事前ノイズ除去手法によるクラスタリング

森恵望　(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有り得るため, 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.

本研究では線形で分けることのできない、複数の群からなる複雑なデータに、一様ノイズを加えたものを、スペクトラルクラスタリングを用いて、複数のクラスタとそれ以外のノイズに分類する手法を検討する。

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格に k-means 法, 非線形クラスタリングの代表格に k-means 法を非線形に拡張した カーネル k-means 法がある. しかしカーネル k-means 法は, 反復演算が必要であったり, 収束解が必ずしも目的関数を最適化するものではない等の欠点があり, 解決策としてスペクトラルクラスタリングがよく用いられているので, 本研究ではスペクトラルクラスタリングで検討する.

2　スペクトラルクラスタリング

スペクトラルクラスタリングとは, 整数計画問題を緩和した固有値問題を解くことによってクラスタリングを行う手法のことである. サンプル点からなるグラフ構造を例に考えると, 各頂点がサンプル点で, 枝にはサンプル点どうしの近さを表す重みがついているとする. サンプル点を 2 つのグループに分けるとすると, それに伴いグラフも 2 分割される. 分割されたグループ間を結ぶ枝のことを分割カットと呼び, グループ内は出来るだけ近いもの同士が集まり, グループ間は遠く離れていることが望ましいので, このカットの重みの合計が小さくなるようにグループ分けを行う. 式で表すと以下のようになる.

$$\min_{\beta} \sum_{i,j} K_{ij} \left(\beta_i - \beta_j\right)^2 = \boldsymbol{\beta}^T P \boldsymbol{\beta}, \ \ \beta_i = \pm 1$$

ここで, P は, 対角行列 Λ を $\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$ として, $P = \Lambda - K$ とおく. $\boldsymbol{\beta}$ が 2 値ベクトルという制約があるため, これは整数計画問題と呼ばれ, 一般に解くのが困難である. そこで, 整数という制約を取り払って任意の実数ベクトルに制約を緩め, また, $\boldsymbol{\beta}$ の大きさを制約するために, $\boldsymbol{\beta}^T \Lambda \boldsymbol{\beta} = 1$ の条件のもとで固有値の推定を行う. この場合, 最小固有値 0 が存在するが, これは全てのサンプルを 1 つにまとめてしまうという意味のない解なので, 実際には 2 番目以降の固有ベクトルの成分符号に基づいてクラスタリングを行う.

3　カーネル法

カーネル法とは、複雑なデータ x, x' があったとき, 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. 式は以下の通りで, $\phi(x), \phi(x')$ は特徴ベクトルを表している.

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

カーネル関数には様々あり, 中でも Gauss カーネル

$$k(x, x') = \exp(-\beta \|x - x'\|^2)$$

を用いて特徴空間に写像した行列は, 相関行列と同じような振る舞いをし, 対応する特徴ベクトルは無限次元に写像したと考えられるため, 本研究ではカーネル関数としてガウスカーネルを採用した.

4　データ間距離

カーネル法ではデータを高次元の特徴空間上へ写像する. 特徴空間の各座標はデータ要素の一つの特徴に対応し, 特徴空間への写像によりデータの集合はユークリッド空間中の点の集合に変換される. そのためカーネル関数は, 特徴空間における内積を評価し, 計算量が抑えられる. 特徴空間上でクラスタの重心とサンプルデータ x_i の距離を測るため, 距離 d_i は以下のように計算される.

$$d_i = \left\|x_i - \mu\right\|_{\phi}^2 = K(x_i, x_i) - \frac{2}{n} \sum_{j=1}^{n} K(x_i, x_j) + \frac{1}{n^2} \sum_{j=1}^{n} \sum_{l=1}^{n} K(x_j, x_l)$$

x_i は各データの値, μ はクラスタの平均, n はクラスタのデータの数である. 本研究では, この値が大きいもの, つまり特徴空間上において, 重心から離れているデータをカットし, ノイズとして推定する手法を用いる.

5　提案手法

本研究では, 一様なノイズが含まれている複数の群から成るデータでスペクトラルクラスタリングを実行すると, 非線形なクラスタリングが一般に困難であることから, 事前にノイズを大まかに取り除いたあと, スペクトラルクラスタリングを行う手法を提案する.

Step1.　まずはデータを 1 つのクラスタとみて, Gauss カーネルによる特徴空間上でのクラスタの重心とデータ間距離を計算する.

Step2.　そのデータ間距離がある閾値より大きいデータをノイズとみなし, 除去する.

Step3.　抽出されたデータだけでガウスカーネル行列を構成し, ここでスペクトラルクラスタリングを行う.

Step4. クラスタリングされた各クラスタに対し, データ間距離を計算し, 残りのノイズを抽出する.

6 実験例

図 1 のような, 線形で分けることの出来ない各 150 個からなる 3 つの群と, ノイズとして一様乱数 250 個を加えた, 計 700 個のサンプルデータを用意する. このデータをスペクトラルクラスタリングで, 3 つのクラスとそれ以外のノイズに分ける.

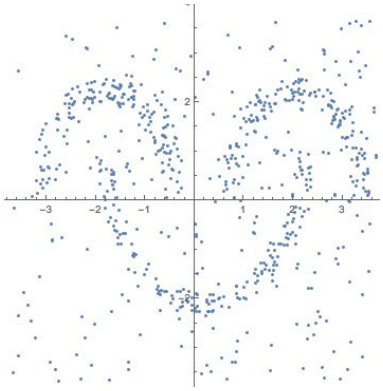


図 1: サンプルデータ

6.1 ノイズを抽出

Step1, Step2 の結果が下の図 2 である.

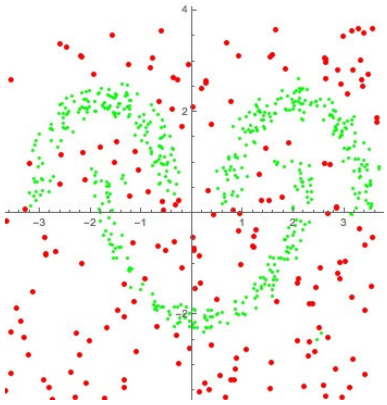


図 2: ノイズの抽出結果

6.2 スペクトラルクラスタリングを実行

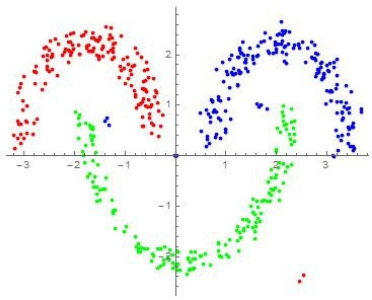


図 3: スペクトラルクラスタリングの実行結果

Step3 の結果が図 3 である. Step2 で取り除けなかったノイズを含め, 非線形に 3 つのクラスに分かれた.

6.3 実験結果

Step4 の結果, 以下のように良好な判別結果が得られた.

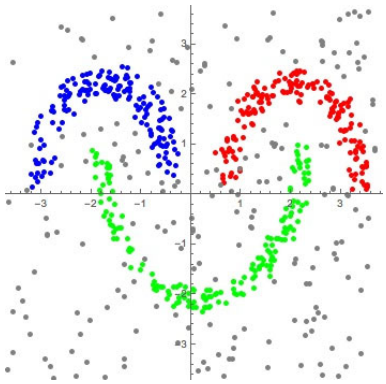
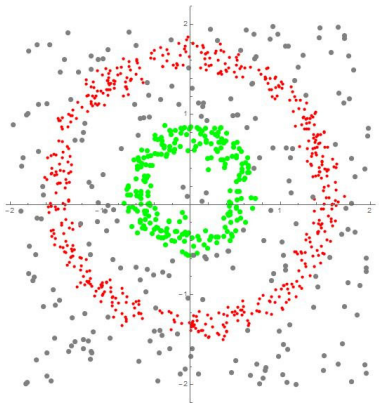


図 4: 実験結果

図 4 では, 非線形に 3 つのクラスとそれ以外のノイズを分けることができた. また, クラスタが二重の輪になっているようなデータでも, 以下のようなクラスタリング結果が得られた.



7 まとめと課題

今回の実験で, 事前にノイズを除去することにより, スペクトラルクラスタリングの精度を高めることが可能である場合があることが, 確認出来た.
今後の課題として, カーネル関数のパラメータ β の適切な値をどのように探し出すか, ノイズを抽出する基準となる閾値をどこに設定するか, の 2 点が挙げられる.

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〜新しい概念と手法, 岩波書店, 2003

[3] 戸井田明, クラス外ノイズを考慮したスペクトラルクラスタリング, お茶の水女子大学理学部情報科学科卒業研究, 2012