スペクトラルクラスタリングにおけるクラスタ数決定手法の提案

矢部　大輔†　　木村　昌臣‡

芝浦工業大学大学院工学研究科†　　芝浦工業大学工学部情報工学科‡

1. はじめに

グラフ構造を分割する手法としてスペクトラルクラスタリングがあげられる．スペクトラルクラスタリングとは Normalized Cut 等の手法でグラフを分割する問題を行列の固有値問題に置き換える手法である．この分割を繰り返し適用することでクラスタを求める．しかしスペクトラルクラスタリングの問題点として，最適な分割数が事前に与えられる必要があるという問題がある．そこで本研究では，赤池情報量規準(Akaike Information Criterion:AIC)を用いてクラスタリングの評価を行うことにより分割数を決定する手法を提案する．

2. スペクトラルクラスタリング

Normalized Cut を用いてグラフを分割する場合，次のように計算を行う．グラフのノード集合 V を2つのグループ A,B に分類するとき，あるノード u,v の間でのエッジの重みを $w(u,v)$ としたときグループ間の類似度を次のように表すことができる．

$$cut(A,B)=\sum_{u\in A,v\in B} w(u,v) \quad (1)$$

また，あるグループ G 内の類似度を

$$assoc(G,V)=\sum_{g\in G,t\in V} w(g,t) \quad (2)$$

とすると，Normalized Cut の重みは

$$Ncut(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)} \quad (3)$$

と表される．この式を最小化することはグループ内の類似度を大きく，グループ間の類似度は小さくするカットとなる．これは一般化固有値問題に帰着することで最小化できることが示されている[1]．データ数が x のとき，W を $x\times x$ の隣接行列，D を W の次数を対角成分に持つ行列とする．このとき $D^{-1/2}(D-W)D^{-1/2}$ の固有値の大きさが $Ncut$ の値となり対応する固有ベクトルがグラフの分割を与える．ただし，最小固有値は0となるので0より大きい固有値をクラスタリングに用いる．小さい方から i 番目の固有値(0 を含む)を第 i 固有値と呼ぶ．得られた固有ベクトルの各成分が対応するノードを識別する値を表すため，同じノードに対応する成分を並べて得られるベクトルに対し K-means 法によってクラスタリングすることでスペクトラルクラスタリングは実現される．

3. スペクトラルクラスタリングの問題点

クラスタリングに用いるベクトルの次元を既存手法では固有値の大きさから決めていたが，それでは対処できないグラフが存在する問題がある．スペクトラルクラスタリングでは固有値が $Ncut$ の値となるため，$Ncut$ を小さくするグラフ構造の分かれ目の数だけ低い値の固有値が存在する．それ以降急激に増大するのは分かれ目以外の部分をカットしようとしたため $Ncut$ の値が大幅に増加したからである．$Ncut$ の値を低くする固有ベクトルだけをクラスタリングに用いればよいので，急激に大きくなる直前までの固有ベクトルを用いればよい．図1にクラスタ間のエッジが少ない明確に分かれ目の見えるグラフ，図2に明確に分かれ目が見えないグラフとそれぞれの $D^{-1/2}(D-W)D^{-1/2}$ の固有値の大きさを縦軸に，横軸に第 n 固有値をとりグラフにした．

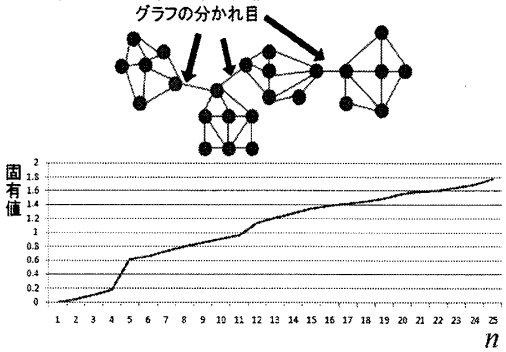

図1. 分かれ目のあるグラフでの固有値

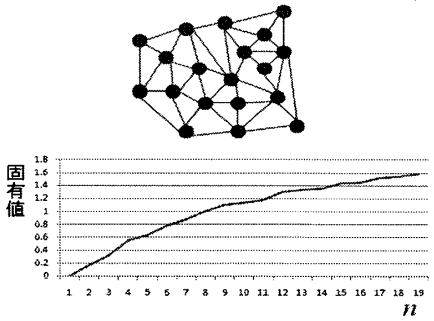

図2. 分かれ目のないグラフでの固有値

図1では固有値は第五固有値から急激に上昇している．しかし，図2では，第二固有値以降大きな増加は見られない．これはクラスタ構造の明確な分かれ目がみられず，グラフ分割の際に切るエッジの数が増えて $Ncut$ の値が大きくなってしまうため固有値の低いものは存在しなくなってしまうからである．これにより固有値が急増しなくなるため既存手法では対処できない．

4. 提案手法

スペクトラルクラスタリングにおいて，クラスタ数とノードを識別するベクトルの次元を決定する手法を提案する．まずグラフの隣接行列を作成し，$D^{-1/2}(D-W)D^{-1/2}$ の固有ベクトルを求める．次にクラスタ数とデータの次元を変更して K-means 法によるクラスタリングを行い，それぞれのクラスタリング結果の AIC を求める．AIC とはモデルが真の確率分布からどれだけ離れているかを測るカルバック・ライブラー情報量を最小にすることを目的とし

The estimation method of the number of cluseters for spectral clustering techiniques
†Yabe Daisuke
†Graduate school of Shibaura Institute Technology
‡Masaomi Kimura
‡Shibaura Institute Technology

て導出されたモデルの評価基準である．スペクトラルクラスタリングは最終的にK-means法によってクラスタリングを行うので，K-means法について提案されている尤度[2]を用いる．AICの定義式にあてはめると次のようになる．

$$AIC = N\left\{1+\ln\left[\frac{2\pi}{N}\sum_{k=1}^{K}\sum_{x\in C_k}(x-X_k)^2\right]\right\}+2KN \quad (4)$$

Kをクラスタ数，Nをノードを識別するベクトルの次元，C_iをi番目のクラスタ，X_iをi番目のクラスタの重心とする．またNは，第N固有ベクトルまでをクラスタリングに用いることを示す．

ノードを識別するベクトルの次元を増やしていくと，固有値の大きい固有ベクトルも用いることになる．Ncutの値が大きくグラフ分割には適さないカットをクラスタリングに用いてしまい，尤度を上げてしまう．その結果ノードを識別するベクトルの次元を増やしていくとAICの値はクラスタ数を変えても増加しつづけてしまう．そこでAICが増加し続ける直前のノードを識別するベクトルの次元におけるAICを最小にするクラスタ数を本研究で求めるクラスタ数とする．

5. 実験

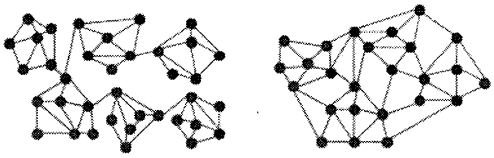

図3.実験対象とする2つのグラフ構造

図3に示す明確に分かれ目の見えるグラフ(ノード数37)と明確に分かれ目の見えないグラフ(ノード数25)を実験対象とし，提案手法を用いてスペクトラルクラスタリング適用時のクラスタ数を求め，クラスタリング結果を調べる．また既存のクラスタリング手法であるReichardtらのspinglassモデルによるクラスタリング手法[3]との比較を行う．これは統計解析ソフトRのパッケージigraphの中のspinglass.community関数を利用した．グラフのエッジの重みは全て1とした．

隣接行列を作成し，$D^{-1/2}(D-W)D^{-1/2}$の固有ベクトルを求める．クラスタ数Kとノードを識別するベクトルの次元Nをそれぞれ2～ノード数/3，1～ノード数/3までの範囲で値を変え，それぞれクラスタリング結果のAICを求める．クラスタの最小ノード数を3以上と考え，クラスタ数の上限をノード数/3と仮定した．ただし，クラスタのノード数が1の場合尤度が$-\infty$となるのでそのクラスタの尤度を0とする．クラスタリング結果を図4に示す．明確に分かれ目が見えるグラフではノードを識別するベクトルの次元7，クラスタ数7，明確に分かれ目の見えないグラフではノードを識別するベクトルの次元4，クラスタ数4であった．図4より，明確に分かれ目の見えないグラフでは同じ結果であったが，明確に分かれ目の見えるグラフでは一つのクラスタになるところが二つに分かれてしまいspinglassの方が良い結果となった．しかし，ノードを識別するベクトルの次元6のときクラスタ数6でAICが最小になりspinglassと同じ結果になった．よってAICにより得られたクラスタ数に誤差があったことが分かる．

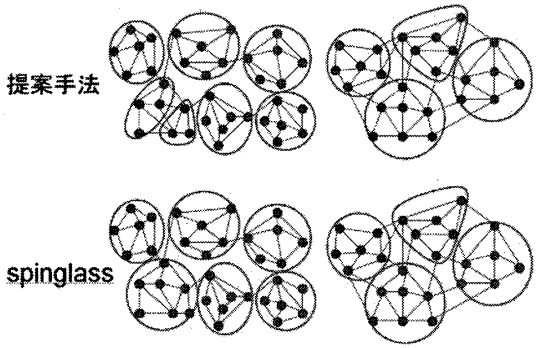

図4.クラスタリング結果

提案手法ではAICの値を求める際にK-means法を繰り返し行うため計算量が多くなる問題がある．スペクトラルクラスタリングでは一般的に$k=N$に近くなるときに良い結果が得られるとされ，実験においても次元数がクラスタ数に近い値となり，その付近でのAICは低くなる傾向があった．このことを利用し，それぞれの次元において2～ノード数/3まで計算していたのを$k=2$，$k=N$のときだけを計算して増減を調べるようにしたところ，クラスタリング結果はそのままに計算量のみを減らすことができた．100回の平均実行時間は明確に分かれ目の見えるグラフでは提案手法は19ミリ秒，spinglassは1607ミリ秒，明確に分かれ目の見えないグラフでは提案手法は15ミリ秒，spinglassは1490ミリ秒となり提案手法の方が早くなった．ノード数が増えた時を調べるためノード数76のグラフでの実行時間を調べたところ，提案手法は59ミリ秒，spinglassは3001ミリ秒となりノードが増えたときも実行時間は少ない．処理時間の増加の割合が提案手法の方が大きいのはK-means法による処理の増加によるものである．

他のグラフ構造での結果を図5に示す．

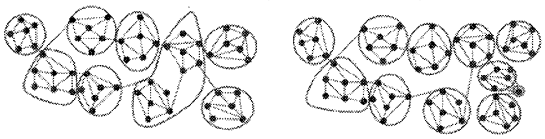

図5.その他のグラフでのクラスタリング結果

6. まとめ

スペクトラルクラスタリングは自動的にクラスタ数を与えないので，本研究ではAICを用いてクラスタリング結果の評価を行いクラスタ数を決定する手法を提案した．問題点としてAICの尤度の改善があげられる．今後改善が行われ，画像分割やデータマイニングに応用されることが期待される．

参考文献

[1] Jianbo Shi, Jitendra Malik:Normalized Cuts and Image Segmentation, IEEE TRANSACTIO NS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8 , pp888-905(2000)

[2] 石岡 恒憲，クラスター数を自動決定するk-meansアルゴリズムの拡張について，応用統計学，Vol.29, No.3, pp.141-149, (2000)

[3] Joerg Reichardt, Stefan Bornhold:Statistical mechanicsof community detection, Physical Review E, vol. 74, 016110, pp.1-14 (2006)