スパースなデータを用いた階層的クラスタリング に対するSinkhorn距離に基づく改良

Improvement of hierarchical clustering for sparse data by using Sinkhorn distance

経営システム工学専攻　須賀原 颯紀

1. 序論

階層的クラスタリングとは，サンプル間距離計算とグルーピングを繰り返し，全サンプルの類似度を階層的に評価する手法である．本研究では，非ゼロ要素数が3%以下のスパースなデータ行列に対する階層的クラスタリングのサンプル間距離計算に改良を行う．サンプル間距離計算はpノルムがよく用いられる．これは，ベクトルの対応する要素同士の差の総和で距離を決める．スパースなデータ行列で同様の計算をする場合，ベクトルの対応する要素が共に非ゼロであることは少ない．特にサンプルが列和1の比率ベクトルであるとき，$p=1$ノルムでは両ベクトルの列和である2がサンプル間距離として多発する．このため，サンプルを有効にグルーピングすることが困難となる．

本研究では，Cuturi [1] によるSinkhorn距離を用いて，ベクトルの非ゼロ要素のうち，類似度の高い属性の要素同士の差の総和でサンプル間距離計算を行うことを提案する．このため，pノルムとsinkhorn距離による実データの階層的クラスタリングを通して性能を比較する．さらに，pノルムに類似する，属性同士の類似度を考慮できる一般化Euclid距離（GED）についても比較する．これにより，Sinkhorn距離を用いたスパースなデータ行列に対する階層的クラスタリングで，差異が明確なグルーピングが得られることを提示する．

2. サンプル間距離のSinkhorn距離活用

列和1のサンプルベクトル$\boldsymbol{r},\boldsymbol{s}\in\mathbb{R}_+^D$間の距離計算では，式 (2.1) の$p$ノルムが多用される．なお，$\boldsymbol{r},\boldsymbol{s}$の要素を$r_i,s_i(i=1,...,D)$とする．

$$dis_p(\boldsymbol{r},\boldsymbol{s})=\left(\sum_{i=1}^{D}|r_i-s_i|^p\right)^{1/p} \tag{2.1}$$

pノルムはベクトル間の対応する要素同士の差の総和で距離を決めている．$p=1$ノルムの場合，要素和1のスパースなベクトルではベクトル間で非ゼロとなる要素が一致しないとき，図2.1のように両ベクトルの要素和2がサンプル間距離となる．

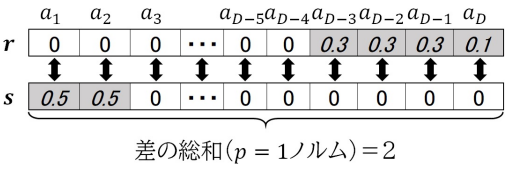

図 2.1: スパースな要素和1の比率ベクトルの例

そこで，非ゼロ要素のうち，類似度の高い属性同士の要素の差の総和で距離計算を行うEMD（Earth Mover's Distance）を考える．EMDは属性同士の類似度をコスト行列C，ベクトルを需要量・供給量と捉えた輸送問題として知られる線形計画問題の最適値である．EMDは定式化 (2.2) のように定義され，コスト行列Cの(i,j)要素$c_{i,j}$が$i=j\Leftrightarrow c_{i,j}=0$と三角不等式を満たすとき，線形計画問題の最適値が距離の公理を満たす．なお，$\langle\bullet,\bullet\rangle$は対応する行列要素の積の総和を表す．

$$dis_{EMD}(\boldsymbol{r},\boldsymbol{s}):=\min_{P\in U(\boldsymbol{r},\boldsymbol{s})}\langle P,C\rangle \tag{2.2}$$

$$U(\boldsymbol{r},\boldsymbol{s}):=\{P\in\mathbb{R}_+^{D\times D}|P\mathbf{1}=\boldsymbol{r},P^\top\mathbf{1}=\boldsymbol{s}\}$$

ここで，定式化 (2.2) の実行可能解Pは確率分布となる．EMDは$O(D^3\log(D))$の時間計算量が掛かり，Pのエントロピーで正則化した定式化 (2.3) がEMDを高速に近似するSinkhorn距離である．

$$dis_{S(\lambda)}(\boldsymbol{r},\boldsymbol{s}):=\langle P_{S(\lambda)},C\rangle \tag{2.3}$$

$$P_{S(\lambda)}=\underset{P\in U(\boldsymbol{r},\boldsymbol{s})}{\operatorname{argmin}}\left\{\langle P,C\rangle-\frac{1}{\lambda}h(P)\right\}$$

$\lambda>0$は大きい値ほど最適解$P^*_{S(\lambda)}$がEMDの最適解P^*に近づく．$\lambda\to\infty$で$P^*_{S(\lambda)}$はP^*に収束するが，$e^{-\lambda c_{i,j}}$の計算限界までしか大きくできない．

3. Sinkhorn距離の高速化性能検証

Sinkhorn距離とEMDの計算時間比較を行う．EMDと$\lambda=1,10,50$のときのSinkhorn距離を計算する．属性数$D=32,64,128,256$ごとに100回ずつ行い，平均計算時間を図3.1にプロットする．

比較に用いるベクトルはD次元ベクトルを列和が1になるように乱数で生成する．さらに，コス

ト行列 C の要素として，ノード数 D のネットワークを考え，ノード間の最短ステップ数を各ノードに対応するベクトルの属性同士の類似度とする．ノード間のエッジは確率 $\frac{1}{2}$ でランダムに設定し，各エッジの重みは 0 と 1 の間で均一に分布させる．なお，計算には Intel(R) Core(TM) i5-4310U CPU @ 2.00GHz 2.59GHz, 実装 RAM4GB を用いる．

λ の増加に対し EMD と Sinkhorn 距離の平均計算時間の差が広がる．また，Sinkhorn 距離は λ が大きくなるほど，平均計算時間も大きくなる．

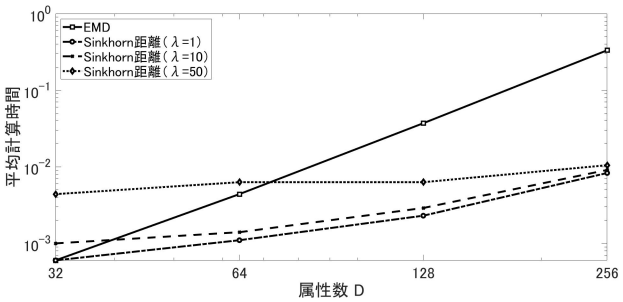

図 3.1: 属性数と平均計算計算時間（両対数グラフ）

4. 活用検証

4.1. 検証方法

Sinkhorn 距離，1 ノルム，一般化 Euclid 距離でサンプル間距離計算を行った場合の階層的クラスタリングを比較する．比較は,「1 ノルム =2」となる（以下，無効距離と呼ぶ.）サンプルペアの割合と図 4.1 のように Sinkhorn 距離や一般化 Euclid 距離で計算するときの，クラスター変化率をみる．

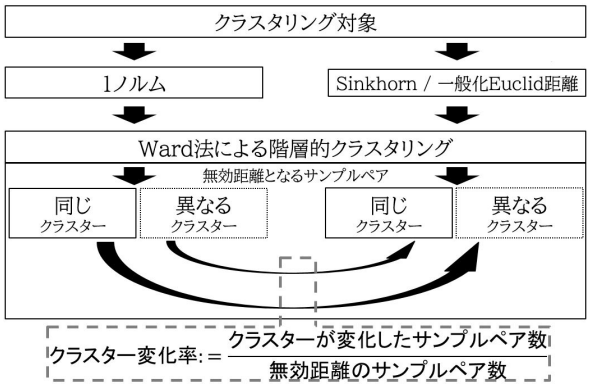

図 4.1: 階層的クラスタリング比較手法

また，シルエット分析も行う．サンプルごとに式 (4.1) を計算し，全サンプルでの平均値（$\bar{e}$）が大きいクラスタリングを優良とする指標である．

$$e_n = \frac{dis_{ot,n} - dis_{ac,n}}{\max\{dis_{ac,n}, dis_{ot,n}\}} \tag{4.1}$$

$dis_{ac,n}$ はサンプル $\boldsymbol{s}_n$ が含まれるクラスターのほかのサンプルまでの平均距離であり，凝集度を示す．また，$dis_{ot,n}$ はサンプル $\boldsymbol{s}_n$ が含まれるクラスターに最も近い別クラスターに含まれるサンプルとの平均距離であり，乖離度を示す．この指標では,「クラスター内のサンプル同士は距離が近く密で，クラスター同士は距離があり疎である状態」を良いクラスタリングとしている．

$\boldsymbol{r}, \boldsymbol{s}$ の一般化 Euclid 距離は正定値対称行列 Σ に対し，式 (4.2) のように定義される．

$$dis_{GED}(\boldsymbol{r}, \boldsymbol{s}) = \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{D} \sigma_{i,j}(r_i - s_i)(r_j - s_j)} \tag{4.2}$$

$\sigma_{i,j}$ は属性間の相互関係をまとめた計量行列 $\Sigma \in \mathbb{R}^{D\times D}$ の要素である．一般化 Euclid 距離は属性同士の類似度を $\sigma_{i,j}$ により考慮できる．しかし，p ノルム同様にベクトルの対応する属性の要素同士の差のみを計算している．

本研究では，サンプル間距離計算に定式化 (2.3) を用いて，スパースなデータ行列に対する階層的クラスタリングを行うことを提案する．[1] で提示される Sinkhorn アルゴリズムを参考にし，$\lambda = 50$ でサンプル間距離計算を行う．そして Ward 法にて階層的クラスタリングを行い，最大連結距離の 70%でクラスターを分割する．活用検証は複数のスパースなデータ行列に対して行うが，抄録ではそのうちの 1 つについて記述する．

4.2. 犯罪経歴保持者クラスタリング

犯罪傾向に基づくクラスタリング

犯罪傾向に基づく犯罪経歴保持者のクラスタリングを行う．これは，犯罪者プロファイリングのための犯人像推定に資する分析である．これに関しては，Taylor 他 [2] の 2 ノルムによる Ward 法を用いた連続殺人犯のクラスタリングなど各国で盛んに研究されている．先行研究を踏まえ，個人の犯罪趣向と逮捕地環境の特徴から犯罪経歴保持者の犯罪傾向を捉えた階層的クラスタリングを行う．

今回，麻薬犯罪，銃器犯罪，家庭内犯罪，傷害事件の各経歴数と逮捕地 77 区分のデータを含む SSL (Strategic Subject List) を活用する．SSL にはシカゴ市 114,807 人分の上記データが含まれる．ここでは，重複がない $N = 4189$ 人分のサンプルで階層的クラスタリングの比較実験を行う．さらに，Sinkhorn 距離を用いたクラスタリングについ

て，重複の対応に基づき 114,807 人分に拡張した結果の考察の概略を記述する．

サンプルベクトルとコスト行列

サンプルベクトルは図 4.2 の形式とする．

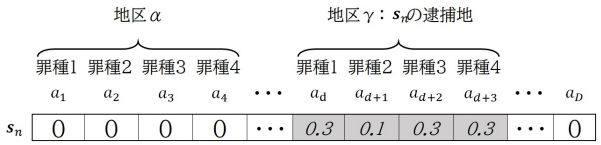

図 4.2: サンプルベクトルの例

犯罪趣向として 4 罪種の犯罪経歴の比率をベクトル要素とする．4 属性を逮捕地区分ごとに設け，$\boldsymbol{s}_n$ の逮捕地区分に対応する 4 属性に 4 罪種の犯罪経歴の比率を入れる．サンプルサイズ $N = 114807$，属性数は 4 罪種 $\times$ 77 区分 で $D = 308$ となる．

また，コスト行列 C は時空間的条件の罪種間類似度から作成する．このため，308 属性単位で集計した各現場状況・発生時間帯の犯罪発生件数ベクトル $\boldsymbol{\xi} \in \mathbb{R}_+^D$ を作成する．そして，ベクトル同士の 2 ノルムの 2 乗をコスト行列 C の要素とする．

計量行列 $\Sigma \in \mathbb{R}^{D\times D}$ は 2 種類作成し，それぞれの (i, j) 要素を $\sigma_{i,j} = \sigma_{i,j}^{\exp}$，$\sigma_{i,j}^{\mathrm{cov}}$ とする．

式 (4.3) はコスト行列 C の要素 $c_{i,j}$ の指数関数に基づく変換による要素である．今回，ハイパーパラメータは $\eta = 1$ とする．

$$\sigma_{i,j} = \sigma_{i,j}^{\exp} = \exp(-\eta c_{i,j}) \tag{4.3}$$

式 (4.4) は $\boldsymbol{\xi}$ の属性間の分散共分散行列に基づく変換による要素である．なお，$\overline{\boldsymbol{\xi}} \in \mathbb{R}_+^D$ は N 個の $\boldsymbol{\xi}$ の属性ごとの平均値によるベクトルとする．式 (4.3) と式 (4.4) の変換から，計量行列は正定値対称行列となる．

$$\sigma_{i,j} = \sigma_{i,j}^{\mathrm{cov}} = \begin{cases} g_{i,j} & (i \neq j) \\ g_{i,j} + 10^{-9} & (i = j) \end{cases} \tag{4.4}$$
$$g_{i,j} = \|\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\|_2^2 + \|\boldsymbol{\xi}_j - \overline{\boldsymbol{\xi}}\|_2^2 - c_{i,j}$$

Sinkhorn 距離の活用

表 4.1 は距離計算 4 手法の比較結果である．

クラスター変化率と平均シルエット値は Sinkhorn 距離を用いた場合の階層的クラスタリングが 1 番高いと分かる．また，非ゼロ要素同士の差を前提としているので，79.43%のサンプルペアが無効距離から適切な距離に改善された．一般化 Euclid 距離を用いた場合については属性の組合せに $\sigma_{i,j}$ を設けたため，サンプル間距離に $\sigma_{i,j}$ 分のばらつきが見られた．このため，クラスター変化率は 34.96%と 42.58%であった．しかし，シルエット分析での良いクラスタリングの基準を満たすほどの改善は Sinkhorn 距離を用いた場合と比べて高くならなかった．

表 4.1: 4 つの距離手法の比較

	無効距離ペア割合 (%)	クラスター変化率 (%)	$\overline{e}$
p(1 ノルム)	79.43	-	0.19
Sinkhorn	0.00	43.91	0.65
GED $(\sigma_{i,j}^{\exp})$	-	34.96	0.38
GED $(\sigma_{i,j}^{\mathrm{cov}})$	-	42.58	0.16

図 4.3〜図 4.6 は距離計算 4 手法を用いた階層的クラスタリングそれぞれのデンドログラムである．破線はクラスターが分割されたことを示している．

Sinkhorn 距離を用いた場合以外ではクラスターに含まれるサンプルの数が極端に偏っている．グルーピングという本来の目的を考えると，結果の価値は低い．Sinkhorn 距離の活用は属性同士の類似度を考慮する上で有効に機能し，クラスター間の差異を明確にする結果となった．

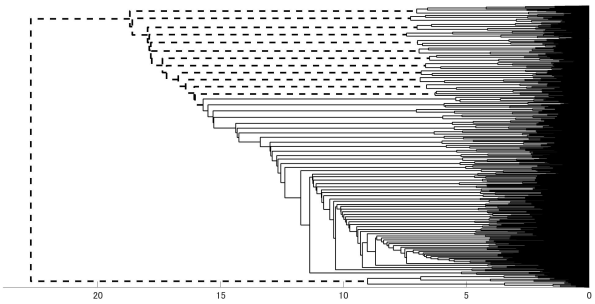

図 4.3: 1 ノルムによるデンドログラム

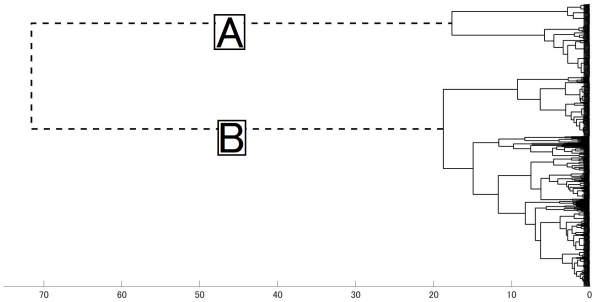

図 4.4: Sinkhorn 距離によるデンドログラム

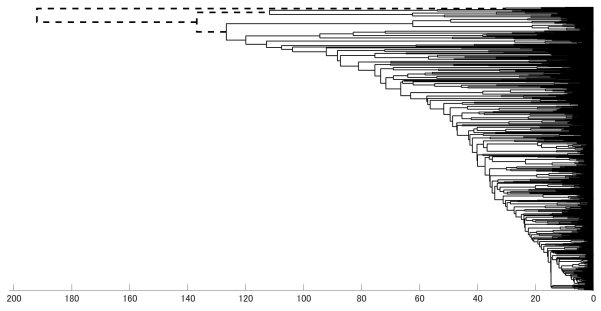

図 4.5: 一般化 Euclid 距離 ($\sigma_{i,j} = \sigma_{i,j}^{\exp}$) によるデンドログラム

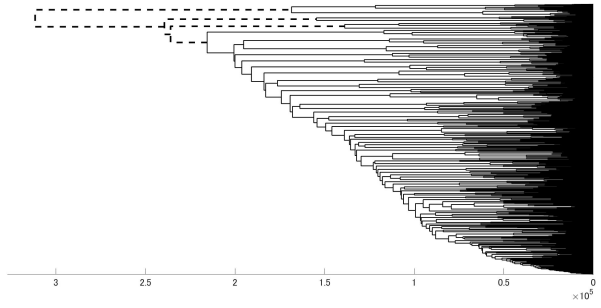

図 4.6: 一般化 Euclid 距離（$\sigma_{i,j} = \sigma_{i,j}^{\mathrm{cov}}$) によるデンドログラム

図 4.7 は Sinkhorn 距離による階層的クラスタリングで，クラスター A に分類された人物のシカゴ市 77 区分ごとの人口密度分布である．

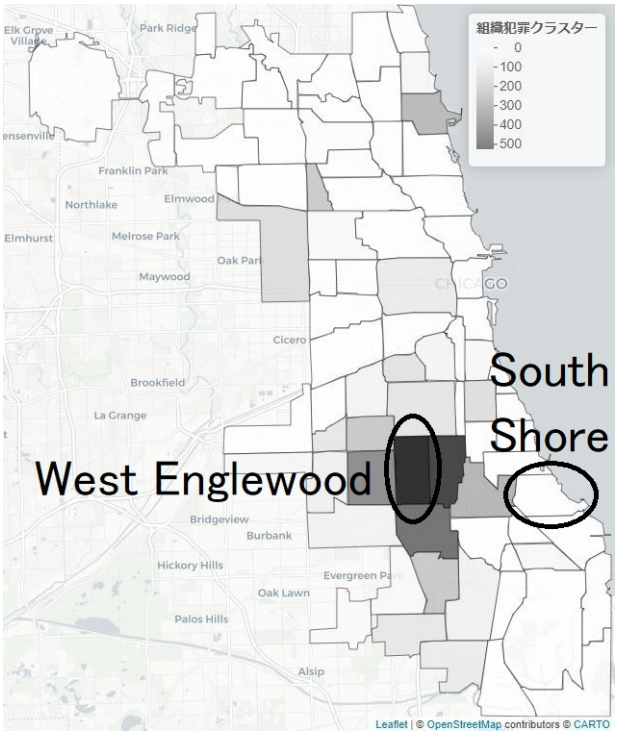

図 4.7: クラスター A の人口密度分布

ここで，人口密度の高密度地区は West Englewood などギャングの縄張りが密集する地区と重なっていた．一方，South Shore など住宅街の多い地区では人口密度が低くなっていた．高密度地区はギャングと関連の強い銃器犯罪の多発地区とも重なる．よって，麻薬や銃器の取引などに関わるギャングとの繋がりが深い犯罪傾向にあるクラスターの可能性が高い．違法行為をより詳細に調べることで，犯罪傾向の把握がより明瞭なものとなると考えられる．クラスター B では別の犯罪傾向をくみ取ることができ，人口密度の分布も大きく異なるものとなった．Sinkhorn 距離を階層的クラスタリングに活用したことで，クラスター間の差異が明確になり傾向を捉えやすくなった．

5. 結論

本研究では，属性同士の類似度を考慮した，Sinkhorn 距離による階層的クラスタリングの改良を提案した．スパースなデータ行列の場合，p ノルムで有効な距離が計算されないことがある点に対し，Sinkhorn 距離の活用が改善に繋がった．これにより，クラスター間の差異を明確にする結果が得られた．また，属性同士の類似度を考慮できる一般化 Euclid 距離を用いた場合との比較でも，Sinkhorn 距離を用いた場合の効果が確認できた．

実用に向けた更なる課題としては，より一層の計算高速化が挙げられる．Sinkhorn アルゴリズムの高速化は近年盛んに研究されており，例えば Lin 他 [3] による高速化アルゴリズムの活用は今後の課題として挙げられる．

参考文献

[1] Marco Cuturi, (2013), “Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances,” Advances in Neural Information Processing Systems, 26, 2292–2300.

[2] Sadie Taylor, Marie Cahillane, Lance Workman, (2017), “Adopting the bottom-up approach and cluster analysis on North American and European male serial killers,” Journal of Forensic Research and Analysis, 1, 1–11.

[3] Tianyi Lin, Nhat Ho, Michael I. Jordan, (2020), “On the Efficiency of the Sinkhorn and Greenkhorn Algorithms and Their Acceleration for Optimal Transport”.