カーネル $k-$ 平均法の拡張によるノイズ除去クラスタリング

新村世梨　(指導教員：吉田裕亮)

1　はじめに

クラスタリングとは, データの集まりをデータ間の類似度に従って, いくつかの集合 (クラスタ) に分類することである. しかし実際にはどのクラスにも属さないノイズが含まれていることもあり得るため, すべてを明確に判別することは現実的ではない場合もある. なお, クラスタリングは線形クラスタリングと非線形クラスタリングの 2 つに大きく分類され, 線形クラスタリングの代表的な手法に $k-$ 平均法, 非線形クラスタリングとして $k-$ 平均法を非線形に拡張したカーネル $k-$ 平均法, スペクトラルクラスタリングがある. しかし, カーネル $k-$ 平均法は初期値依存が強く, 適切な解を得るのに何度も初期値を変えて繰り返さなければならない場合もあるという欠点がある.

本研究では, このカーネル $k-$ 平均法を拡張し, 線形で分けることの出来ない複数の群からなるデータを, カーネル $k-$ 平均法の欠点を改善し, 初期値に強く依存することなく, 複数のクラスタとノイズに分類する手法を考察する.

2　$k-$ 平均法

$k-$ 平均法は, クラスの代表点からメンバーへの二乗距離の総和, つまり以下の式の値が最小となるように, クラスを分ける方法である. ($x^{(j)}$：サンプル点, μ_i：グループ i の代表点, N_i：μ_i を代表点とするグループのメンバーの集合)

$$R = \sum_{i=1}^{c} \sum_{x^{(j)} \in N_i} \left\| x^{(j)} - \mu_i \right\|^2$$

あらかじめグループ数 c を決めておき, クラスの代表点の初期値を適当に選ぶ. そしてサンプル点を最も距離の近い代表点のクラスへ分類し, 改めて代表点を更新する. 以上の手続きを収束するまで繰り返す. $k-$ 平均法は線形クラスタリングであるため, クラスタの形が非線形の場合は, 適切にクラスタリングすることができない.

3　カーネル法

カーネル法とは, データ x, x' が与えられたとき, それらの間の関係を $k(x, x')$ という実数値関数であるカーネル関数によって要約し, 全てを数値の世界に置き換えて処理する方法である. カーネル関数は, 2 つの対象 x, x' の類似度を表していると考えることもでき, 2 つの要素 x, x' に対し, それぞれの特徴ベクトルどうしの内積として定義される. すなわち, $\phi(x), \phi(x')$ を特徴ベクトルとして

$$k(x, x') = \phi(x)^{\mathrm{T}} \phi(x')$$

と表される. カーネル関数には様々あるが, 本研究では以下の Gauss カーネル

$$k(x, x') = \exp\left(-b\|x - x'\|^2\right)$$

を用いた.

4　カーネル $k-$ 平均法

カーネル $k-$ 平均法とは, $k-$ 平均法をカーネルを用いて一般化したものである. サンプル点集合 $x^{(1)}, x^{(2)}, ..., x^{(n)}$ が与えられ, その特徴ベクトルを $\phi(x^{(1)}), \phi(x^{(2)}), ..., \phi(x^{(n)})$ とするとき, それらをグループ分けすることを考える.

まず, グループ数 c はあらかじめ決めておき, 代表点を $\mu_1, \mu_2, ..., \mu_c$ とする. それぞれのサンプル点は, その点に最も近い代表点 μ_i のグループのメンバーの集合 N_i に入り, 代表点はグループに属するサンプル点の重心に取ることにする. このとき,

$$L = \sum_{i=1}^{c} \sum_{x^{(j)} \in N_i} \left\| \phi(x^{(j)}) - \mu_i \right\|^2$$

を最小とするように N_i, μ_i を決める.

特徴ベクトルと代表ベクトルとの距離は

$$\begin{aligned} \left\| \phi(x^{(j)}) - \mu_i \right\|^2 &= \left\| \phi(x^{(j)}) - \frac{1}{|N_i|} \sum_{x^{(l)} \in N_i} \phi(x^{(l)}) \right\|^2 \\ &= k(x^{(j)}, x^{(j)}) - \frac{2}{|N_i|} \sum_{x^{(l)} \in N_i} k(x^{(j)}, x^{(l)}) \\ &\quad + \frac{1}{|N_i|^2} \sum_{x^{(l)} \in N_i} \sum_{x^{(m)} \in N_i} k(x^{(l)}, x^{(m)}) \end{aligned} \tag{1}$$

となり, カーネル関数だけを使って書き表される. アルゴリズムは以下のようになる.

1. サンプルを適当に c 個のグループに分け, N_i を初期化する.
2. 式 (1) に基づいて N_i を更新する.
3. グループ分けが収束するまでステップ 2 を繰り返す.

目的関数 L が単調に減少していくことが示されるので, このアルゴリズムは局所最適解に収束することは保証されるが, 一般に大域的な最適解に収束するとは限らない.

5　提案手法

本研究では, カーネル $k-$ 平均法のクラス分けにおいて, Gibbs 分布による確率的な処理をし、温度変化による状態遷移を可能にする手法を提案する. こうすることにより, 局所解に陥ることを防ぎ, 初期値への依存性を弱めることを考える. また, 温度を下げたときの帰属確率の収束を, ノイズ判断に役立てる.

特徴ベクトル $\phi(x^{(j)})$ と代表ベクトル μ_i との距離の 2 乗を m_{ij} としたとき, この m_{ij} の重みに基づく Gibbs

分布を用いて, 特徴ベクトル $\phi(x^{(j)})$ におけるクラス i への帰属確率 α_{ij} を

$$\alpha_{ij} = \frac{\exp\left(-\beta\, m_{ij}\right)}{\sum_{p=1}^{c} \exp\left(-\beta\, m_{pj}\right)} \qquad \left(\beta = \frac{1}{T}\right)$$

とする. ここで T は物理モデルの温度に対応する. また, これに伴い代表ベクトルを

$$\mu_i = \frac{1}{\sum_l \alpha_{il}} \sum_l \alpha_{il}\, \phi(x^{(l)})$$

と求める. 特徴ベクトル $\phi(x^{(j)})$ と代表ベクトル μ_i との距離の 2 乗 m_{ij} は通常のカーネル $k-$ 平均法と同様にカーネル関数だけで表すことができる.

こうすることにより, 例えばクラス数を 2 としたとき, $T \to \infty$ のとき, 2 つのクラスへの帰属確率は $(\alpha_{1j}, \alpha_{2j}) \to (0.5, 0.5)$ となる. $T \to 0$ のとき, $(\alpha_{1j}, \alpha_{2j}) \to (1, 0) \mathrm{or} (0, 1)$ となり, 通常のカーネル $k-$ 平均法と同様になる.

また, 温度を下げたときの, この比への収束の状況でノイズと判別することを考える.

6 実験例

6.1 実験 1

図 1 のような, それぞれ 100 個ずつからなる大小の円を組み合わせた計 200 個のサンプルデータを用意する.

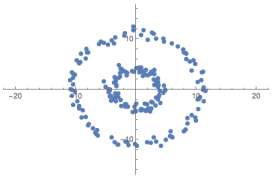

図 1： サンプルデータ

図 1 をランダムにクラス分けした, 3 つの初期値を図 2 のように用意する.

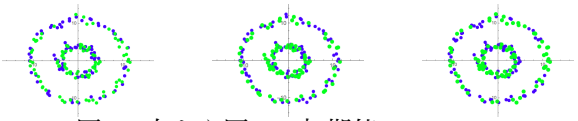

図 2：左から図 1 の初期値 1, 2, 3

図 2 の初期値 1, 2, 3 を用いて, 通常のカーネル $k-$ 平均法でクラスタリングを行った結果が, 図 3 である. 次に, 同じく初期値 1, 2, 3 を用いて, 本研究での提案手法でクラスタリングを行った結果が, 図 4 である. 温度のスケジュールは, $\frac{1}{T} = 30$ から始め, 帰属確率が $(0.5, 0.5)$ になるまで温度を上げ, その後収束するまで下げた.

図 3 では初期値によって結果が異なっており, 適切にクラス分けができていない. 一方, 図 4 では, 初期値に依存することなく, 適切にクラス分けができていることが分かる.

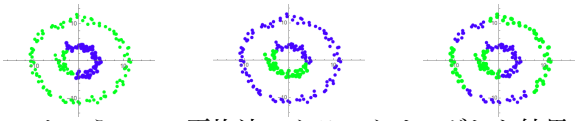

図 3：カーネル $k-$ 平均法でクラスタリングした結果

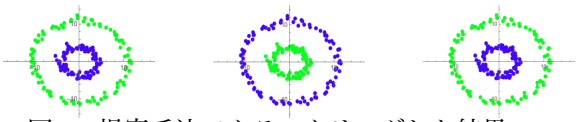

図 4：提案手法でクラスタリングした結果

6.2 実験 2

図 5 のような, 2 つの群からなる各 100 個ずつのデータに, ノイズとして一様乱数 100 個を加えた計 300 個から構成されるサンプルデータを用意する. このデータを, 提案手法を用いて 2 つのクラスとそれ以外のノイズに分ける.

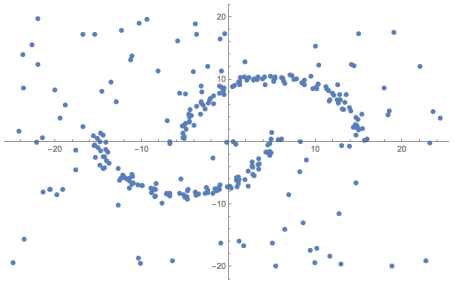

図 5： サンプルデータ

まず初期値をランダムに決め, クラス数を 2 としてノイズを含めてクラスタリングを行い, 温度を下げたときの帰属確率の収束が, ある閾値より遅いものをノイズとした結果が図 6 である. 温度 T のスケジュールは実験 1 と同様にした. おおむね良好な判別結果が得られた.

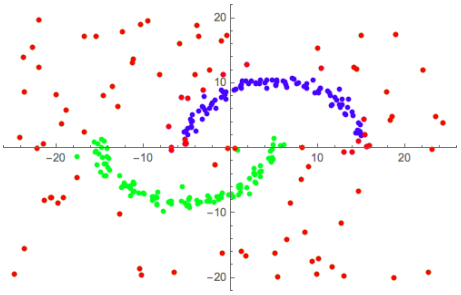

図 6： 実験結果

7 まとめと課題

カーネル $k-$ 平均法を拡張し, 線形で分けることのできない複数の群からなるデータを, 初期値に強く依存することなく, 複数のクラスタとノイズに分類することができた.

今後の課題として, カーネル関数のパラメータ b の適切な値をどのように探し出すか, ノイズであると判断する基準となる閾値をどこに設定するか, 温度パラメータ T の適切なスケジュールをどのように決めるか, が挙げられる.

参考文献

1. 赤穂昭太郎, カーネル多変量解析〜非線形データ解析の新しい展開, 岩波書店, 2008
2. 麻生英樹・津田宏治・村田昇, パターン認識と学習の統計学〜新しい概念と手法, 岩波書店, 2003