1F3-05

ラフクラスタリングによる医療データの類型化の試み

Rough Clustering and Its Application to Time-series Medical Data Analysis

平野章二　　津本周作
Shoji Hirano　Shusaku Tsumoto

島根大学医学部医学科医療情報学講座
Department of Medical Informatics, Shimane University, School of Medicine

In this paper, we present an indiscernibility-based clustering method called rough clustering, which clusters objects according to their relative proximity. Experimental results on the artificially created numerical datasets demonstrated that this method could produce good clusters even when the proximity of the objects did satisfy the triangular inequality. Clustering results on chronic hepatitis dataset also demonstrate that this method could absorb local disturbance in the proximity matrix and produce interpretable clusters containing time series that have similar patterns.

1. はじめに

クラスタリングは，分類対象の性質に応じて適当な類似度（若しくは相違度）と群化の手続きを定め，類似した対象を同一の群としてまとめる処理であり，これまでに k-means，EM algorithm，BIRCH など様々な方法が提案されている [1]。数値データを対象としたクラスタリング法では一般に，個々の対象の属性値から対象間類似度を導出するとともに，重心や分散など群としての性質を示す代表的な特徴量を導出し，クラスタの内的均一性を最大化しつつクラスタの相互分離性を最大化するように基準を定めて最適分割を決定していく。しかしながら，個々の対象の属性値を直接参照できず，対象間の関係から相対的な類似度のみが与えられる場合，まとまりを評価する尺度の定義が難しく，適切なクラスタを生成することは容易ではない。また，選好データのように類似度が主観的評価から生成される場合，それが三角不等式を満足することは保証されず，重心等の幾何的特徴量の利用が制限される。古典的な階層的クラスタリングは相対的相違度を扱うことができるが，方法の特性として処理順序に対する結果の不定性，空間縮退や拡張等の問題が知られている [2]。

本稿では，(1) 類似度行列のみが与えられる場合，(2) 類似度が相対的である場合，において可読性の高いクラスタを生成する方法として，対象間の識別不能度に基づくラフクラスタリングを提案すると共に，このような性質をもつデータの例として医療データをとりあげ，その類型化を試みた結果を報告する。

2. ラフクラスタリング

ラフクラスタリングは，相対的類似性に基づく局所的分類を各対象において個別に実施し，その分類の共通性から対象間の大局的類似性を判断してクラスタリングを行う手法である [3]。その手続きは，(1) 初期同値関係の構築，(2) 同値関係の再帰的更新，の 2 ステップから構成される。

2.1 初期同値関係

まず，N 個の対象の各々について，全体集合を「同じもの」と「異なるもの」の 2 つのカテゴリへ 2 値分類する同値関係（初期同値関係）を定める。対象の全体集合を $U = \{x_1, x_2, ..., x_N\}$ とするとき，対象 x_i に対する初期同値関係 R_i は次式により定義される。

$$U/R_i = \{P_i,\ U - P_i\}, \tag{1}$$

$$P_i = \{x_j |\ d(x_i, x_j) \leq Th_{d_i}\},\ \ \forall x_j \in U. \tag{2}$$

ここで，$d(x_i, x_j)$ は対象 x_i と x_j の相違度を示し，Th_{d_i} は x_j を x_i と識別不能とみなす相違度の上限閾値を示す。同値関係 R_i は，U を 2 つのカテゴリ P_i と $U - P_i$ に類別する。P_i は x_i と類似した対象の集合であり，$U - P_i$ は相違した対象の集合である。相違度 $d(x_i, x_j)$ が上限閾値 Th_{d_i} 以下のとき，x_j は x_i と識別不能とみなされる。

相違度 $d(x_i, x_j)$ の定義は基本的に任意であるが，ここでは多重スケールマッチングにより出力される系列間相違度を用いる。また，同値類とみなす相違度の上限閾値 Th_{d_i} については，対象の密度から自動的に定める方法等を文献 [3] において紹介しているが，初期同値関係 R_i の役割は U の 2 値分類であり，それを実現する他の分類方法を選択しても良い。

2.2 識別不能度

初期同値関係の決定後，任意の 2 つの対象組を取り上げ，それらを同一カテゴリに類別する同値関係の数を数え上げる。この数の N に対する比を識別不能度と呼び，ある 2 つの対象が，N 個の同値関係の中で識別不能とみなされる割合を表現する。任意の対象 x_i と x_j について，識別不能度 $\gamma(x_i, x_j)$ は次式のとおり定義される。

$$\gamma(x_i, x_j) = \frac{\sum_{k=1}^{|U|} \delta_k^{indis}(x_i, x_j)}{\sum_{k=1}^{|U|} \delta_k^{indis}(x_i, x_j) + \sum_{k=1}^{|U|} \delta_k^{dis}(x_i, x_j)}, \tag{3}$$

ここで，

$$\delta_k^{indis}(x_i, x_j) = \begin{cases} 1, & \text{if } (x_i \in [x_k]_{R_k} \wedge x_j \in [x_k]_{R_k}) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\delta_k^{dis}(x_i, x_j) = \begin{cases} 1, & \text{if } (x_i \in [x_k]_{R_k} \wedge x_j \notin [x_k]_{R_k}) \\ & \text{or } (x_i \notin [x_k]_{R_k} \wedge x_j \in [x_k]_{R_k}) \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

式 4 は，同値関係 R_k の下で対象 x_i と x_j が識別不能なとき $\delta_k^{indis}(x_i, x_j)$ が 1 となることを示している。条件として，x_i 及び x_j がいずれも x_k と同一カテゴリに属さなければならな

連絡先: 島根大学医学部医療情報学講座 平野章二
〒 693-8501 島根県出雲市塩冶町 89-1
Phone:(0853)20-2173，E-mail:hirano@ieee.org

い。一方，式 5 は，同値関係 R_k の下で対象 x_i と x_j が識別可能なとき $\delta_k^{dis}(x_i, x_j)$ が 1 となることを示している。条件として，x_i あるいは x_j が排他的に x_k と同一カテゴリに属さなければならない。式 3 に示すように，すべての同値関係 $k(1 \leq k \leq |U|)$ について $\delta_k^{indis}(x_i, x_j)$ および $\delta_k^{dis}(x_i, x_j)$ を積算することで，対象 x_i と x_j を識別不能とみなす同値関係の割合を得る。

2.3 初期同値関係の再帰的更新

ここで，高い識別不能度をもつ，すなわち，多くの同値関係により同値類と類別される対象組に対して異なる類別を与える同値関係があるとき，その同値関係は詳細すぎる類別知識を与えると捉えることができる。そこで，予め識別不能度に対して閾値を設け，識別不能度がその閾値より高い対象組については，全ての同値関係が同一カテゴリに類別するように初期同値関係を更新する。$R_i \in \mathbf{R}$ を U における同値関係とするとき，R_i の修正後の同値関係 $R_i' \in \mathbf{R}'$ は次式により定義される。

$$U/R_i' = \{P_i',\ U - P_i'\}, \tag{6}$$

ここで，P_i' は

$$P_i' = \{x_j | \gamma(x_i, x_j) \geq T_h\}, \quad \forall x_j \in U. \tag{7}$$

により規定される対象集合である。また，T_h は x_i と x_j を識別不能とみなす識別不能度の下限閾値で，ユーザにより与えられる。もし $\gamma(x_i, x_j)$ が T_h よりも大きい場合，R_i は x_j を x_i と同一カテゴリに類別するように更新される。この更新処理を全ての対象組について分類が固定するまで再帰的に繰り返すことで，所与の識別不能度に応じた「粗い」クラスタを生成することが出来る。1 にラフクラスタリングの概念図を示す。

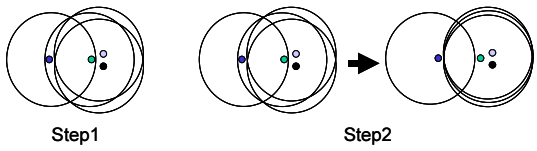

図 1: ラフクラスタリングの概念図

3. 実験結果

3.1 人工データの分類

相対的相違度の取り扱いにおける提案方法の有効性を検証するため，人工的に生成した数値データのクラスタリング実験を行った。最初に 2 次元正規分布に従うデータを生成し，ユークリッド距離に基づく相違度行列を生成した後，行列中の任意の要素をランダムに選択して 0 へ置換する。これにより，局所的に相違度が三角不等式を満足しない状態を意図的に作りだし，クラスタリング結果がどの程度影響を受けるかを調べた。

実験用データの生成手続きを以下に示す。まず，Neyman-Scott 法 [4] に基づき 2 次元データを生成する。ここでは簡単のためクラスタ数を 3 とし，各々のクラスタに約 100 点，計 310 点のデータを生成した。次に，各点間のユークリッド距離を計算し，310×310 要素からなる相違度行列を作成する。続いて，相違度行列から一定数の要素をランダムに選択し，それらの値を全て 0 へ置換する。置換する要素数は，全要素数の 10%, 20%, 30%, 40%, 50% の 5 種類とした。5 種類のそれぞれについて，十分なランダム性を持たせるために 10 回ずつ独立した選択・置換作業を行い，計 50 個の置換済み相違度行列を用意した。

上記操作により得られた相違度行列を入力とし，提案法によるクラスタリングを行った。初期同値関係の構築には文献 [3] の密度に基づく方法を用い，パラメータは予備実験により $\sigma = 15.0$ 及び $T_h = 0.3$ と定めた。また，比較対象として，同じく相違度行列を入力にとることができる階層的クラスタリング法（群間平均結合 AL-AHC 及び最遠隣結合 CL-AHC) [2] を採り上げた。AHC では，クラスタ数が 3 に到達した時点で併合を終えている。なお，ここではランダムに要素を置換した相違度行列を入力として実験を行うため，データの属性値を直接的に参照する他のクラスタリング法は採り上げていない。クラスタリング結果の妥当性は以下の尺度により評価した。

$$\text{Validity } v_{\mathbf{R}}(C) = \min\left(\frac{|X_{\mathbf{R}} \cap C|}{|X_{\mathbf{R}}|}, \frac{|X_{\mathbf{R}} \cap C|}{|C|}\right),$$

ここで，$X_{\mathbf{R}}$ は生成されたクラスタの集合を表し，C はデータ生成に用いたクラスタを示す。$v_{\mathbf{R}}(C)$ により，両者の重なり度合いを評価している。

表 1 に比較結果を示す。表の第 1 行目は相違度行列の要素のうち，置換されたものの比率を示す。例えば 30 は相違度行列の全要素のうち 30%が 0 へ置換されたことを示している。続く 3 つの行は，AL-AHC, CL-AHC 及び提案法によりそれぞれ生成したクラスタの妥当性 $v_{\mathbf{R}}(C)$ を示す。無置換（割合が 0）の場合を除き，妥当性は各割合に於いて 10 個の異なる置換済相違度行列から得られたものの平均 ± 標準偏差の形で表記した。

無置換の場合，相違度行列はユークリッド距離行列と完全に一致する。原データは比較的分離度の高い 2 次元正規分布に従う分布であるため，AL-AHC, CL-AHC のいずれも妥当性 0.99 を超える適正なクラスタを生成した。また，提案法においても妥当性 0.98 を超える適正なクラスタを生成した。しかしながら，相違度の置換が生じると，AL-AHC 及び CL-AHC の生成するクラスタの妥当性は例えば 10%の場合でそれぞれ 0.688，0.874 と大きく低下し，その度合いは置換割合の増加とともに更に大きくなる。局所的な置換の発生によりこのような妥当性の低下が生じる原因は以下のとおりである。ある 2 つの対象間の相違度を 0 に置換すると，相違度行列を距離行列ととらえた場合，対象空間に局所的な縮退が生じると考えることができる。ペアごと相違度を評価する結合方法では，初期段階で 2 対象間の相違度が最小となる組を探索するため，これらは他の対象との相違度の大小にかかわらず最初の結合対象となる。本来異なるクラスタに属するべき対象が初期段階に於いて結合された場合，特に AL-AHC 及び CL-AHC のように結合階層の反転を許さないクラスタリング法では，それらを再び分離することはできず，最終的に不適当に内的連結された図 2 あるいは図 3 の様なクラスタが生成される。

これに比べ，提案法では 50%まで置換が進んだ場合にも，クラスタリング結果の高い妥当性を維持することが出来ている。これは，本方法が局所的分類の結果を直接利用するのではなく，分類結果の全体を見渡して大局的な類似性を評価し分類を行うことによる。相違度が 0 に置換された 2 つの対象は，それらに割り当てられる初期同値関係において局所的には識別不能とみなされる。しかしながら，それ以外の大多数の対象は，これら 2 対象間の相違度ではなく，自分自身とこれらとの相違度に基づき同値関係を定めるため，その大多数に置換が生じない限り，もとの対象間相違度が適切に反映される。個々の同値関係により規定される分類の全体的な類似性を評価する識別

表 1: Comparison of the clustering results

Mutation Ratio[%]	0	10	20	30	40	50
AL-AHC	0.990	0.688 ± 0.011	0.670 ± 0.011	0.660 ± 0.011	0.633 ± 0.013	0.633 ± 0.018
CL-AHC	0.990	0.874 ± 0.076	0.792 ± 0.093	0.760 ± 0.095	0.707 ± 0.098	0.729 ± 0.082
Our method	0.981	0.980 ± 0.002	0.979 ± 0.003	0.980 ± 0.003	0.977 ± 0.003	0.966 ± 0.040

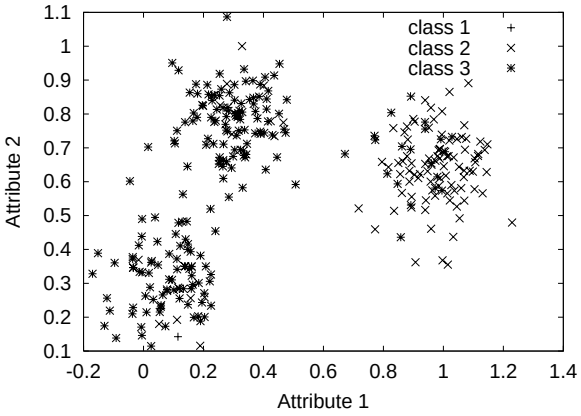

図 2: AL-AHC による分類結果。置換率 40%。

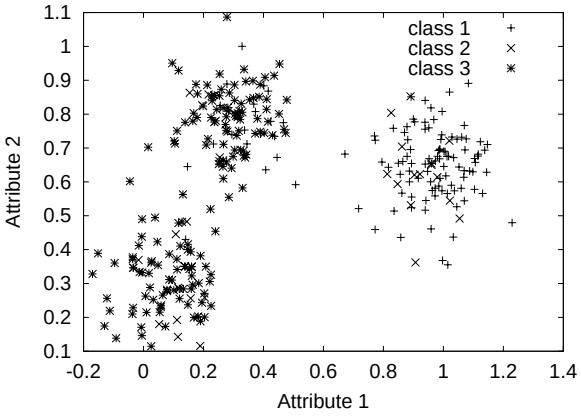

図 3: CL-AHC による分類結果。置換率 40%。

不能度は高くなり，これらを類別する方向に初期同値関係が更新されるため，結果として図 4 に示すとおり妥当性の高いクラスタを生成することができる。以上から，本方法が，局所的に三角不等式が満足されない相違度行列においても，対象間の大局的関係に基づき妥当な分類を実現できることが示された。

3.2 医療データの分類

続いて，慢性ウイルス性肝炎に関する検査記録をデータ化した慢性肝炎データセット [5] に提案法を適用し，分類実験を行った。同データは PKDD Discovery Challenge の共通データとして提供されているもので，771 名の B 型および C 型肝炎患者の数年から十数年にわたる臨床検査結果が時系列として記録されている。このような慢性疾患の検査データでは，長い期間の中で患者の病態に変化が生じ，それに応じて受診間隔が変化するためデータの収集間隔が変化すること，また，収集期間も転帰に応じて患者ごとに異なることから，不等間隔・不等系列長データを対象とした比較法の導入が要求される。ここ

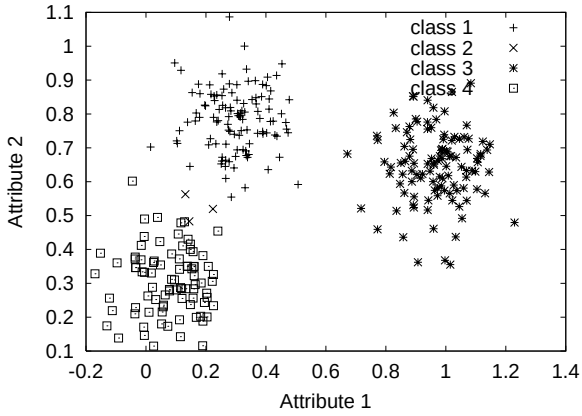

図 4: 提案法による分類結果。置換率 40%。同値関係の更新は 4 順目で収束。

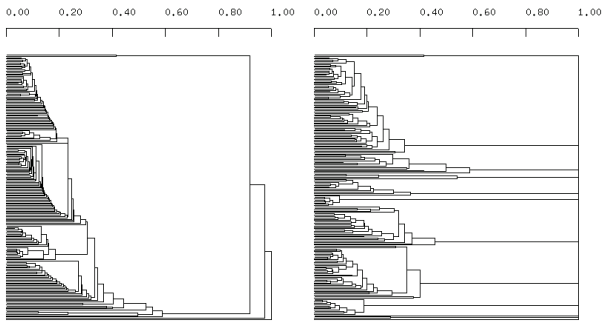

図 5: AHC による GPT 系列の樹状図. 左: AL-AHC. 右: CL-AHC.

では，時系列の多重スケール構造比較法 [6] を採用した。この方法は，時系列を平滑化度を変えて多重スケール表現し，部分系列の最適な対応関係を探索するもので，生成される相違度には，(1) 系列組ごとに異なる視野で比較を行うため，3 系列の相違度間で必ずしも三角不等の関係が満たされない，(2) 複数の形状パラメータに基づく比較であるため，相違度と原系列値との間に直接的な関係がない，(3) スケール数の制約等により最適対応を獲得できない場合，相違度に例外値を割り振る必要がある，等の特徴があり，相違度行列は前項の人工データの場合と類似した性質を有する。

図 5 に AHC を用いて C 型インターフェロン適用例の GPT 系列 196 例から生成された樹状図を示す。本実験では，上述した最適対応獲得失敗時の相違度例外値として，最適対応獲得に成功した系列の中での最大値を代入している。AL-AHC の樹状図では，非対応系列を含むクラスタ間の相違度が非常に大きくなるため，相違度の平均値を適切に算出することができず，適切な構造が得られていないことがわかる。一方，CL-AHC の樹状図では，一つ以上の相手系列において最適対応の獲得に

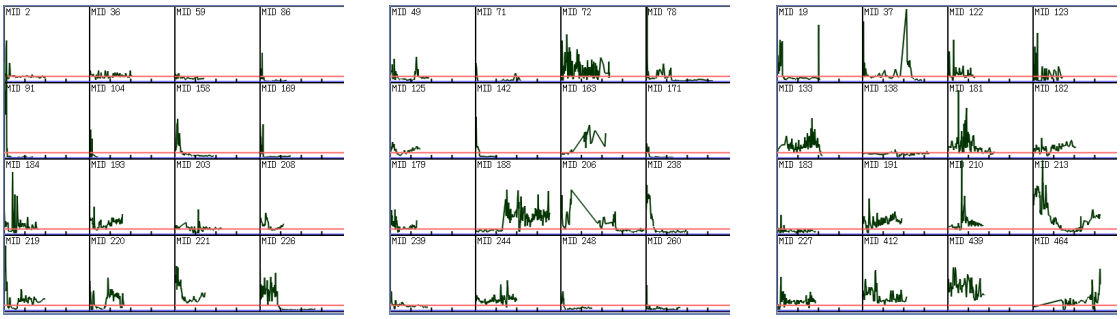

図 6: CL-AHC による分類結果。左から第 1 クラスタ (71 例), 第 2 クラスタ (39 例), 第 3 クラスタ (28 例)。いずれも ID 順に 16 例抜粋。

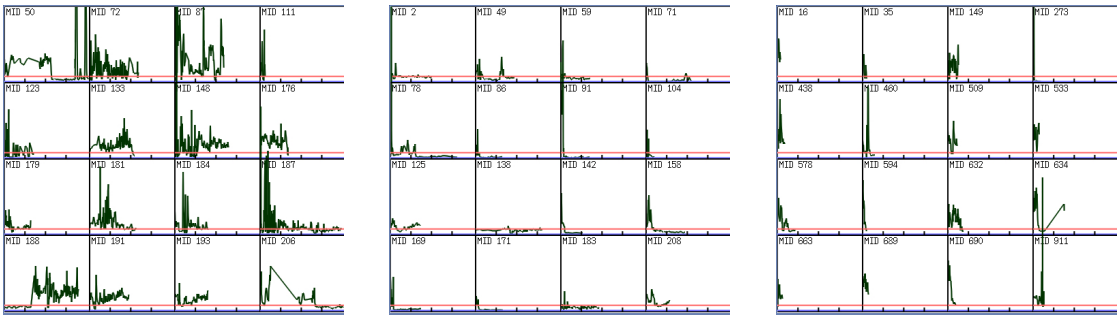

図 7: ラフクラスタリングによる分類結果。左から第 1 クラスタ (80 例)，第 2 クラスタ (60 例)，第 3 クラスタ (18 例)。いずれも ID 順に 16 例抜粋。

失敗した系列を含むクラスタが最後まで結合されず，それ以上の大局的な階層構造が得られないことがわかる。相違度が著増する（前段との差がその mean+1SD を超える）時点へ分割点を定めると，AL-AHC では 8 個のクラスタが生成されたが，大部分の系列 (182/196) が図 6 左上に示す単一のクラスタに分類され，興味ある特徴は見いだせなかった。一方，CL-AHC ではそれぞれ 71, 39, 28, . . . 例を含む計 16 個のクラスタが生成された。図 6 に最大のものから 3 番目までの各クラスタに類別された系列の例をそれぞれ示す。類似した系列が同一クラスタに見られるが，同時に明らかに異なる系列もそれらと同じクラスタに含まれている。

一方，ラフクラスタリングに基づく分類では，それぞれ 80, 60, 18, 6. . . 例からなる計 25 クラスタが生成された。図 7 に最大のものから 3 番目までの各クラスタに分類された系列の例を示す。第 1 クラスタには，GPT 値が持続的に乱高下するパターンが多数含まれることが確認できる。一方，第 2 クラスタには高値から低下した後に平坦化するパターンが多く見られる。また，第 3 クラスタには短期系列がまとめられている。前 2 者はウイルスの活動が長期に継続するパターン及び沈静化するパターンをそれぞれ示すと考えられ，局所的に特異な値をもつ相違度行列においても，他の分類を考慮して大局的な分類を行える本方法によって興味深い傾向が得られたと言える。

4. おわりに

本稿では，(1) 相違度行列のみが与えられ，データの属性値を直接参照できない場合，かつ (2) 相違度が相対的なものとして定義され，必ずしも三角不等の条件を満足しない場合，においても適切な類型化が可能なラフクラスタリングを提案した。これらの条件を満たす例として人工データ及び医療データを採り上げ，その分類実験を通じて局所的に三角不等式が満たされない場合においても提案法により妥当性の高いクラスタが生成されることを示した。今後の課題として，識別不能度の算出手続きにおける計算量の削減があげられる。

参考文献

[1] P. Berkhin (2002): Survey of Clustering Data Mining Techniques. Accrue Software Research Paper. URL: http://www.accrue.com/products/researchpapers.html.

[2] B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.

[3] S. Hirano and S. Tsumoto (2003): An Indiscernibility-based Clustering Method with Iterative Refinement of Equivalence Relations. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(2):169-177.

[4] J. Neyman and E. L. Scott (1958): Statistical Approach to Problems of Cosmology. Journal of the Royal Statistical Society, Series B20: 1–43.

[5] URL: http://lisp.vse.cz/challenge/

[6] S. Hirano and S. Tsumoto (2003): Multiscale Analysis of Long Time-series Medical Databases. Proceedings of AMIA Annual Symposium 2003, 289–293.