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

戸井田明　(指導教員：吉田裕亮)

1　はじめに

クラスタリングとは, 与えられたいくつかのデータをいくつかのある集合に明確に分類することである. つまり, あるクラスタを考えるときに, 1つ1つのデータがそのクラスタに属するか属さないかが明確に判別される. しかし, 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. 実際にはどのクラスにも属さないノイズが混入していることもあり得ると考えられる.

そこで本研究ではこのような不明確なデータをどのクラスタにも属さないような, クラスタリングの手法について考察する.

2　クラスタリング

クラスタリングは, 階層的手法と非階層的手法の2つに大きく分類され, 非階層的手法の代表例に $K-$ 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はあるが, 初期値に依存するアルゴリズムなので, 収束解が必ずしも目的関数を最適にするものではないという点と, 反復演算を必要とするという欠点がある.

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上に写像してからクラスタリングを行うので, 非線形なクラスタリングに拡張され, 非線形性なクラス形状をもつデータでも上手くクラスタリングすることができる.

3　カーネル関数

非線形なクラス形状をもつ複雑なデータを扱うために, カーネル関数を用いる. カーネル関数 $k(x,x')$ とは, データ変数の集合の2つの要素 x,x' に対し, x,x' のそれぞれの特徴ベクトル $\phi(x),\phi(x')$ どうしの内積

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

として定義される. カーネルには様々なものがあるが, その中でもガウスカーネル

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

を用いて特徴空間上に写像した行列は, 相関行列と同じような振る舞いをすることが知られている. ここで, $||\cdot||^2$ は通常のユークリッド距離2乗で, $\beta \in R$ は適当なパラメータである.

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.

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

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, ${}^T\beta\Lambda\beta = 1$ 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の成分符号に基づいてクラスタリングを行うことになる.

5　データ間距離の計算

カーネル法では, 非線形なデータを一度高次元の特徴空間上に表現し, 写像することで, 解析しやすいデータに変換し, その特徴空間上で線形なモデルを組み立て, 問題を解く. このとき, 特徴空間上での内積をカーネル関数を用いて計算することにより, 計算量を抑えることができるという利点がある.

各クラスの重心とサンプルデータの距離を測るために, 距離 d_i は以下のようにカーネルを用いて特徴空間上で計算する.

$$\begin{aligned} d_i &= ||x_i-\mu||^2 \\ &= K(x_i,x_i) - \frac{2}{n}\sum_{j=1}^{n}K(x_i,x_j) \\ &\quad + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}K(x_j,x_l) \end{aligned}$$

x_i は各データの値, μ は各クラスタの平均, n はその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.

6　提案手法

まずデータを与え, そのデータから与えられるカーネル行列に非線形にクラスタリングする. クラス分け

された各クラスタに対し, もう一度適当なパラメータを用いてカーネル行列を計算し, 特徴空間上での各クラスタの重心とデータの距離を計算する. そしてその重心から遠いデータをノイズとし, 抽出するという手法を試みる.

本研究では, 以下のような手法を提案する.

1. 与えられたデータのガウスカーネル行列 $\mathbf{K}$

$$\mathbf{K} = \begin{bmatrix} k(x_1,x_1) & k(x_2,x_1) & \dots & k(x_n,x_1) \\ k(x_1,x_2) & k(x_2,x_2) & \dots & k(x_n,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_1,x_n) & k(x_2,x_n) & \dots & k(x_n,x_n) \end{bmatrix}$$

   を構成し, スペクトラルクラスタリングを行う.
2. クラス分けされたうちの各クラスタに対し, 適当なパラメータを選び, もう一度ガウスカーネル行列を計算する.
3. この新たなガウスカーネル行列を用い, クラスタの重心とデータの距離を計算し, そのデータ間距離の値がある閾値より大きいデータをノイズと推定する.
4. これをすべてのクラスタに対して実行し, すべてのノイズを抽出する.

7 実験例

図 1 のような, 線形で分けることのできない 2 つの群からなるサンプルデータを用意する. 250 個ずつの 2 つの群に, ノイズとして-0.8 から 0.8 の一様乱数 300 個を加えた, 計 800 個のサンプルデータとなっている. このデータをスペクトラルクラスタリングで, 2 つのクラスとそれ以外のノイズに分ける.

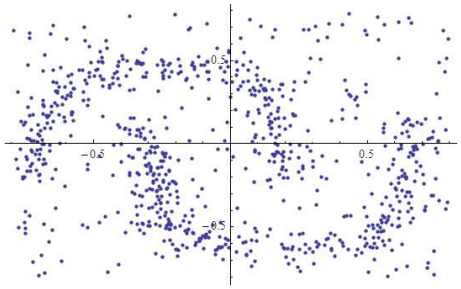

図 1 ：サンプルデータ

7.1 スペクトラルクラスタリング実行

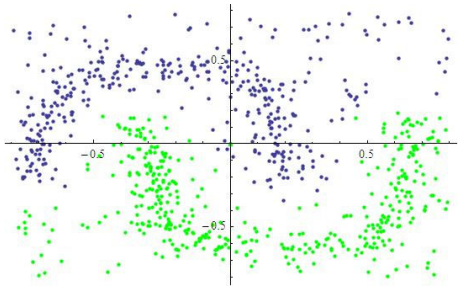

図 2 ：スペクトラルクラスタリングの実行結果

サンプルデータのガウスカーネル行列を計算し, スペクトラルクラスタリングを実行した結果, 図 2 のようにノイズを含め, 非線形に 2 つのクラスに分かれた. このときパラメータ β の値を 100 とした.

7.2 ノイズを抽出

2 つに分かれたクラスタのうちの 1 つのクラスタに対し, データ間距離を計算し, ある閾値より大きいものをノイズとした結果下の図のようになった.

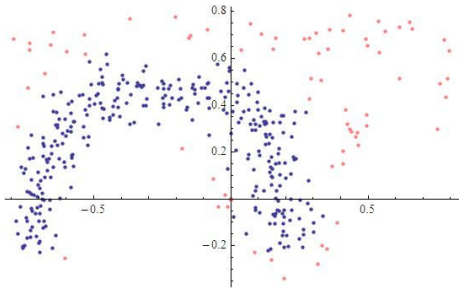

図 3 ：ノイズの抽出結果

この例では 1.000435 より大きいものをノイズとして抽出した. またもう 1 つのクラスタに対しても同じ操作を施す.

8 実験結果

最終的には以下のような良好な判別結果が得られた.

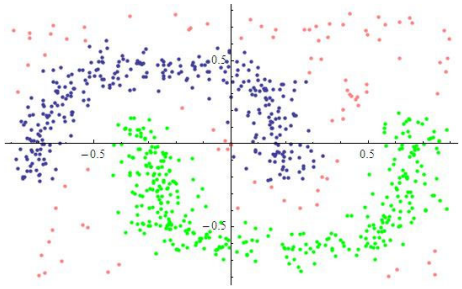

図 4 ：実験結果

9 まとめ

複雑なサンプルデータを非線形にクラス分けすることができ, ある程度上手くノイズを抽出することができた.

しかし, スペクトラルクラスタリングを実行する際に, パラメータ β の値によって結果が大きく左右するので, 適切な β の値を探すのが困難である. また, サンプルデータをどの程度ノイズとするか, つまり閾値をどこに設定するか, という課題が残されている.

参考文献

1. 赤沼昭太郎, 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. 西田英郎, クラスター分析とその応用, 内田老鶴圃, 東京, 1988.
3. 茨木志織, モーメント法によるノイズ推定を用いたスペクトラルクラスタリング, お茶の水女子大学大学院理学専攻情報科学コース修士論文,2011.