 n が大きくなると計算量が増大しますが，軽量なクラスタリング法を利用すれば現実的な時間で計算可能です．現在大規模データへの適用に関する後続研究を執筆中です．

Q3: **$x_i \in \mathcal{X}_i$ の更新にペナルティは必要ないのか?**

$\mathcal{X}_i$ の中心に近いほど小さく，端に近いほど大きくなるペナルティを課して提案法を計算してみると，ペナルティ項は "わずかに入れると結果がよくなるが，大きすぎると悪くなる" ことが実験により分かっています．

Q4: **楽観的に更新すると，なぜ結果がよくなるのか?**

理論的にはわかっていません．クラスタリングではありませんが，変数誤差モデルによる回帰の最尤推定では，提案法 GOC と似た楽観的な変量更新を介した最適化問題が自然に導びかれます．同様に提案法 GOC が何らかの確率モデルの最尤法と対応付いている可能性を考えています．

3 実際の天文データへの応用

ではここで，実際の天文データのクラスタリングを考えてみましょう．まず Roederer et al. (2018) では観測された特殊な性質を持つ 83 天体のうち，不確実性の大きな 48 天体を取り除いた 35 天体についてクラスタリングを行いました．一方でこの解析では取り除かれた 48 天体のクラスタが分かりません．また 2018 年以降で 78 天体が追加で観測されたので，様々な不確実性を持った 161 天体が観測されています．

そこで，我々の論文 (Hattori et al., 2023) では提案法をこれら 161 天体の不確実性集合に適用しました．すると 161 天体のクラスタリング結果のうち，Roederer et al. (2018) で考慮された 35 天体は Roederer et al. (2018) による結果とほぼ合致し，また残りの 126 天体についても新たなクラスタを割り当てることができまし

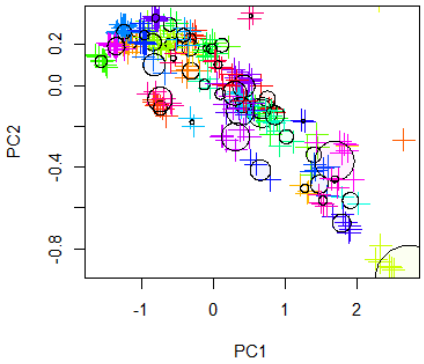

図 2: 既存法 (K-means)

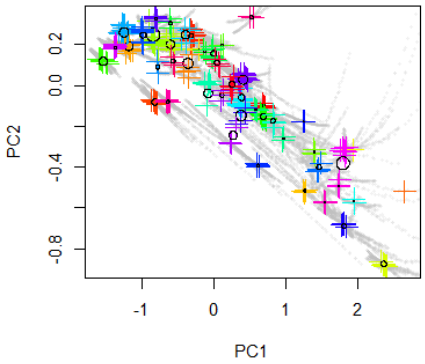

図 3: 提案法 (GOC + K-means)

た．得られたクラスタリング結果を各天体の化学組成に関するデータと照らし合わせてみると，天文学的に見て尤もらしい結果となりました．

4　まとめと今後の展望

本研究では変量 x_i の値が不確実であり，不確実性集合 $\mathcal{X}_i$ に値を取ることが期待できる場合のクラスタリング法を提案しました．貪欲かつ楽観的なクラスタリング (Greedy and Optimistic; GOC) 法を提案し，天体の物理シミュレーションを介して提案法が既存法より高いスコアを示すことを確認しました．実際に，Roederer et al. (2018) ではあらかじめ削除していた 48 天体を含む 161 天体をクラスタリングし，天文学的に見ても尤もらしい結果が得られました．

今後の課題は色々と残っていますが，大きく (1) より大規模なデータへ適用をすること，(2) 何故スコアが上がるのか理論的な解明をすること，の 2 点があります．(1) については軽量なクラスタリング法を利用し，代表点の更新を各反復では荒く行うなどして計算を高速化する研究が進行中です．(2) についても，なぜ楽観的な最適化が統計的推定を改善しうるのか，一般的な枠組みでの理論構築を進めています．

参考文献

Hattori, K., Okuno, A., and Roederer, I. U. (2023). Finding r-II sibling stars in the milky way with the greedy optimistic clustering algorithm. *The Astrophysical Journal*, 946(1):48.

Okuno, A. and Hattori, K. (2025). A greedy and optimistic clustering for leveraging individual covariate uncertainty. *Annals of the Institute of Statistical Mathematics*. accepted.

Roederer, I. U., Hattori, K., and Valluri, M. (2018). Kinematics of highly r-process-enhanced field stars: Evidence for an accretion origin and detection of several groups from disrupted satellites. *The Astronomical Journal*, 156(4):179.