縮約類似度行列を用いたスペクトラル手法による クラスタリング結果の改善

Refinement of Clustering Results by Spectral Method Using the Reduced Similarity Matrix

新納浩幸[1*]　佐々木稔[1]
Hiroyuki Shinnou[1] and Minoru Sasaki[1]

[1] 茨城大学 工学部 情報工学科
[1] Department of Computer and Information Sciences, Ibaraki University

Abstract:
Spectral clustering is actually powerful, but needs to solve the eigenvalue problem of the Laplacian matrix converted from the similarity matrix corresponding to the given data set. Therefore, we cannot use spectral clustering for a large data set. In this paper, we propose the method to reduce the similarity matrix. use spectral clustering for a large data set. However our methods needs a clustering result for that reduction. Therefore, our method is regarded as the method to improve the given clustering result.
In the experiment, we used seven data sets to evaluate our method. We compared our method with k-means and spectral clustering, Mcut. The experiment showed our method improves the clustering result generated by k-means.
In future we will investigate the proper reduction degree, and improve the similarity definition.

1　はじめに

スペクトラルクラスタリングは精度の高いクラスタリング手法である。しかしそこでは類似度行列のサイズの固有値問題を解く必要があり、データ数が多い場合、現実的には利用できない。本論文では縮約された類似度行列（縮約類似度行列）を作成することで、対象データセットに対するスペクトラルクラスタリングを行う。ただしここで提案する縮約類似度行列の作成方法は、ある程度妥当なクラスタリング結果を必要とする。そのため提案する縮約類似度行列を利用してスペクトラルクラスタリングを行うことは、最初に得られたクラスタリング結果を改善するという位置づけになる。

スペクトラルクラスタリングは、グラフ分割の観点からクラスタリングを行う手法であり、その精度の高さから近年活発に研究されている [8][3]。そこではグラフの最適な分割を求めるために、ある評価関数を設定する。この評価関数の最適解がある固有値問題の解に対応することを利用して、クラスタリングを行うのがスペクトルクラスタリングである。

スペクトラルクラスタリングの精度は高いが、実際の処理には類似度行列のサイズ（つまりデータ数の自乗のサイズ）の固有値問題を解く必要がある。このためデータ数が大きい場合、スペクトラルクラスタリングは現実的には利用できない [2]。ここではデータ数が概ね 10,000 以下のデータセットを想定しているが、この程度のデータ数でもスペクトラルクラスタリングを利用するのは無理がある。

本手法の概略を述べると、まずデータセットを k-means でクラスタリングし、そこから各クラスタの重心に近いデータだけをまとめて、１つのデータと見なす。１点にまとめられるデータの集合を、論文 [7] に従って、ここでは Committee と呼ぶ。Committee 以外のデータはそのまま 1 点となる。それらデータに対して類似度行列を作成する。この際、Committee の大きさ分だけデータサイズが縮約されるので、類似度行列も縮約される。この縮約された類似度行列をここでは縮約類似度行列と呼ぶ。

本手法の主張点はこの縮約類似度行列の作成方法にある。Committee が真に同じクラスタに属するデータで構成されることを仮定すれば、データ間の類似度は、

*連絡先：茨城大学工学部情報工学科
〒 316-8511 茨城県日立市中成沢 4-12-1
shinnou@mx.ibaraki.ac.jp

スペクトラルクラスタリングで利用する類似度行列の拡張になるように設定できる。本論文ではこの点をまず示す。ただし上記の仮定は実際には成立していないので、類似度を近似する必要がある。その近似式を提案する。

以上により、縮約類似度行列が作成でき、それを用いてスペクトラルクラスタリングを行う。縮約類似度行列を作成するために、k-means 等でクラスタリングを既に行っているので、スペクトラルクラスタリングにより得られたクラスタリング結果は、最初のクラスタリング結果の改良と見なされる。

実験では 7 つの文書データセットを用いて、本手法の有効性を示す。本手法により大規模なデータに対してもスペクトラルクラスタリング手法を適用することが可能になる。適切な縮約の度合いの推定と近似式の改良が今後の課題である。

2 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、データをグラフのノードとして表現し、ノード間のエッジの重みには両端のデータ間の類似度を与える。類似度が 0 の場合は、エッジを張らない。このようにデータの集合をグラフとして表した場合、クラスタリングとはエッジをカットして、全体のグラフをいくつかのサブグラフに分割することに対応する。その際に、サブグラフ内のエッジは密になり、サブグラフ間でカットしたエッジは疎になるようなカットが望ましい。望ましいカットを見つけるために，評価関数を設定する。この評価関数の最適解がある固有値問題の解に対応することを利用して，クラスタリングを行うのがスペクトラルクラスタリングである。評価関数はいくつか提案されているが、ここでは Mcut [3] で提案されているものを利用する。

まずサブグラフ A と B の類似度 $cut(A,B)$ を以下で定義する。

$$cut(A,B) = W(A,B) \tag{1}$$

ここで関数 $W(A,B)$ はサブグラフ A と B 間にあるエッジの重みの総和である。エッジの重みはノード（データ）間の類似度を表すので、結局、関数 $W(A,B)$ はサブグラフ A と B の類似度を表している。また，$W(A) = W(A,A)$ と定義しておく。

Mcut の評価関数は以下である。この式を最小化するようなサブグラフ A と B を見つけることが課題である。

$$Mcut = \frac{cut(A,B)}{W(A)} + \frac{cut(A,B)}{W(B)} \tag{2}$$

スペクトラルクラスタリングは 2 つのクラスタに分割するのが基本である。目的のクラスタ数を得るまで、上記の処理を再帰的に繰り返す。

式 (2) の最小化の問題は、以下の式を最小化するベクトル y を求める問題と等価である。

$$J_m = \frac{y^T(D-W)y}{y^T W y} \tag{3}$$

ここで W はデータ間の類似度行列、また $D = diag(We)$ である。e は $e = (1,1,\cdots,1)^t$、$diag$ は対角要素の行列を意味する。つまり、

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_N \end{pmatrix}$$

であり、d_i は i 番目のデータとその他のデータ（i 番目のデータも含む）との類似度の和である。つまり d_i はノード i の degree である。

また q は N 次元ベクトルであり、各要素は a か $-b$ の離散的な値を取る。i 番目のデータがクラスタ A に含まれるなら、q の i 番目の要素は a、クラスタ B に含まれるなら $-b$ を取る。ここで $a = \sqrt{\frac{d_B}{d_A d}}$、$b = \sqrt{\frac{d_A}{d_B d}}$ であり、$d_X = \sum_{i \in X} d_i$ を意味し、$d = d_A + d_B$ である。

次に式 (3) の最小化問題を解くために、y を連続値の要素をとるベクトルと考える。このとき式 (3) の最小化問題は、以下の固有値問題を解くことで得られる。

$$I - D^{-1/2} W D^{-1/2} \tag{4}$$

この最小の固有値に対する固有ベクトルが式 (3) を最小にするが、最小の固有値は 0 であり、その対応するベクトルは $u_1 = D^{1/2}e$ である。これは我々には意味がないので、ここから 2 番目に小さな固有値に対応する固有ベクトル u_2 を求め、それを近似解とする。このベクトルは Fielder ベクトルと呼ばれる。

次に $\hat{q} = D^{-1/2}u_2$ から $\hat{q}$ を得る。次に $\hat{q}$ の要素をソートして、ある値以上をクラスタ A に、それより小さい部分をクラスタ B に対応させることでクラスタリングが行える。実際には、$\hat{q}$ の要素が正ならクラスタ A に、負ならクラスタ B に対応させる簡易な方法でも概ねうまくゆく。

3 縮約類似度行列

データセット X のデータ数が N のとき、X に対する類似度行列 W は $N \times N$ のサイズとなる。スペクトラルクラスタリングを用いて、X がクラスタ A とクラスタ B に分割されたとする。このときこの分割によって式 (2) の値が最小になる。

そしてクラスタ A のある部分集合 A' を 1 点 a' で表すことを考える。

$$A' = \{a_1, a_2, \cdots, a_m\} \subset A$$

まず a' とクラスタ B 内の点 b との類似度を以下で定義する。

$$sim(a', b) = \sum_{i=1}^{m} sim(a_i, b) \tag{5}$$

次に a' とクラスタ A 内の点で A' に含まれていない点、つまり $A - A'$ 内の点 a との類似度を以下で定義する。

$$sim(a', a) = \sum_{i=1}^{m} sim(a_i, a) \tag{6}$$

最後に a' と a' の類似度を以下で定義する。

$$sim(a', a') = \sum_{i=1}^{m} \sum_{j=1}^{m} sim(a_i, a_j) \tag{7}$$

式 (5)、(6)、(7) を用いて、X 中の A' を 1 点 a' で表し、その結果できるデータセットを X' とする。この X' に対する縮約類似度行列 W' を作成する。このとき W' のサイズは $(N - m + 1) \times (N - m + 1)$ に縮約されている。

X に対するクラスタリング結果を X' のクラスタリング結果に当てはめる。つまり A' 以外のデータのクラスタリング結果は X のクラスタリング結果と同じであり、データ a' のクラスタは A と考える。そして W' を用いて、式 (2) の値を計算する。するとこの値は、X に対するクラスタリング結果を W を用いて計算した式 (2) の値と同じであることが容易に確認できる。

X に対してはクラスタ A とクラスタ B の分割によって、式 (2) の値が最小になることを考えれば、縮約類似度行列 W' を用いて、スペクトラルクラスタリングを行った場合でも先と同じクラスタ A とクラスタ B の分割が得られる。

以上より式 (5)、(6)、(7) を用いて、クラスタ内の一部のデータを縮約した縮約類似度行列が作成可能である。これによって行列のサイズが小さくなり、スペクトラルクラスタリングが可能となる。

ただし現実的にはクラスタ A とクラスタ B の分割が不明であるので、縮約類似度行列 W' を用いてスペクトラルクラスタリングを行っても妥当な結果は得られない。この問題への対処を以下に説明する。

3.1 荒いクラスタリングによる Committee の作成

縮約類似度行列を作成するためにはデータセット X に対する正しいクラスタリング結果が必要である。しかし一般に正しいクラスタリング結果を得ることはできない。

ここでは縮約するデータの集合はクラスタ全体でなくても良いことに注目する。つまり先の A' は A の部分集合であり、A そのものである必要はない。必要なことは A' 内のデータが同じクラスタに属することである。このような A' をここでは、論文 [7] に従って Committee と呼ぶ。

そこで本論文では、まず k-means などの荒いクラスタリング手法を用いることで、あるクラスタリング結果を得る。次に、そのクラスタリング結果の信頼性の高い部分を選出することで、Committee を作成する。具体的には各クラスタの重心から近い距離のデータは、そのクラスタに真に属すると考え、Committee のメンバと見なすことにする。

ここでどの程度、重心から近ければ Committee のメンバとするかによって、縮約の程度が変化する。例えば、重心から近い順に上位 9 割を Committee のメンバと考えれば、データ数を約 9 割カットした縮約が可能となる。ただしその一方で実際には同じクラスタに属さないデータも Committee に属するようになる。その部分の誤りは以後の処理でも引き継ぎ、結果として、最終的な精度が低くなる。

3.2 重心を利用した類似度の近似

Committee 群が作成できた場合、各 Committee を縮約した縮約類似度行列を作成するためには式 (5)、(6)、(7) の計算を行わなくてはならない。しかし現実的には Committee 内に誤りが含まれるために、そのままの式を使うと誤りが増大してしまう。

このために本論文では、A' の重心 $\bar{a}$ を利用して、以下の近似式を用いる。

$$sim(a', b) \approx m \cdot sim(\bar{a}, b) \tag{8}$$

$$sim(a', a) \approx m \cdot sim(\bar{a}, a) \tag{9}$$

$$sim(a', a') \approx m/2 \tag{10}$$

4 クラスタリングの手順

ここでは提案するクラスタリングの手順をまとめる。

step 1. データセット X と目的とするクラスタ数 K を与え、k-means によりクラスタリングを行う。

step 2. step 1. によって得られた各クラスタに対して、そのクラスタの重心を求め、重心から距離の

近い順にそのクラスタの 8 割のデータを取り出し、それをそのクラスタの Committee とする。

step 3. step 2. によって作成された Committee は K 個あるが、それぞれを 1 点に縮約し、新たなデータセットを作成する。そのデータセットに対して、式 (8)、(9)、(10) を用いて縮約類似度行列 W' を作成する。

step 4. W' に対してスペクトラルクラスタリングを行い、得られたクラスタリング結果中の 1 点に縮約されたデータをもとに戻すことで、全体のクラスタリング結果を得る。

step 4. には注意が必要である。1 点に縮約されたデータが複数存在するが、それらはスペクトラルクラスタリングによってそれぞれ別個のクラスタに割り当てられなくてはならないからである。

通常、1 点に縮約されたデータは別個のクラスタに割り当てられるはずであるが、1 点に縮約されたデータどうしが同じクラスタに割り当てられることも考えられる。そのため、ここではスペクトラルクラスタリングで 2 分割を行う際に、1 点に縮約されたデータが少なくとも 1 点クラスタに含まれるように調整する。

具体的には Fielder ベクトルをソートして、ある点でカットしたときに、片側だけに縮約されたデータが集まった場合は、反対側のクラスタに最も近い縮約されたデータを 1 点だけ反対側のクラスタに移動させる。この処理によって最終的なクラスタリング結果には、各クラスタに 1 点だけ縮約されたデータが入ることになる。

5 実験

提案手法の効果を確認するために、 CLUTO のサイト[1]で提供されている 7 つのデータセット (tr12, tr31, mm, la12, sports, ohscal, cacmcisi) を用いる。これらのデータセットのデータ数、次元数、非ゼロ要素数、クラスタ数を表 1 にまとめる。

表 1: データセット

データ名	データ数	次元数	非ゼロ要素数	クラス数
tr12	313	5804	85640	8
tr31	927	10128	248903	7
mm	2521	126373	490062	2
cacmcisi	4663	41681	83181	2
la12	6279	31472	939407	6
sports	8580	126373	1107980	7
ohscal	11162	11465	674365	10

[1]http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

まず最初に k-means によりクラスタリングを行う。得られた各クラスタからそのクラスタの重心までの距離を利用して距離の近い上位 8 割を Committee とした。次に式 (8)、(9)、(10) を用いて、縮約類似度行列を作成した。

作成された縮約類似度行列を用いて、スペクトラルクラスタリングを行い、最終的に得られたクラスタリング結果をエントロピーと純度で評価した結果が表 3 と表 4 である。当然、本手法はスペクトラルクラスタリング (Mcut) の結果よりも精度は悪いが、k-means よりも精度は改善されている。また cacmcisi、la12、sports、ohscal の 4 つのデータセットに対してはデータ数が大きく、スペクトラルクラスタリングを行うには現実的には不可能である。

表 2: 実験結果（エントロピー）

データ名	Mcut	k-means	本手法
tr12	**0.3800**	0.4366	0.3840
tr31	**0.2946**	0.3419	0.3414
mm	**0.9715**	0.9847	0.9837
cacmcisi	—	0.6768	0.6744
la12	—	0.4523	0.4575
sports	—	0.3142	0.3049
ohscal	—	0.5678	0.5722

表 3: 実験結果（純度）

データ名	Mcut	k-means	本手法
tr12	**0.7061**	0.6550	0.6741
tr31	**0.8037**	0.7605	0.7702
mm	**0.5799**	0.5601	0.5688
cacmcisi	—	0.6869	0.6869
la12	—	0.7015	0.7019
sports	—	0.7735	0.7871
ohscal	—	0.5434	0.5440

6 考察

6.1 縮約の度合い

実験では縮約の度合いを 8 割減になるようにとった。この 8 割というのは適当である。データセット mm に対して、この縮約の度合いを 1 割づつ変えながらエントロピーと純度を調べた。実験結果を図 1 と図 2 に示す。

100% 縮約を行うと、それは k-means の結果をそのまま使うことになる。これがグラフの最も左側の位置を表す。また全く縮約を行わない場合、それは全データを対象にしてスペクトラルクラスタリング（Mcut）を行うことを意味する。これがグラフの最も右側の位置を表す。

理論的には縮約の度合いが大きいほど、結果は k-means に近くなり、縮約の度合いが小さいほど、結果は Mcut に近くなり、しかもその間は単調に増減する

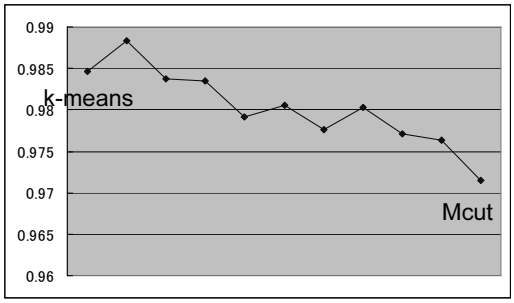

図 1: 縮約に対するエントロピーの変化

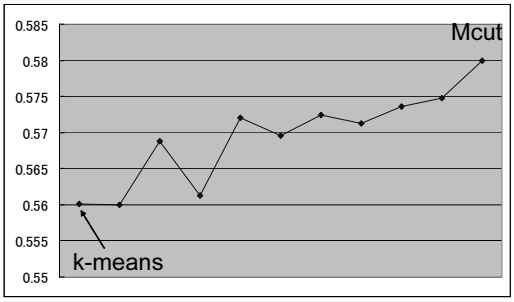

図 2: 縮約に対する純度の変化

はずである。図 1 と図 2 に単調性はないが、その傾向があることは確認できる。

実際に単調にならないのは、Committee 内に誤りが含まれる点と、類似度の算出が近似であることが原因である。

また、一般に Mcut の方が精度が高いので、縮約の程度は小さくした方が精度が高いが、縮約の程度が小さいと計算の負荷を減らす効果も小さく、本来の意味が失われる。どの程度の縮約の度合いにすれば良いか、また現実的に有効な類似度の近似式の提案は、今後の課題である。

6.2 クラスタリングの改善

縮約類似度行列を用いてスペクトラルクラスタリングが可能になるが、縮約類似度行列を作成するために、荒いクラスタリングが必要となる。そのため縮約類似度行列を用いたスペクトラルクラスタリングは縮約類似度行列を作成するために行ったクラスタリング結果を改善しないと意味がない。

クラスタリング結果の改善手法という位置づけで本手法を見た場合、まず最初のクラスタリングで正解と思われるものを固定して、分類が曖昧になるデータに対してだけ、高精度のクラスタリング手法を行っている形になる。

そのためここでのアプローチは CBC (Clustering by Committee) の一種とも考えられる。CBC では Committee と呼ばれる各クラスタの核となるデータセットを作り、各データは Committee との距離によってどの Committee に属するかを判定することでクラスタリングを行う。各データの Committee への割り当ては、単連結法（最短距離法）によるクラスタリングと見なせる。一方、本手法においては k-means で得られたクラスタ中の信頼性のある集合を Committee とし、各データの振り分け部分にスペクトラルクラスタリングを利用している。

またあるクラスタリング結果を部分的に調整することで、クラスタリング結果を改善する手法がいくつか提案されている。論文 [1] の first validation や論文 [3] の Link based refinement は、基本的には、クラスタ内のデータを別のクラスタに移動させた場合の評価値を現在得られているクラスタリング結果から算出し、クラスタ間で一部のデータを移動させることでクラスタリング結果を改善する。

また k-means 等の初期値依存性のあるクラスタリング手法は最初に荒いクラスタリング結果を得ることで、初期値を構成する手法が存在する [4]。その場合、それらの手法はクラスタリング結果の改善手法と位置づけられる。

また混合分布モデルを用いたクラスタリングにおいても、分散共分散行列のモデルを推定するために、最初にクラスタリングを行う場合があり、これもクラスタリング結果の改善手法と見なせる。

6.3 精度が改善できない問題

本来、Committee に属さないデータはクラスタが曖昧なデータであり、それらを正確にクラスタに分類すること自体が困難なタスクである。そのため、本手法を用いても精度が改善できない場合も存在する。

本手法では最初のクラスタリングとして k-means を用いているが、k-means 自体はそれほど荒いクラスタリングではなく、標準的な精度が得られる。スペクトラルクラスタリングが精度の高い手法であったとしても、k-means と精度的に差が生じないようなデータは存在する。このような場合は、k-means の結果を改善することは難しい。スペクトラルクラスタリングが k-means よりも良い結果を出すようなタイプのデータセットに対して本手法は効果を発揮するといえる。

6.4 大規模データセットに対するスペクトラルクラスタリング

ここではデータ数が概ね 10,000 以下のデータセットに対するクラスタリングを考えたが、更に大きなデータセットに対しても利用可能である。

大規模データセットに対する従来のクラスタリング手法は、サンプリングのアプローチと小規模クラスタ生成のアプローチに大別できる。

サンプリングのアプローチでの代表的研究は Ng らの CLARANS である [6]。これはその前身の CLARA の改良版である。CLARA ではランダムサンプルから得たデータを PAM と呼ばれる k-means に類似のクラスタリング手法でクラスタリングする。次にランダムサンプルで選ばれなかったデータを各クラスタとの距離に基づいて割り振ることでクラスタリングを行う。CLARANS は CLARA が最初にランダムサンプルしている部分を、PAM の実行中に行うことで精度の向上を図っている。

小規模クラスタ生成のアプローチの代表的研究は Zhang らの BIRCH [9] と Hinneburg らの格子を用いた手法 [5] がある。BIRCH では最初にデータを走査し、CF-tree というデータの要約情報を作成する。以後の処理は CF-tree に対して行うことで、クラスタリングが可能となる。CF-tree が小規模クラスタ群に対応する。また格子を用いた手法とは各次元を L 個の区間に離散化し、空間を L^d 個の格子に分割する。そして各格子内のデータから小規模クラスタを生成する。L^d がデータ数よりも小くなるように設定すれば、大規模データに対するクラスタリングが可能となる。

どちらのアプローチにしても、中規模のクラスタリングが核となるので、そこで本手法が利用可能になる。

7 おわりに

本論文では縮約類似度行列を用いたスペクトラルクラスタリングを提案した。

スペクトラルクラスタリングは高精度のクラスタリング手法であるが、その実行には、類似度行列のサイズの固有値問題を解く必要があり、計算負荷が高い。ここでは最初に k-means で荒いクラスタリングを行い、各クラスタに対する Committee を作成し、それを 1 点で表すことで縮約類似度行列を作成する。類似度の設定をスペクトラルクラスタリングに特化した形にしている点が特徴である。

これによって大規模データセットに対してもスペクトラルクラスタリングが可能となる。ただし最初に荒いクラスタリング結果を得ているので、本手法はクラスタリング結果の改善手法と位置づけられる。

実験によって提案する縮約類似度行列によりスペクトラルクラスタリングが可能となることを示した。また合計 7 個のデータセットを用いた実験では、最初に行った k-means のクラスタリング結果を改善することができた。

最適な縮約の度合いを推定することと類似度の近似式を精緻にすることを今後の課題とする。

謝辞

本研究の一部は、日本学術振興会 科学研究費補助金 特定研究「日本語コーパス」（課題番号 19011001）による補助のもとで行われた。

参考文献

[1] Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. Iterative Clustering of High Dimentional Text Data Augmented by Local Search. In *The 2002 IEEE International Conference on Data Mining*, pp. 131–138, 2002.

[2] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts. In *The University of Texas at Austin, Department of Computer Sciences. Technical Report TR-04-25*, 2005.

[3] Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. Spectral Min-max Cut for Graph Partitioning and Data Clustering. In *Lawrence Berkeley National Lab. Tech. report 47848*, 2001.

[4] J. He, M. Lan, C-L. Tan, S-Y. Sung, and H-B. Low. Initialization of Cluster Refinement Algorithms: A Review and Comparative Study. In *IEEE Int. Joint Conf. Neural Networks*, pp. 297–302, 2004.

[5] A. Hinneburg and D. A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Knowledge Discovery and Data Mining*, pp. 58–65, 1998.

[6] R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. In *20th International Conference on Very Large Data Bases*, pp. 144–155, 1994.

[7] P. Pantel and D. Lin. Document clustering with committees. In *Proceedings of SIGIR-02*, 2002.

[8] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888–905, 2000.

[9] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, Vol. 1, No. 2, pp. 141–182, 1994.