クラスタリングにおける最適クラスタ数の推定に関する研究

理学専攻・情報科学コース　谷本 聡子

1　概要

データ解析において類似の要素同士をまとめるクラスタリングは有効な手法のひとつであり，そのクラスタリング手法のひとつに k-means 法がある．k-means 法は，クラスタ数の決定は利用者に委ねられている．従って最適なクラスタ数を探す必要があるが，どのようなクラスタ数を最適とするかの判断基準は重要な問題である．

本論文では，k-means 法における最適クラスタ数を自動決定に関するひとつの手法を述べる．

k-means 法によってクラスタリングできるデータであっても，そのクラスタリング結果の評価が成功するとは限らない．線形手法では最適なクラスタ数を推定できるデータが限定されてしまうため，カーネル法を使ったクラスタリングの評価を提案する．

2　k-means 法

k-means 法は非階層的クラスタリング手法のひとつである．

k-means 法では，クラスタ数を k，i 番目のクラスタ C_i の重心を μ_i とした時，次の式を最小化するようにデータを分割してクラスタリングを行う．

$$\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|^2 \tag{1}$$

はじめに，ランダムにデータを k 個のクラスタに割り当て，それぞれの重心を得る．すべてのデータについて，これら k 個の重心との距離を求め，最も重心との距離が小さくなるクラスタに再度割り当てる．このアルゴリズムを繰り返すことで k 個のクラスタに分割することができる．

3　提案手法

最適なクラスタ数とはどのようなものか考えると，最適なクラスタ数とは，それぞれのクラスタ内の要素が密接にまとまっているものと言える．つまり，類似の要素が同じクラスタに属していて，かつクラスタ同士の距離が離れているものとなる．これを基にクラスタ数を求める基準を考える．

クラスタ内のそれぞれの要素の座標を x，i 番目のクラスタ C_i の座標の平均を μ_i とする．クラスタ C_i のまとまりを表す式は以下の式で表される．[1]

$$S_i=\frac{1}{\text{クラスタ内の要素数}}\sum_{x\in S_i}\|x-\mu_i\|^2 \tag{2}$$

クラスタ S_i と S_j の距離は各クラスタの重心同士の距離で求められる．

$$d_{ij}=\|\mu_i-\mu_j\|^2 \tag{3}$$

(2)，(3) より i 番目のクラスタ S_i がどの程度まとまっているかを表す r_i を得る．

$$r_i=max_j\frac{S_i+S_j}{d_{ij}} \tag{4}$$

すべてのクラスタから得られた r_i の平均がそのクラスタリングを評価する r となる．

クラスタ数を k として

$$r=\frac{1}{k}\sum_{i=1}^{k}r_i \tag{5}$$

この r が小さいほどクラスタ内の要素が密集し，クラスタ同士の距離が大きく，よいクラスタリング結果と言える．それを満たすクラスタ数を探索する．

ただしこれは各クラスタをユークリッド距離に基づいてまとまりを考える場合には有効である．クラスタによって要素の数や散布の方向が違うデータにはこの評価は有効とは限らないため．このようなデータのクラスタリングの評価を行う場合は非線形手法の導入が必要である．

4　カーネル k-means 法

前節で述べたように，線形手法である k-means 法では，提案した手法でのクラスタリングの評価が難しい場合がある．そこで本手法では，k-means 法にカーネル法を組み合わせたカーネル k-means 法を用いる．

カーネル法とはデータ解析手法のひとつであり，データを高次元の特徴空間に写像することによって非線形特徴を抽出するものである．また，データ x_i が特徴空間中で $\phi(x_i)$ と表されるとすると，特徴空間中での内積はカーネル関数 $k(x,y)$ の形で表すことができる．

$$k(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle$$

これによって，高次元の特徴空間において内積の計算が容易になる．本研究では以下のカーネル関数で表されるガウスカーネルを利用した．

$$k(x_i,x_j)=\exp(-\beta\cdot\|x_i-x_j\|^2)$$

カーネル法を利用すると (1) はカーネル関数のみを使って表すことができる．

$$\sum_{c=1}^{k}\sum_{x_i\in X_c}[k(x_i,x_j)-\frac{2}{|X_i|}\sum_{x_l\in X_i}k(x_j,x_l)+\frac{1}{|X_i|^2}\sum_{x_l\in X_i}\sum_{x_m\in X_i}k(x_l,x_m)]$$

5　計算結果

本手法を実行するにあたり図 1 のような 3 つのクラスタの集まったサンプルデータを使用した．このデー

タに対し，クラスタ数 $k = 2$，$k = 3$，$k = 4$ と設定してカーネル k-means 法によるクラスタリングを行った．$k = 2$ の時のクラスタリングの結果は図 2，$k = 3$，$k = 4$ の時のクラスタリング結果はそれぞれ図 3，図 4 のようになった．

このクラスタリング結果それぞれについて第 3 節の手法で評価を行った．結果は以下の通りである．

$k = 2$ のとき，$r = 1.12143$

$k = 3$ のとき，$r = 0.51503$

$k = 4$ のとき，$r = 0.66644$

$k = 3$ の時が最も r の値が小さくなり，適切な評価ができていると言える結果が得られた．

また，同じサンプルデータを使って k-means 法によるクラスタリング、およびその評価も行った．クラスタリングの結果はカーネル k-means 法でクラスタリングを行った場合と同様の結果となった．しかし，その評価結果は以下のようになった．

$k = 2$ のとき，$r = 35.7157$

$k = 3$ のとき，$r = 28.1968$

$k = 4$ のとき，$r = 27.3978$

$k = 3$ が最適なクラスタ数であるが，$k = 4$ の時に r は最も低い値を取り，最適クラスタ数は 4 であるという結果となった．

このように，線形手法である k-means 法ではクラスタリングが成功していてもその評価は必ずしも成功するわけではなく，最適クラスタ数を計算によって求められない場合がある．このようなデータに対して，カーネル k-means 法を使うことによって最適なクラスタ数を求めることができることが示された．

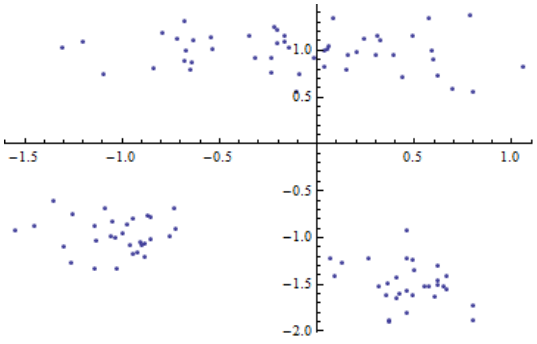

図 1: サンプルデータ

6 まとめ

本論文では k-means 法でのクラスタリングにおける最適クラスタ数を求めるために非線形手法のカーネル k-means 法を導入する手法を提案した．クラスタの要素数に差があるデータなど，線形手法でのクラスタリングの評価が難しい場合にも，カーネル法 k-means 法を用いて距離を測ることにより，適切なクラスタリングの推定が可能になった．

参考文献

[1] 嘉村準弥，小柳滋，x-means 法における分割停止規準の改良，第 8 回情報科学技術フォーラム，2009

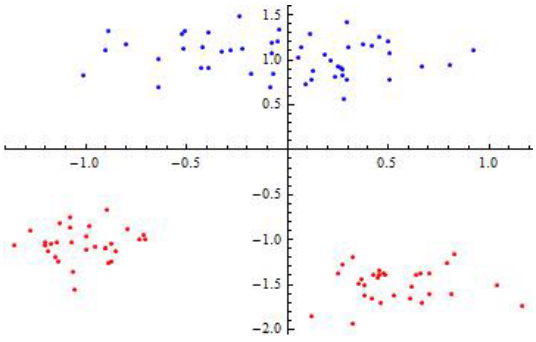

図 2: クラスタリング結果 ($k = 2$)

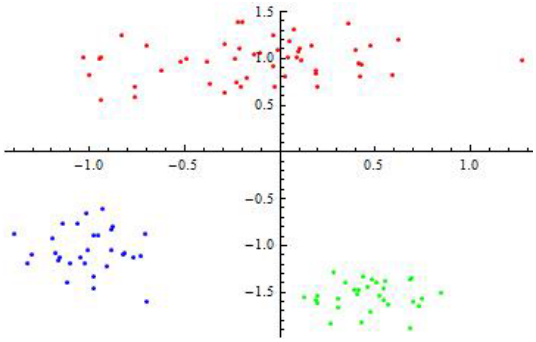

図 3: クラスタリング結果 ($k = 3$)

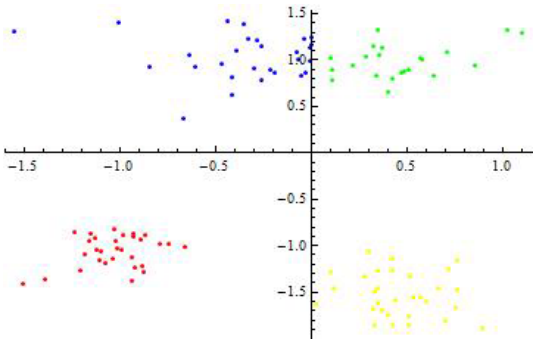

図 4: クラスタリング結果 ($k = 3$)