固有ベクトルのクラスタリングを援用したノイズ除去多群クラスタリング

伊藤真友 (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タにも属さないノイズが混入していることもあり得るため, 全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, ノイズが混入した 3 つの群からなるデータを, 固有ベクトルの成分判別をすることにより, 3 つのクラスタとノイズに判別するスペクトラルクラスタリングの拡張手法を提案する.

2　先行研究

竹田ら [1] は, 固有ベクトルの成分判別によるノイズ除去スペクトラルクラスタリングを提案した. 竹田らのクラスタリングは, 固有ベクトルをヒストグラムで表示し, 手動でノイズの識別を行い, 2 つのクラスタとノイズを判別するクラスタリングの手法を示している.

本研究では, 竹田らの研究を拡張し, 3 つの群からなるノイズ入りのデータに対して, 固有ベクトルの成分判別を自動で行うクラスタリング手法を提案する.

本研究と同様にノイズ除去多群クラスタリングを行っている宮原らの研究 [2] もある.

3　クラスタリング

クラスタリングは大きく階層的クラスタリングと非階層的クラスタリングの 2 つに分類される. 非階層的クラスタリングの代表に K-平均法がある. K-平均法は非常に有用であるが, 初期値依存が強く収束解が必ずしも目的関数を最適にするものでない点と, 反復演算を必要とするという欠点がある. 本研究で用いるスペクトラルクラスタリングは, クラスタリングの問題を固有値問題として定式化することによって, これらの問題を避けることができる.

また, K-平均法は, データを最も近いクラスタに分類するという線形なクラスタリング手法なので, データの形によってはうまくいかない場合もある. 一方スペクトラルクラスタリングでは, 与えられたデータをカーネル法を用いて高次元の特徴空間に写像してからクラスタリングを行うので, 非線形なクラスタ形状を持つデータでもうまくクラスタリングすることが可能となる.

4　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, 全てを数値に置き換えて処理する方法である. カーネル関数は特徴量で見たときの x と 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同士の内積として定義される. すなわち, $\phi(x), \phi(x')$ を高次元空間の特徴ベクトルとして

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

と表される. カーネル関数にはさまざまな種類があるが, 本研究では以下の Gauss カーネル

$$k(x, x') = \exp(-\beta ||x - x'||^2), \quad \beta > 0$$

を用いた.

5　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は, サンプル点をグラフ構造として考える. 各頂点がサンプル点で枝にはサンプル点同士の近さを表す重みがついているとし, 例えばサンプル点を 2 つのグループに分けると, 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カットと呼び, 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.

$$\min_{\boldsymbol{\beta}} \sum_{i,j} A_{i,j} (\beta_i - \beta_j)^2 = \min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T \Lambda \boldsymbol{\beta}, \quad \beta_i = \pm 1$$

ここで Λ は対角行列 D を $D_{ii} = \sum_{j=1}^{n} A_{i,j}$ として $\Lambda = D - A$ と書ける. $\boldsymbol{\beta}$ は 2 値ベクトルという制限がある. これは整数計画化問題と呼ばれ, 一般には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ベクトルに ${}^T\boldsymbol{\beta} D \boldsymbol{\beta} = 1$ という条件の下, 制約を緩めることにより推定を行う. この場合, 最小固有値 0 が存在するが, これはすべてのサンプルを 1 つにまとめてしまうという意味のない解のため, 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.

6　提案手法

一様なノイズが含まれている複数の群からな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. そこで本研究では, 固有ベクトルの判別により自動でノイズを取り除き, スペクトラルクラスタリングを行う拡張手法を提案する. 以下の手順を繰り返す.

1. 与えられたデータから, パラメータを設定し Gauss カーネル行列を計算
2. $\Lambda = D - A$ を構成
3. ラプラシアン Λ の固有値と固有ベクトルを算出
4. 第 2 固有ベクトル, 第 3 固有ベクトルの成分を判別しノイズ推定を行う.

図 1 は第 2, 第 3 固有ベクトルの成分をヒストグラムで表示したものである. ノイズがない場合は, はっきりとピークが分かれているが, ノイズがある場合は, 成分分布にピークはあるがはっきりとは分かれていない. 成分分布のピークから外れているデータをノイズであると判断する.

本研究では, 固有ベクトル成分を密度準拠クラスタリングである DBSCAN で適当な数のクラスタに分ける.

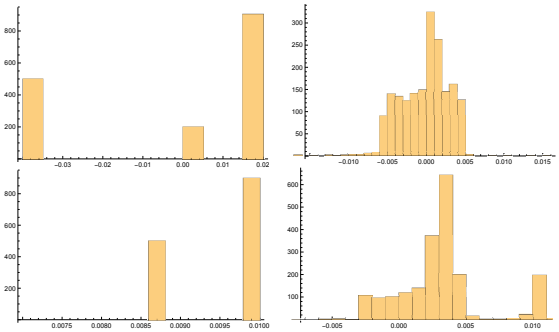

図 1: 固有ベクトルの成分分布
(左側ノイズなし, 右側ノイズあり)

そして, それぞれの固有ベクトルの成分群の平均値から一定数以上離れているデータをノイズとみなす.

7 実験例

7.1 実験 1

図 2 のような, 3 つの群からなる各 900 個ずつのデータに, ノイズとして一様乱数 200 個を加えた計 2900 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 3 つのクラスタとそれ以外のノイズに判別する.

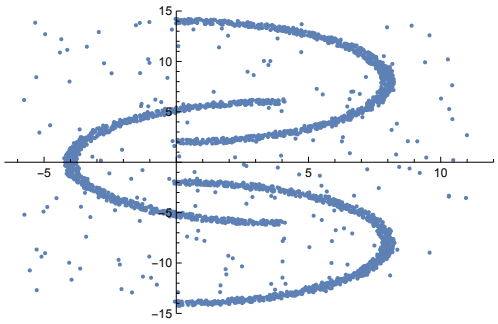

図 2: サンプルデータ

この実験では $\beta = 8.0$ に設定した. 最終的な結果を図 3 に示す. 黒い点がノイズと推定されたデータである. 良好な判別結果が得られた.

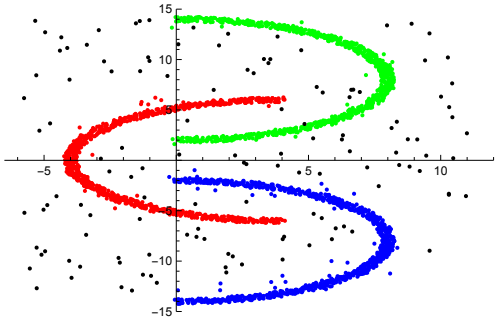

図 3: ノイズの推定結果

7.2 実験 2

図 4 のような, 外側の円から 2400 個, 1400 個, 800 個のデータに, ノイズとして一様乱数 200 個を加えた計 4800 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, 実験 1 と同様に 3 つのクラスとそれ以外のノイズに判別する.

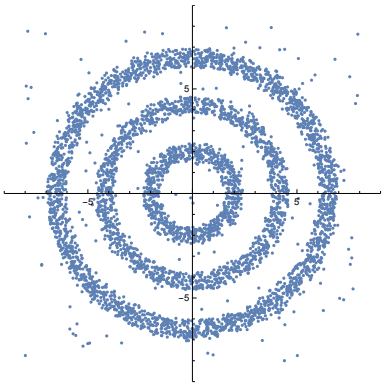

図 4: サンプルデータ

この実験では $\beta = 8.0$ に設定した. 最終的な結果を図 5 に示す. 黒い点がノイズと推定されたデータである. 比較的難しいと言われる形状のデータであるが, 良好な判別結果が得られた.

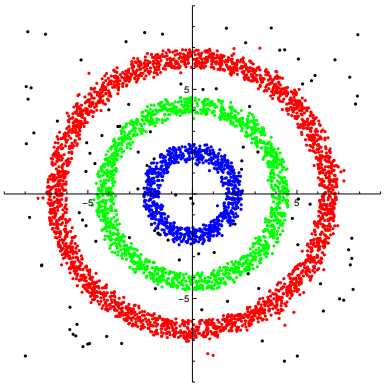

図 5: ノイズの推定結果

8 まとめ・課題

線形で分けることのできない複数の群からなるノイズが混入したデータを, スペクトラルクラスタリングで用いられる固有ベクトルを精査してノイズを自動除去するクラスタリングを行うことができた.

今後の課題として, カーネル関数のパラメータ β の適切な値をどのように設定するか, ノイズであると判断する閾値をどのように設定するか, が挙げられる.

参考文献

[1] 竹田ほのか. 固有ベクトルの成分判別によるノイズ除去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, 2019.

[2] 宮原颯, 駒崎幸之, and 宮本定明. コアポイント抽出によるスペクトラルクラスタリングの効率化. In 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集 第 *29* 回ファジィシステムシンポジウム, pages 21–21. 日本知能情報ファジィ学会, 2013.