Vol.2010-MPS-81 No.39
Vol.2010-BIO-23 No.39
2010/12/17

ランダム行列を用いたスペクトラルクラスタリング

茨　木　志　織[†1]

カーネル法を用いたクラスタリングの 1 つにスペクトラルクラスタリングがある. 本研究では, ノイズを含んだデータに対し, そのノイズをランダム行列の手法を用いて取り除くことにより, スペクトラルクラスタリングの精度を上げる手法を提示する. カーネルにはガウスカーネルを用いて, Wishart 行列の固有値分布とガウスカーネルで写像した特徴空間における内積行列の固有値分布が等価であることを利用し, ノイズを推定する.

Spectral clustering using random matrices

Shiori Ibaraki[†1]

The spectral clustering is known as one of methods for clustering by using kernel technique. In this study, we shall show a method of improving spectral clustering by removing the noise from data with the theory of random matrices. We will use the Gaussian kernel and estimate the noise since the spectral distribution of Wishart matrix is equivalent to one of the matrix constituted from inner products of Gaussian kernel.

1. はじめに

データをまとまりごとに集めてグループ分けをすることは, クラスタリングと呼ばれている. 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, 反復演算を必要とする点と, 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, これらの問題点のないアルゴリズムを構成することが出来る. 本研究ではノイズを含んだデータに対し, そのノイズをランダム行列の手法を用いて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上げる手法を提示する.

†1 お茶の水女子大学大学院　人間文化創成科学研究科
Graduate School of Humanities and Sciences, Ochanomizu University

2. ランダム行列理論

一般に, ランダム行列とは確率変数を要素に持つ行列であり, その代表例として Wishart 行列が挙げられる.

2.1 Wishart 行列

各成分が独立に $N(0,1)$ の標準正規分布に従う変数をもつ $n \times p$ の行列を C とする. このランダム行列 C から

$$S = \frac{1}{n} C\, {}^{T}C \tag{1}$$

で求められる $n \times n$ 対称ランダム行列 S を Wishart 行列という. $p/n = \lambda$ を保ちながら, $n \to \infty, p \to \infty$ の極限をとると, Wishart 行列 S の固有値の経験分布は, $\lambda_{min} \leq t \leq \lambda_{max}$ のときに以下の確率密度関数に収束することが知られている.

$$p(t) = \frac{1}{2\pi} \frac{\sqrt{-(t-\lambda_{max})(t-\lambda_{min})}}{\lambda t} \quad , \quad \lambda_{min}^{max} = (1 \pm \sqrt{\lambda})^2 \tag{2}$$

また, このような確率密度関数を持つ分布は Marcenko-Pastur 分布と呼ばれている.

2.2 ガウスカーネル

変数の集合の二つの要素 x, x' に対し、カーネル関数 $k(x, x')$ は x, x' それぞれの特徴ベクトルどうしの内積

$$k(x, x') = \phi(x)^T \phi(x') \tag{3}$$

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

$$k(x, x') = \exp(-\beta ||x - x'||^2) \tag{4}$$

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することが知られている. ここで, $||\ ||^2$ は通常のユークリッド 2 乗距離で, $\beta \in R$ は適当なパラメータである.

また, Wishart 行列の固有値分布と, ガウスカーネルで写像した特徴空間における内積行列の固有値分布は等価であると知られており, これによりガウスカーネル行列におけるノイズ部に相当する固有値分布も Marcenko-Pastur 分布と同様の性質を持つことがわかる.

Vol.2010-MPS-81 No.39
Vol.2010-BIO-23 No.39
2010/12/17

3. 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では, 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. 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のカットと呼ぶ. グループ内は出来るだけ近いものどうしが集まり, グループ間は遠く離れていることが望ましいので, 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.

$$\min_{\beta} \sum_{i,j} K_{ij}(\beta_i - \beta_j)^2 = 2\ {}^T\beta P\beta, \quad \beta_i = \pm 1 \tag{5}$$

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. そ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.

4. 提 案 手 法

本研究ではランダム行列の理論を用いてスペクトラルクラスタリングを以下のように行うことを提案する.

(1) サンプルデータのガウスカーネル行列 K を構成し, そのスペクトルを求める.
(2) 求めたスペクトル分布と Marcenko-Pastur 分布を比較し, ノイズに相当すると考えられる部分を取り除く. その残ったスペクトルを用いて, カーネル行列 K' を再構成する.
(3) この新たな K' を与えられたカーネル行列と見なし, スペクトラルクラスタリングを行う.

5. 実 験 例

図 1 のような, 線形で分けることの出来ない 2 群データを用意する. 1 群が 300 点, 2 群が 250 点の, 合計 550 点のサンプルデータとなっている. これらをスペクトラルクラスタリングで 2 つにクラス分けする.

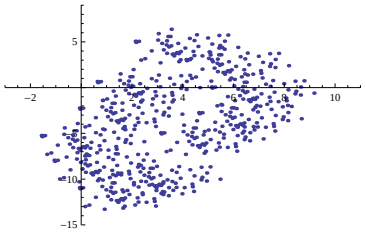

図 1 サンプルデータ
Fig. 1 sample data

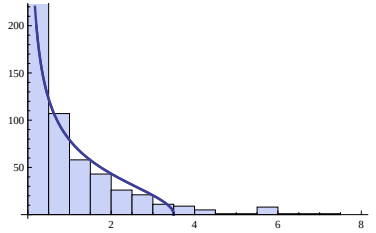

図 2 ヒストグラムと Marcenko-Pastur 分布
Fig. 2 histogram and Marcenko-Pastur distribution

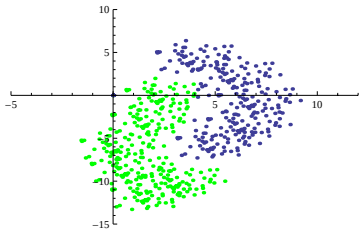

図 3 ノイズを除いた場合
Fig. 3 with reduction

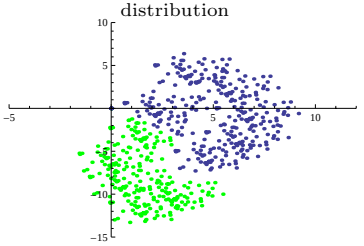

図 4 除いていない場合
Fig. 4 without reduction

5.1 ノイズ部スペクトルの推定

サンプルデータのガウスカーネル行列のヒストグラムと Marcenko-Pastur 分布を重ね合わせると, 図 2 のようになった. この分布から外れた部分がこのデータにおける取り出すべき情報, つまりデータからノイズを除いた構造部であると考えた. 今回の実験では, この構造部のスペクトルは 27 個であり, このスペクトルのみを用いてカーネル行列を再編成し, スペクトラルクラスタリングを行った.

5.2 結　　果

図 3 の結果が得られた. 比較の為, 右に通常のスペクトラルクラスタリングの結果 (図 4) を並べた. 図 3 は, ほぼ設定した通りにクラス分けが出来ていることが読み取れる.

6. ま と め

ランダム行列理論を用いて、ノイズを除去することで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確認された. また今後の課題として, Marcenko-Pastur 分布の適切なパラメータの推定を行いたいと考えている.