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　(指導教員：小林 一郎)

1　はじめに

近年，ビッグデータ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，これまで文書分類においても大量データに対する様々な手法が提案されている．

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．具体的には，観測されたデータに含まれているノイズ，つまりデータを分析する上で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．それにより，得られたデータからカーネル行列のスペクトル分布を調べ，ランダム行列理論においてノイズが従うとされている分布のモーメントと照らし合わせることによって定量的にデータの構造部を推定する．そして，構造部のみを用いて新たな行列を作成し，スペクトラルクラスタリングを用いてクラスタリングを行う．ノイズ除去の有無における精度を比較することにより提案手法の有効性を検証する．

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，$n \times n$ 対称ランダム行列 S で $p/n = \lambda$ を保ちながら，$n \to \infty$，$p \to \infty$ 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 Marchenco-Pastur 分布と呼ばれている．

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

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

3　ガウスカーネル

対象とするデータを文書データとするため，本研究では Wishart 行列の代わりにガウスカーネル行列を用いる．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

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

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

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

を用いて特徴空間に写像した行列

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

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 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 Marchenco-Pastur 分布と同様の性質をもつことがわかる．

4　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き，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．

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，以下のようになる．

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

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とこになる．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．

6 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カーネル行列 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．

7 実験

7.1 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．このときガウスカーネルにおける β の値は 2×10^{-3} を用いることとする．

7.2 実験結果および評価

ガウスカーネル行列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

固有値の小さい順に 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．

$$MI(L,A)=\sum_{l_i\in L,\alpha_j\in A}P(l_i,\alpha_j)\log_2\frac{P(l_i,\alpha_j)}{P(l_i)P(\alpha_j)}$$

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
$\overline{MI}$	0.0462	**0.0466**

7.3 考察

ノイズを除いていないデータと比較して，結果は多少改善しているように見えるが、誤差の範囲に過ぎず，文書データにおいてノイズを除くことで良いクラスタリング結果が得られるかは明確に判断するに至らなかった．また，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カーネルを作る際のパラメータ β の調整が考えられる．さらに，データ数によって精度が変化するなどが複数回の実験を通して見受けられた．データが少ない場合はスペクトラルクラスタリングそのものの精度が高いため，ノイズの除去を行う利点はなく，データ数を増やすとより理論値に合う固有値の分布を見つけることができ，ノイズの除去を明確に行うことができたが，クラスタリングの精度は低くなった．今回，クラスタリング手法としてスペクトラルクラスタリングのみを用いたが，様々な手法を試し，精度の向上を目指す必要があると考える．

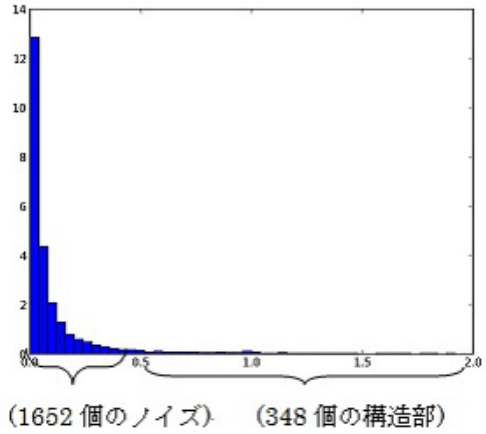

図 1: 固有値分布によるノイズ判定

8 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．実験によって多少分類精度向上は見受けられたが，今回用いた文書サイズではランダム行列理論を十分に生かし切れなかった可能性がある．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，ガウスカーネルにおけるパラメータ値の設定方法について検討を重ねていくつもりである．

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.