

ランダム行列理論を用いた大量文書クラスタリング

郷治 雅[†] 小林 一郎[‡]

[†]お茶の水女子大学理学部情報科学科

[‡]お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース

1 はじめに

近年、ビッグデータなどと呼ばれる高次元で巨大なデータが増えている。これらのデータを効率良く処理することが今後必要不可欠であると考えられる。本研究では高次元データに対して有用であることが知られているランダム行列理論を用いて大量文書のクラスタリング精度を高めることを目的とする。観測されたデータに含まれているノイズ、つまりデータ内の不要な部分を取り除き、新たに重要部分のデータを再構築する手法である。このデータを用いてスペクトラルクラスタリングを行い、ノイズ除去の有無における精度を比較する。

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 行列が挙げられるが、本研究ではガウスカネル行列を用いる。 $n \times n$ 対称ランダム行列 S で $p/n = \lambda$ を保ちながら、 $n \rightarrow \infty$, $p \rightarrow \infty$ の極限を取ると、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 $p(t)$ に収束する。この分布は Marchenko-Pastur 分布と呼ばれている。

$$p(t) = \frac{1}{2\pi} \frac{\sqrt{-(t - \lambda_{max})(t - \lambda_{min})}}{\lambda t}$$

$$\lambda_{min}^{max} = (1 \pm \sqrt{\lambda})^2$$

2.1 ガウスカネル

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同士の内積として定義される。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することが知られている。ここで、 $\|\cdot\|^2$ は通常のコークリッド 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ネル行列におけるノイズ部に相当する固有値分布も Marchenko-Pastur 分布と同様の性質をもつことがわかる。

3 モーメント法

$f(x)$ を連続確率変数 X の密度関数とすると、原点まわりの k 次モーメント m_k は、

$$m_k = \int_{-\infty}^{\infty} x^k f(x) dx$$

と表せ、これらの値は確率分布の特徴を与える。また、先にランダム行列理論で述べた Marchenko-Pastur 分布のモーメントは、

$$m_k = \frac{(2k)!}{k!(k+1)!} m_1^k$$

で与えられる。本研究では、観測データからの標本モーメント列を理論値と比較することにより、最適なノイズの推定を行う。目視でスペクトルを比較するより、定量的な推定が行える。

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では、サンプル点をグラフ構造として考え、各頂点をサンプル点、枝にはサンプル同士の類似度を表す重みがついているとする。例えば、サンプル点を二つのグループに分けると、それに伴いグラフも二分割される。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、以下のようなになる。

$$\min_{\beta} \sum_{i,j} K_{ij} (x_i - x_j)^2 = 2^T X P X, \quad x_i = \pm 1$$

ここでは、 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。 X は 2 値ベクトルという制約が

Large Document Set Clustering using Random Matrix Theory

[†]Miyabi GOUJI(g1020511@is.ocha.ac.jp),

[‡]Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

[†]Dept. of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

[‡]Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

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， $^T X \Lambda X = 1$ 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 番目以降の固有値ベクトルの成分符号に基づいてクラスタリングを行う．

5 提案手法

本研究では，以下の手続きでスペクトラルクラスタリングを行うことを提案する．

1. 文書データから単語の出現頻度に基づく文書行列を作成する．
2. 文書行列からガウスクーネル行列 K を構成し，その固有値分布を求める．
3. Marchenko-Pastur 分布のモーメント列に最も適合するように，2 で求めた固有値分布からモーメント法を用いてノイズ部と構造部を推定する．
4. 構造部の固有値のみを用いてカーネル行列 K' を再構成する．
5. この新たな K' を与えられたカーネル行列とみなし，スペクトラルクラスタリングを行う．

6 実験

6.1 実験設定

20Newsgroups[1] 中の 4 つのカテゴリから 500 文書ずつ抽出し，計 2000 文書を用いてクラスタリングを行う．このときガウスクーネルにおける β の値は 5×10^{-3} を用いることとする．

6.2 実験結果および評価

ガウスクーネルの固有値分布とそのモーメントを調べた結果を表 1 に，固有値分布を図 1 に示す．

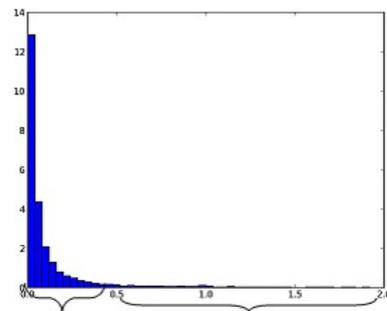
表 1: 固有値のモーメント

	1 乗	2 乗	3 乗	4 乗	5 乗	6 乗
理論値	1	2	5	14	42	132
モーメント	1	1.85	4.59	13.13	40.67	132.22

固有値 1652 個のモーメント値が理論値に適合したため，残りの 348 個の固有値を用いてクラスタリングを行った．

クラスタリング結果を以下に示す相互情報量を用いて評価する．

$$MI(L, A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i, \alpha_j) \log s \frac{P(l_i, \alpha_j)}{P(l_i)P(\alpha_j)}$$



(1652 個のノイズ) (348 個の構造部)

図 1: 固有値分布によるノイズ判定

ここで， L および A は分類結果と正解を示し， l_i および α_j はそれぞれ分類結果のカテゴリ，正解カテゴリを指す．得られた情報量を $[0,1]$ の範囲になるように正規化を行った結果を表 2 に示す．

表 2: 相互情報量による精度評価

	ノイズを除かない場合	提案手法
MI	0.0462	0.0466

6.3 考察

ノイズを除いていないデータと比較して微小ではあるが結果が改善した．しかし，提案手法のクラスタリングの精度は低く，その原因としてガウスクーネルを作る際のパラメータの調整が考えられる．また，クラスタリング方法も検討が必要である．

7 おわりに

本研究では，ランダム行列理論を用いてデータ内に含まれるノイズを削減し，文書クラスタリングに応用した．ランダム行列理論は高次元データに対して有用であるため，より大きなデータを使用して実験を行い精度を比較する必要がある．今後は，更なる高次元データに対して実験を行い，パラメータの値を検討していく．更に，様々なクラスタリング手法を試し，精度の向上を目指す．

参考文献

- [1] <http://qwone.com/~jason/20Newsgroups/>
- [2] 茨木志織，吉田裕亮，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．
- [3] 相馬亘，藤原 義久，尹 熙元，経済物理とランダム行列 - 株式市場にある本質的な構造の抽出 - ，特集：「ランダム行列の広がり」 - その多彩な応用 - ，数理科学 2007 年 2 月号 No.524，2007．
- [4] 伊藤里江，ランダム行列理論を用いた Gaussian カーネルにおける雑音の推定，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，2009．
- [5] 赤穂昭太郎，カーネル多変量解析，岩波書店，2008．