数理解析研究所講究録
第1603巻 2008年 111-119

k-nearest neighbours 判別を用いたクラスター解析のバリデーション

独立行政法人医薬品医療機器総合機構　浦田 正夫（Masao Urata）

1 はじめに

クラスター解析は、与えられたデータに含まれる個体をその特性値に基づいていくつかのクラスター（類似した個体の集団）のいずれかに分類する操作・手法である。判別分析とは異なり、外的に正解が与えられないことを特徴とする [1]。その目的は、一般にデータに含まれるクラスター個数を調べること、各個体がどのクラスターに属するかを判定することで、クラスターに階層構造（複数のクラスターがより上位のクラスターを構成する）が仮定される場合にはその階層構造を推定することも目的とされる。一般に、個体の分類に先立ち、与えられたデータに含まれるクラスター個数の判定が必要となるが、広く用いられている階層型クラスタリングや k-means 法などのクラスター解析手法では、クラスター個数の判定手順が手法そのものには含まれていない。そこで、クラスター個数を判定するために Clest [2] や prediction strength [3] などの多くの手法・指標が提案されている。これらの手法は、いずれもクラスター個数を指定し、得られるクラスター構造の「もっともらしさ」や安定性をなんらかの指標で評価して、適切と考えられるクラスター個数を判定するものである。一般にクラスター解析のバリデーションというと、これらのクラスター個数判定を意味することが多いが、クラスターの個数が正しく判定できていたとしても、各個体のクラスターへの帰属がどこまで正しく判定されているか、また、どのクラスター解析手法が最も望ましいかという疑問は残される。ここでは、クラスター構造の類似性の指標として classification error（CE）、判別の手法として k-NN 法（k-nearest neighbours 法）を用いたクロスバリデーション手順を用いて生成されるクラスター構造の安定性を評価することにより、この疑問に答えることを試みた。以下、次節で提案する手法の説明を行い、3節では求められる CE の性質について述べる。4節では、手法の特性を調べるために行なったシミュレーション（階層型クラスタリングにおけるリンケージ

の選択）の内容とその結果を示す。最後に、本研究のまとめと今後の課題を5節で述べる。

2 クラスター解析結果のバリデーション

クラスター解析結果を安定性をクロスバリデーション的な手順で調べる方法として、以下を考える。

1. 与えられたデータをトレーニングデータ、テストデータにランダムに分割する。
2. 所定のクラスタリング手法でトレーニングデータをクラスタリングする。
3. 所定のクラスタリング手法でテストデータをクラスタリングする。
4. テストデータの各個体について、トレーニングデータのクラスター解析結果に基づき、どのクラスターに帰属するかを適当な判別手法で判定する。
5. テストデータについて、手順３．と４．で得られた２通りのクラスター分類結果の類似性を適当な指標で評価する。
6. 手順１．～５．を多数回繰り返し、５．で求めた指標を集計する。

この手順は、Handl [4] らの分類による predictive power/stability アプローチに該当するものであり、所定のクラスタリング手法が生成するクラスター構造がどれだけ安定しているかをクロスバリデーション的に評価する。なお、評価されるのはクラスター構造の安定性であり、個体が真のクラスターに分類されているかどうかではないことに注意が必要である。ただし、安定性が低ければ真のクラスターに分類されている個体の割合も低いことが考えられる。なお、評価された安定性が低すぎる場合にはクラスターの個数を過大あるいは過小に判定したことが疑われ、クラスター個数判定の問題と重なる部分があるが、ここでは取り上げない。ここで検討が必要なのは、手順４．の判別手法と手順５．の指標である。判別分析の手法は多数あるが、その選択にあたっては用いるクラスタリング手法とそれが前提としている仮定を考慮すべきだろう。具体的には、k-means クラスタリングのバリデーションを行なう際、テストデータの各個体を k 個あるトレーニングデータのクラスターのうち、クラスター中心（含む個体の特性値の平均値）が当該個体の特性値に最も近いものに分類することが考えられる。本研究では、クラスタリング手法が強い仮定を前提としない場合、例えば階層型クラスタリングに対して、同様に強い仮定を必要としない k-nearest neighbours 判別（k-NN 判別）を用いることを検討する。

k-NN 判別は、正しく分類が行われている既存のデータから判別を行いたい個体に最も「近い」k 個の個体（k-NN）を選び出し、それらの個体が最も多く属しているクラスに当該個体を分類するものである。判別を行いたい個体の特性値を x とし、既存データの特

性値 $x_1, x_2, ..., x_n$ が可分距離空間 X 上の独立同一分布に従う確率変数としたとき、集合 $\{x_1, x_2, ..., x_n\}$ における 1-NN は $n \to \infty$ のとき x に確率 1 で収束することが知られている [5] 。

クラスター分類結果の類似性の指標としては、 prediction strength や classification error（CE）など多くのものが提案されているが、本研究ではその値が意味するところを感覚的に理解しやすい CE を用いることとした。

3 Classification error（CE）とその性質

CE は、以下のように定義される。

$$\mathrm{CE} = 1 - \frac{1}{n} \max_{\sigma} \sum_{k=1}^{K} n_{k,\sigma(k)} \tag{1}$$

ここで、n はサンプルサイズ、K はクラスターの個数、n_{ij} は第一のクラスタリングによりクラスター i、第二のクラスタリングによりクラスター j に分類された個体の個数であり、σ は $\{1, 2, ..., K\}$ から $\{1, 2, ..., K\}$ への単射である。これは、二種類のクラスタリング結果を分割表に集計し、適当にその行を入れ替えて対角要素の和を最大にしたときの非対角要素の割合を意味しており、異なるクラスターに分類された個体の割合を最も楽観的に評価したものである。

クラスター解析結果に対して上述の方法で 1-NN 判別（再近隣法とも呼ばれる）と CE を用いたクロスバリデーションを適用した場合に、得られる CE が持つ性質について、クラスターの個数が 2 個の場合で検討を行なった。

最初に、以下の表記を導入する。

$f(x)$	与えられたデータの特性値が従う確率密度関数
$\xi(x)$	特性値 x を持つ個体がクラス 1 に属する条件付き確率
$\eta_n(x)$	サンプルサイズが n である場合に特性値 x を持つ個体がクラスター 1（クラス 1 に対応する）に分類される確率
$\tilde{\eta}_n(x)$	サンプルサイズが n である場合に特性値 x を持つ個体の最近隣個体がクラスター 1（クラス 1 に対応する）に分類される確率

ここで、x は p 次元ユークリッド空間上の特性値ベクトルと考える。
これらの表記を用いると、クラスター解析手法の誤判別率 R_n は以下のように表現さ

れる。

$$R_n = \int \{\xi(x)[1-\eta_n(x)] + [1-\xi(x)]\eta_n(x)\} f(x)dx \tag{2}$$

一般に $\xi(x)$ を推定することはできないため、誤判別率 R_n を推測することはできない。そこで、かわりに以下の量（stability index）を考える。

$$\begin{aligned} R_{Cn} &\equiv \int \{\eta_n(x)[1-\eta_n(x)] + [1-\eta_n(x)]\eta_n(x)\} f(x)dx \\ &= \int 2\eta_n(x)[1-\eta_n(x)] f(x)dx \end{aligned} \tag{3}$$

Stability index は、ある個体のみを共通して含む 2 つのデータ（同一母集団からのサンプル）をそれぞれクラスタリングしたときにその個体が異なるクラスターに分類される確率であり、値が小さい場合には得られるクラスターの構造が安定していると考えることができる。また、$\eta_n(x)$ と $\xi(x)$ が連続である場合には、$R_n \geq R_{Cn}/2$ が成り立つ（補遺参照）。

サンプルサイズが n のときに上述の方法で求めた CE を $\hat{R}_{Cn}$ と表記し、これと R_{Cn} の関係を考える。$\hat{R}_{Cn}$ は、以下の期待値を持つ。

$$E[\hat{R}_{Cn}] = \int \{\tilde{\eta}_{n/2}(x)[1-\eta_{n/2}(x)] + [1-\tilde{\eta}_{n/2}(x)]\eta_{n/2}(x)\} f(x)dx, \tag{4}$$

次に、$\hat{R}_C$ の性質を考える。$\lim_{n\to\infty} E[\hat{R}_{C_n}]$ を評価するために、$\lim_{n\to\infty} \tilde{\eta}_{n/2}(x)$ を求める。ここで $\lim_{n\to\infty} \tilde{\eta}_{n/2}(x)$ が存在すると仮定すると、

$$\begin{aligned} \lim_{n\to\infty} \tilde{\eta}_{n/2}(x) &= \lim_{n\to\infty} \tilde{\eta}_n(x) \\ &= \lim_{n\to\infty} E[\eta_n(\mathrm{NN}_n(x))] \end{aligned} \tag{5}$$

ここで、$\mathrm{NN}_n(x)$ は n 個の個体のうちの特性値 x の最近隣値である。これは以下のように表すことができる。

$$= \lim_{s\to\infty} \lim_{t\to\infty} E[\eta_s(\mathrm{NN}_t(x))] \tag{6}$$

$|\ \eta_s(\mathrm{NN}_t(x))\ | \leq 1$ なので、ルベーグの有界収束定理より極限と期待値の順番が交換できて、

$$= \lim_{s\to\infty} E[\lim_{t\to\infty} \eta_s(\mathrm{NN}_t(x))] \tag{7}$$

である。$\eta_s(x)$ が連続である点 x において、$\lim_{t\to\infty} \eta_s(\mathrm{NN}_t(x))$ は $n\to\infty$ のとき $\eta_s(x)$ に収束する。ここで、x の最近隣値が確率 1 で x に収束する性質を用いた [5]。これより、

$$\lim_{n\to\infty} \tilde{\eta}_n(x) = \lim_{s\to\infty} E[\eta_s(x)] = \lim_{s\to\infty} \eta_s(x) \tag{8}$$

この結果とルベーグの支配収束定理をもう一度用いると、以下を示すことができる。

$$\lim_{n\to\infty} E[\hat{R}_{Cn}] = \lim_{n\to\infty} R_{Cn} \tag{9}$$

つまり、$\hat{R}_{Cn}$ は R_{Cn} の漸近不偏推定量である。また、$R_n \geq R_{Cn}/2$ より、求められた CE の概ね半分以上の割合で誤判別が生じているだろうとする解釈が可能である。

4 クロスバリデーションによるリンケージの選択

クラスタリングの手法として階層型クラスタリングを用いる場合には、個体間の距離の尺度の他に、クラスター間の距離を測定する尺度（リンケージと呼ばれる）も与える必要がある。ここでは、上述のクロスバリデーション手順で得られる CE を最小とするリンケージを選択することにより、クラスター解析の誤判別率を最も小さくするリンケージが選択できるかどうかをシミュレーションにより調べた。手順は以下の通りである。

1. 仮想データの発生
2. 階層型クラスタリング（6 種類のリンケージで）、真の誤判別率の算出
3. 各リンケージでクロスバリデーション手順により CE を算出（20 回）
4. 手順１．～３．を 100 回繰り返す

検討したリンケージは、average、median、complete、Ward's、single 及び centroid である。各データについて、真の誤判別率（MCR）が最小のリンケージと CE（20 回算出したものの平均）が最小であるリンケージを集計することとした。最初に、以下の設定でシミュレーションを行なった。

設定１

特性値の空間:	3 次元ユークリッド空間
クラスターの分布:	$(-2,0,0)$ 及び $(2,0,0)$ を中心とする二つの 3 次元標準正規分布
サンプルサイズ:	各クラスター 60

これは、3 次元空間上に 2 個の球状のクラスターが存在している構造である。結果を以下に示す。なお、タイが存在するため、合計は 100 を超えている。

表 1. リンケージ比較結果 設定 1

CE 最小 / MCR 最小	Average	Median	Complete	Ward's	Single	Centroid	計
Average	25	0	1	12	0	0	38
Median	7	0	0	0	0	0	7
Complete	19	0	3	10	0	0	32
Ward's	23	0	1	20	0	0	44
Single	0	0	0	0	0	0	0
Centroid	8	0	0	5	0	0	13
計	82	0	5	47	0	0	

真の誤判別率（MCR）についてみると、Ward's、average 及び complete リンケージは、それぞれ 100 回の繰り返しのうち 30〜40 回程度「最良」のクラスタリング結果を与えたことがわかる。これは、母集団の分布が同一であってもデータにより真の誤判別率を与えるリンケージが異なることを意味している。次に、以下の設定を検討した。

設定 2

特性値の空間: 3 次元ユークリッド空間

クラスターの分布: $(-0.8, 0, 0)$ 及び $(0.8, 0, 0)$ を中心とする二つの 3 次元標準正規分布で、共分散行列が diag(0.1, 1, 1)

サンプルサイズ: 90 及び 30

結果を以下に示す。ここでもタイが存在したため、合計は 100 を超えている。

表 2. リンケージ比較結果 設定 2

CE 最小 / MCR 最小	Average	Median	Complete	Ward's	Single	Centroid	計
Average	12	2	2	9	5	2	32
Median	4	2	1	9	2	1	19
Complete	2	1	0	1	0	0	4
Ward's	10	5	2	3	3	0	23
Single	21	9	2	9	16	4	61
Centroid	13	3	2	7	3	2	30
計	62	22	9	38	29	9	

実際にはsingleリンケージが最も低い誤判別率を達成することが多いが、CEの最小化を基準に選択を行うと、Ward'sリンケージやaverageリンケージの方が選択されやすいことがわかる。

2種類の設定のみでの結果ではあるが、設定1、2を通じ、averageリンケージはCEで評価されたように高い安定性を持つと考えられた。しかしながら、誤判別率は必ずしも低いとは言えない。真のクラスター構造を捉えるかどうかは別にして、安定したクラスター構造を見出しやすいことが考えられる。Ward'sリンケージは、いずれの設定でも低い誤判別率を示し、CEにもそれが反映されたものと考えられる。Singleリンケージは、低い誤判別率を達成できるデータにおいてもCEの値が比較的高く、CEを最小にするリンケージを選択するという基準では選ばれにくいと考えられた。

5 考察

マイクロアレイデータの解析等で階層型クラスタリングが使われることが多いが、リンケージの選択は解析者に委ねられている。リンケージの性質については詳細な研究がなされており [6]、その幾何学的な性質は明らかにされているが、実用上は誤判別の割合に興味があると考えられる。本研究では、k-NN法（k-nearest neighbours法）を用いたクロスバリデーション的な手順により評価したclassification errorと誤判別率の関係をクラスター数が2個の場合で示し、シミュレーションにより誤判別率を最も小さくするリンケージが選択できるかどうかを調べた。その結果、CEの最小化という基準では、誤ってはいるが安定したクラスターを見出しやすいリンケージ（averageリンケージ）を過大評価する傾向があると考えられた。、そのため、実際にリンケージの選択を行う際には、クラスターの安定性だけではなく、クラスター構造の幾何学的なもっともらしさを評価するバリデーション（Handl [4] らの分類によるcompactness, connectedness, separationの評価など）を組み合わせて用いる必要があると考えられる。また、singleリンケージが選択されにくいことが判明した。本手法ではクロスバリデーションのためにデータを分割し、サンプルサイズが半分になるため、性能がサンプルサイズの影響を受けやすいリンケージは見かけの安定性が低下することなどが考えられるが、その確認は行えていない。場当たり的な対応とはなるが、CEの最小化ではなく、サンプルサイズとリンケージの性質を考慮した上でCEを比較することが考えられる。なお、実用上、本バリデーション手法はクラスター個数の判定を行った上で用いられるものであるが、行ったシミュレーションはその点を考慮していないため、その影響について今後の検討が必要である。

補遺

以下で、$R_n \geq R_{Cn}/2$ を示す。$\eta_n(x)$ と $\xi(x)$ は連続と仮定する。

$$\begin{aligned} R_{Cn} &\equiv \int \{\eta_n(x)[1-\eta_n(x)] + [1-\eta_n(x)]\eta_n(x)\} f(x)dx \\ &= \int 2\eta_n(x)[1-\eta_n(x)] f(x)dx \end{aligned} \tag{10}$$

を以下のように分解する。

$$R_{CAn} = \int_{\eta_n \leq 0.5} 2\eta_n(x)[1-\eta_n(x)] f(x)dx \tag{11}$$

$$R_{CBn} = \int_{\eta_n > 0.5} 2\eta_n(x)[1-\eta_n(x)] f(x)dx \tag{12}$$

同様に、

$$R_n = \int \{\xi(x)[1-\eta_n(x)] + [1-\xi(x)]\eta_n(x)\} f(x)dx \tag{13}$$

を以下のように書く。

$$R_{An} = \int_{\eta_n \leq 0.5} \{\xi(x)[1-\eta_n(x)] + [1-\xi(x)]\eta_n(x)\} f(x)dx \tag{14}$$

$$R_{Bn} = \int_{\eta_n > 0.5} \{\xi(x)[1-\eta_n(x)] + [1-\xi(x)]\eta_n(x)\} f(x)dx \tag{15}$$

$\eta_n(x)$ を固定すると、R_{An} は $\eta_n \leq 0.5$ なる領域で $\xi(x) = 0$ である場合に最小値をとる。最小値は、

$$\min_{\xi(x)} R_{An} = \int_{\eta_n \leq 0.5} \eta_n(x) f(x)dx \tag{16}$$

(11) と (16) より、

$$2 \min_{\xi(x)} R_{An} = R_{CAn} + \int_{\eta_n \leq 0.5} 2\eta_n^2(x) f(x)dx \tag{17}$$

よって

$$R_{An} \geq \frac{1}{2} R_{CAn} \tag{18}$$

R_{Bn} と R_{CBn} についても同様の議論を行うと、

$$R_n \geq \frac{1}{2} R_{Cn} \tag{19}$$

が得られる。

参考文献

[1] Hastie, T., Tibshirani, R., Friedman, J. (2001) 'The Elements of Statistical learning', Springer (Chap 14)

[2] Dudoit, S. and Fridlyand, J. (2002) 'A prediction-based resampling methods for estimating the number of clusters in a dataset', Genome Biology. 3,0036:1-21

[3] Tibshirani, R., Walther, G., Botstein, D., and Brown, P. (2001) 'Cluster validation by prediction strength', Technical Report, Department of Statistics, Stanford University.

[4] Handl, J.,Knowles, J. and Kell, D. B. (2005) 'Computational cluster validation in post-genomic data analysis', Bioinformatics, 21, no.15, 3201-3212

[5] Cover, T.M. and Hart, P.E. (1967) 'Nearest Neighbor Pattern Classification', IEEE TRANSACTIONS ON INFORMATION THEORY, 13, NO.1, 21-27

[6] B. S. Everitt, et al. (2001) 'Cluster Analysis', A Hodder Arnold Publication