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

竹田ほのか　(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, 線形で分けることの出来ない 2 つの群からなる不明確なデータを, 固有ベクトルの成分判別をすることにより, 2 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手法を考察する.

2　クラスタリング

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に K-平均法がある. K-平均法は非常に有効なクラスタリング手法であるが, 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.

一方スペクトラルクラスタリングでは, 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, これらの問題点を避けるアルゴリズムを構成することができる.

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. しかし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, 非線形性なクラスタ形状をも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.

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 (直感的には x と x' の近さ) 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. 本研究では, カーネル関数として以下の Gauss カーネル

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

を用いた. β は非線形性を調整するパラメータの一種と考えられる.

4　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, 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する. したがって, 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分割カットと呼び, 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.

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

ここで, P は対角行列 Λ を $\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$ として, $P = \Lambda - K$ と書ける. β は 2 値ベクトルという制約がある. これは整数計画問題と呼ばれ, 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実数ベクトルに, $\beta^T \Lambda \beta = 1$ という条件の下, 制約を緩めることにより推定を行うことになる. 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成分符号に基づいてクラスタリングを行う. 2 群の場合は第 2 固有ベクトルを用いることになる.

5　提案手法

以下の手順をカーネルパラメータ β を変え繰り返す.

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. $P = \Lambda - K$ を構成
3. ラプラシアン P の固有値と固有ベクトルを算出
4. 下から 2 番目の固有ベクトルの成分を判別しノイズの推定を行う

本研究では, これで得られた弱い学習を強い学習にすべく, 個々の学習を重ね合わせ, 学習を強化させるため多数決の手法を用いる. これにより, 1 回では除去しきれないノイズであっても, 多数のノイズ推定を重ね合わせることで除去できるのではないかと考える.

6　実験例

6.1　実験 1

図 1 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 150 個を加えた計 350 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

イズに判別する.

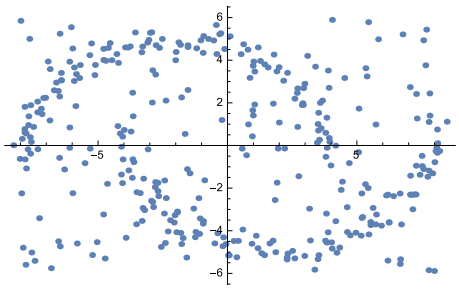

図 1: サンプルデータ

まず, カーネルパラメータ β を 3.0 に設定する. 固有ベクトル成分をヒストグラムで表示し判別を行う. ここで図 2 のように, 成分分布ピークから外れている中心部はノイズと判断する. クラスタが 2 つあるため分布ピークは 2 つ現れる. 比較のためノイズがない場合のヒストグラムも図 2 右で示したが, はっきりとピークが 2 つ現れている. クラスタの範囲を設定しクラス分けする. このような過程を経てカーネルパラメータ $\beta = 3.0$ におけるクラスタリング結果を得る.

以上の実験を $\beta = 3.5, 4.0, 4.5, 5.0, 5.1$ に関しても実行した. ひとつひとつに着目すると, データにばらつきがありノイズをノイズとして抽出できていない. しかし, 6 個を重ね合わせることで得られた最終的な結果が図 3 である (赤い点が最終的に推定されたノイズといえる). 良好な判別結果が得られた.

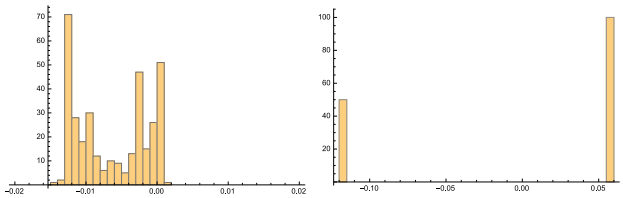

図 2: 左からノイズあり, ノイズなしの成分分布図

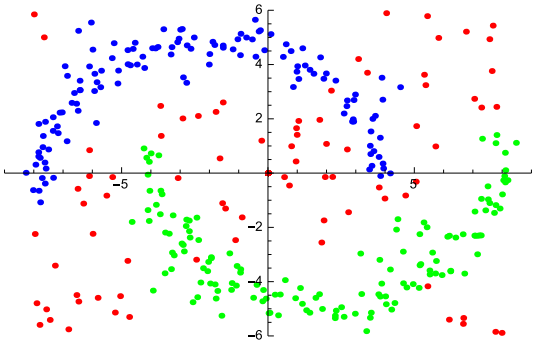

図 3: ノイズの推定結果

6.2 実験 2

図 4 のような, 2 つの群からなる外円内円それぞれ 300 個と 50 個のデータに, ノイズとして一様乱数 200 個を加えた計 550 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 と同様の手法で 2 つのクラスとノイズに分ける.

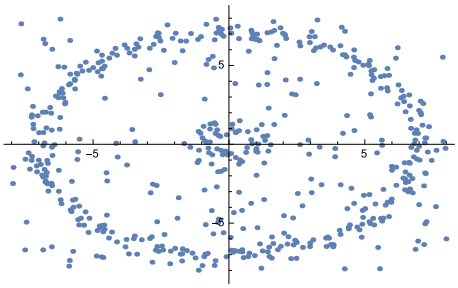

図 4: サンプルデータ

比較的難しいと言われる形状のデータのため, カーネルパラメータ β を 17 回変え多数決の候補を増やし, 結果を重ね合わせた. また, β について $\beta = 3.0 \sim 4.0$ で安定したため, なるべくこの範囲で多数決の候補をとった.

最終的に図 5 のおおむね良好な判別結果が得られた.

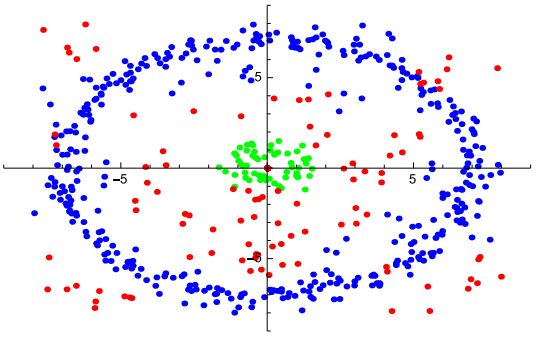

図 5: ノイズの推定結果

7 まとめ・課題

複雑なサンプルデータは 1 回のクラスタリングでは良い結果を得られないが, 弱い学習を重ね合わせることでうまくノイズを抽出して, 非線形クラスタリングを行うことができた.

しかし, カーネルパラメータ β の値の設定, そして固有ベクトル成分判別で結果が左右されるので, 適切な値の範囲を探すのが重要である. 今後の課題として, 固有ベクトル成分の範囲設定を自動化する手法の導入が挙げられる.

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, 東京, 2009.

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, 東京, 2003.