解　説

データマイニング分野のクラスタリング手法(1)

—クラスタリングを使ってみよう！—

A Survey of Recent Clustering Methods for Data Mining (Part 1)

— Try Clustering! —

神嶌　敏弘　　産業技術総合研究所
Toshihiro Kamishima　National Institue of Advanced Industrial Science and Technology (AIST).
mail@kamishima.net, http://www.kamishima.net/

Keywords: clustering, unsupvervised learning, survey, data mining.

1. はじめに

本稿では，代表的なデータ解析手法であるクラスタリングの最新手法を，二回にわたって紹介する．クラスタリングとは，内的結合（internal cohesion）と外的分離（external isolation）が達成されるようなクラスタと呼ぶ部分集合に，データの集合を分割すること [Everitt 93, 大橋 85] である．クラスタリングはデータマイニングの重要なツールとして利用され [Fayyad 96]，大規模データ処理などの新たな要求が生じている．近年，これらの要求に対処するさまざまな手法が研究されるようになっている．

今回の前編（1）には7章まで，以降を後編（2）(次号掲載予定）に記す．内容的には2～5章の第1部と残りの第2部で構成される．第1部では，クラスタリングを利用するうえで参考になる事柄をまとめた．2章ではクラスタリング全般の参考文献を紹介する．3章には基本的なクラスタリング手法についての簡単な説明を，4章にはこれらの手法を利用するときの注意点をまとめる．5章ではクラスタリングに関するWebサイトを紹介する．

第2部では，データマイニング分野のクラスタリングの最近の研究を紹介する．6章には文書，WWW，バスケットデータなどの解析のため需要が増したカテゴリ属性を扱う手法，7章には確率モデルに基づく手法，8章と9章にはデータマイニングのために需要が増した，対象数や次元数が多い場合の手法，10章には超楕円体状以外の形状のクラスタを抽出できる手法，11章にはクラスタリングでは対象の類似性の定義が分割結果に大きく影響するが，それを制約や事例を用いて定める手法を紹介する．最後の12章ではまとめを述べる．

本稿で用いる表記を表1にまとめたので，参照されたい．

表1　本稿の記号表記一覧

x_i	:対象（object，分類されるもの）
N	:対象数
$X=\{x_1,..,x_N\}$	:対象集合
d	:対象を表す属性ベクトルの次元数
$x_i=(x_{i1},...,x_{id})$	: x_i の属性値ベクトル
$\mathcal{D}_1,...,\mathcal{D}_d$	:属性の値域
k	:クラスタ数
$C_1,...,C_k$	:分割されたクラスタ
n_i	: C_i 中の対象数
c_i	: C_i のセントロイド（平均）
$D(x_i, x_j)$	: x_i と x_j の間の距離（非類似度）

第1部　基本的なクラスタリング手法の利用方法

2. クラスタリング全般についての参考資料

クラスタリング全般についての参考文献をいくつか紹介する．Everittの [Everitt 93] は基本的な事項をまとめた教科書で，クラスタリングの利用者向けによい．JainとDubesの [Jain 88] は，網羅的かつ詳細な文献で，クラスタリングの研究者には非常に参考になる．Jainによるサーベイ論文 [Jain 99] もよくまとめられている．[Jain 00] はクラスタリングをパターン認識の一分野として簡潔に取り上げている．国内の文献では，宮本の [宮本 99] がある．分量的には [Everitt 93] と [Jain 88] の教科書の中間に位置する．

クラスタリングは，統計，パターン認識，データベース，データマイニング，ファジィ，そして人工知能などの分野で研究されているが，本稿では，データマイニングを中心にデータベースや人工知能の分野の研究を取り上げる．統計やパターン認識での研究については前述の [Jain 88] を，ファジィについては [宮本 99] を参考にされたい．また，人工知能分野では，各クラスタの概念記述を分割と同時に獲得する概念クラスタリングがある

が，これについては [Fisher 91] に詳しく，国内では，CLUSTER [Michalski 83] と COBWEB [Fisher 87] について解説した [樋木 96] がある．

本稿で紹介する最近の研究の参考資料をあげておく．国際会議 KDD [Keim 99]や ACM SIGMOD [Hinneburg 99] での Hinneburg と Keim によるチュートリアルの資料は，基本的な事項から最近の研究までを幅広く紹介している．発表資料が http://hawaii.informatik.uni-halle.de/~hinnebur/ClusterTutorial/で公開されているので，本稿とともに参考にされることをすすめる．国内では，福田らの [福田 01] に BIRCH などいくつかの最新手法の解説がある．

3. 基本的なクラスタリング手法

クラスタリング手法は大きく，最短距離法などの階層的手法（hierarchical）と，k-means などの分割最適化手法（partitioning-optimization）に分けられるが，これらの基本的手法を紹介する．

階層的手法は，さらに分枝型（divisive）と凝集型（agglomerative）に分けられるが，ここでは後者のみを扱う．この手法は，1個の対象だけを含む N 個のクラスタがある初期状態から，クラスタ間の距離（非類似度）関数に基づき，最も距離の近い二つのクラスタを逐次的に併合する．そして，この併合を，すべての対象が一つのクラスタに併合されるまで繰り返すことで階層構造を獲得する．この階層構造は図3（b）のようなデンドログラムによって表示する．デンドログラムは，各終端ノードが各対象を表し，併合されてできたクラスタを非終端ノードで表した2分木である．非終端ノードの横軸は，併合されたときのクラスタ間の距離を表す．

クラスタ C_1 と C_2 の距離関数 $D(C_1, C_2)$ の違いにより以下のような手法がある．

最短距離法（nearest neighbor method）
または
単連結法（single linkage method）

$$D(C_1, C_2) = \min_{x_1 \in C_1,\, x_2 \in C_2} D(x_1, x_2)$$

最長距離法（furthest neighbor method）
または
完全連結法（complete linkage method）

$$D(C_1, C_2) = \max_{x_1 \in C_1,\, x_2 \in C_2} D(x_1, x_2)$$

群平均法（group average method）

$$D(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} D(x_1, x_2)$$

ウォード法（Ward's method）

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

ただし，$E(C_1) = \sum_{x \in C_1} (D(x, c_i))^2$

Ward 法は，各対象から，その対象を含むクラスタのセントロイドまでの距離の二乗の総和を最小化する．

なお，最短距離法，最長距離法および群平均法は任意の対象間の距離 $D(x_i, x_j)$ が与えられている場合に適用できる．もし，対象が属性ベクトルで記述されている場合は，属性ベクトル間のユークリッド距離などを求めて適用する．Ward 法は対象が属性ベクトルで与えられている場合にのみ適用できる．

分割最適化手法は，非階層的手法のほか，partitional や optimization など多くの呼び方がある．この手法は，分割の良さの評価関数を定め，その評価関数を最適にする分割を探索する．可能な分割の総数は N に対して指数的なので，実際は準最適解を求める．代表的な k-means 法は，セントロイドをクラスタの代表点とし

$$\sum_{i=1}^{k} \sum_{x \in C_i} (D(x, c_i))^2$$

の評価関数を最大化する．最適解の探索は図1のように，対象のクラスタへの割当てと代表点の再計算を交互に繰り返して行う．この手法は山登り法で，局所最適解しか求められないため，ランダムに初期値を変更して，評価関数を最大にする結果を選択する．

1. k 個の代表点 $c_1,\ldots,c_k$ をランダムに選択
2. $\forall x \in X$ を $\min_i D(x, c_i)$ なる代表点に割当て
3. **if** 代表点への割当てが変化しない **then** 終了
   else 各クラスタのセントロイドを代表点にして
   ステップ2.へ

図1 k-means 法

4. クラスタリングの注意点

前章の基本的手法を実装したソフトウェアは，Web などから容易に入手して利用できる．しかし，各手法の特徴や傾向を無視したために，不適切な結果が導かれている利用例がときどき見られる．ここでは，このような問題を回避するための，主な注意点をまとめる．

4·1 クラスタリング結果の解釈

最も重要な点は，クラスタリングは探索的（exploratory）なデータ解析手法であって，分割は必ず何らかの主観や視点に基づいているということである．よって，クラスタリングした結果は，データの要約などの知見を得るために用い，客観的な証拠として用いてはならない．この「データの要約」を直観的に理解するのに役立つように，Cutting らの研究 [Cutting 92] を紹介する．

データベースから明確な目的に適合する文書を検索する場合，キーワードを用いた文書検索手法は有効である．しかし，明確な目的がなく，データベース全体の傾向を

知りたい場合はどうであろうか？　この場合，具体的なキーワードを示すことは困難なので，文書検索手法の利用は不適当である．そこで，クラスタリングによって，その要約，すなわち，データベース中の主な話題を表すカテゴリの一覧を取り出す．Cuttingらは，ニューヨーク・タイムス紙1990年8月の約5 000件の記事のデータベースの傾向を抽出する問題にクラスタリングを用いた．話題が類似している文書をまとめたクラスタを生成した結果，以下の話題を含むクラスタが発見された．

教育，国内，イラク，芸術，スポーツ，石油，
ドイツ統合，裁判

利用者は，内容が全く不明であった新聞記事のデータベースのおおまかな内容を，これらのクラスタから知ることができるであろう．この要約は，一つのクラスタ，例えばイラクをさらに分割して，パキスタンやアフリカといったより詳細な要約を得たり，文書検索のためのキーワードを決める目的にも利用できる．クラスタリングは，このように未知のデータベースの内容に見当をつける目的で利用できるため探索的であるといえる．

ここで注意すべき点は，この例では8個のクラスタに記事を分類し，データベースの「正しい」要約を得ることができた．しかし，イラクと石油はどちらも湾岸戦争に関する話題なので，これらをまとめても，データベースの「正しい」要約といえる．すなわち，どちらにも，それを正当化する視点が存在する．このように，クラスタリングの結果は絶対的でも，普遍的でも，客観的でもないので，分割結果は結論を導く証拠にはなり得ない．例えば，教育と国内が違うクラスタに分類されているが，これは実社会で二つの問題に関連性が皆無であることを意味しない．クラスタリング結果の妥当性は，その分割の利用目的など，外的な知識によって判断するしかない．例えば，新聞記事の話題の抽出という目的であれば，国内とドイツ統合を同じクラスタに分類していれば，妥当な要約とはいえないであろう．だが，同じ週に起きた事件をまとめるという目的ならば，これらをまとめるのも妥当かもしれない．クラスタリングの結果は，その利用目的などに応じて，妥当性を常に検証する必要がある．

ただし，均一に分布するデータを分割する行為は，多くの場合，妥当ではない．この観点での妥当性に関する議論は [Dubes 79] に詳しい．また，妥当性の問題に関連するものとして，クラスタ数の決定問題があるが，よく分離されたクラスタ構造が存在する場合に，クラスタ数の決定基準を比較した研究 [Milligan 85] などもある．

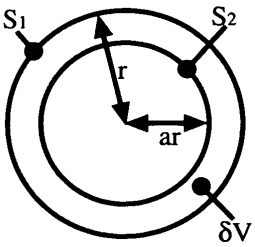


図2　球面集中現象 [石井 98]

4·2　次元の呪い

高次元空間の対象を扱う場合，その高次元性に起因した問題は「次元の呪い」と呼ばれる．クラス分類問題では，高次元の対象を分類する場合に，十分な精度の推定に必要な事例数が不足することが問題となる [坂野 02]．クラスタリングでも次元の呪いの問題は存在し，その原因は次の球面集中現象 [石井 98] にある．

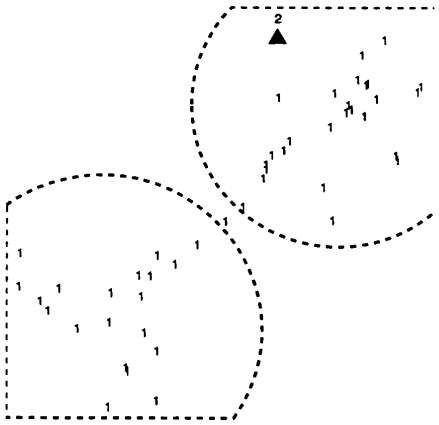
図2のように，ある対象を中心にした，半径がそれぞれ r と $ar(0<a<1)$ の d 次元超球 S_1 と S_2 があるとする．S_1 の体積 V に対する，二つの超球の体積の差 δV（図のグレー部分）の比は $\delta V/V=1-a^d$ となる．ここで，対象が均一分布していると，すると空間中に存在する対象数は体積に比例する．また，$\delta V/V$ は d の増加に伴い1に近づくことから，d が大きな場合は S_1 中の対象は，ほとんど二つの超球の隙間に存在することがわかる．これは，中心の対象からほかの対象までの距離は，次元の増加に従って急速に大きくなることを意味する．すなわち，どの対象も互いに似ていないことになる．クラスタリングは，似ている点をまとめる操作なので，有意な解を得ることができなくなる．

この次元の呪いに対する抜本的な解決法は，外的な知識によって不要と考えられる属性を排除し，次元数を小さくすることである．しかし，データの性質が不明で，どの属性が不要か判断できない場合も多いため，9章のような高次元データを処理する手法も研究されている．

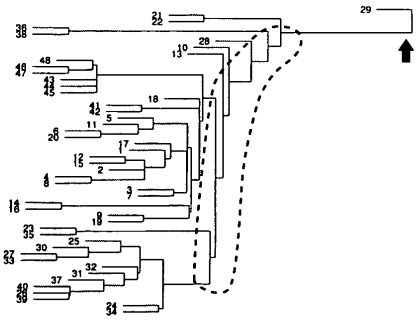
4·3　手法固有の注意点

階層的手法の注意点について述べる．数学的に優れたある種の性質をもつ最短距離法は，空間濃縮という性質のため，極めて外乱に弱く，実データではあまり良い結果は得られない．空間濃縮とは，併合されてできた新しいクラスタは，以後の併合の対象として選ばれる可能性が加速度的に増加する現象である．図3 (a) のデータには，視覚的なまとまりからすれば点線で囲んだ部分に二つのクラスタとみなせる部分と，これらをつなぐ外乱とみなせる対象が存在する．このデータに最短距離法を適用すると，図3 (b) のデンドログラムが得られる．このデンドログラムを用いて，図3 (b) の矢印の部分で二つのクラスタに分割すると，一方は1個の対象（図3 (a) の▲印），もう一方はそれ以外すべてを含むクラスタが生成され，直観に反した結果が得られる．また，この空間濃縮は，つながった外乱を一つずつ併合することがあるため，チェイニング効果とも呼ばれる．この効果によって図3 (b) の点線部分のような階段状構造ができるが，恣意的な構造であることが多いので注意されたい．

逆に，最長距離法には，併合されてできたクラスタは，



(a) データ



(b) デンドログラム

図3 チェイニング効果の例 [Everitt 93]

以後，併合されにくくなる空間拡散という性質があり，本来のクラスタから偶然離れてしまった対象があると，過剰に分割される傾向がある．群平均法は，空間濃縮や拡散は生じないが，デンドログラムの反転現象 [宮本 99] があり，階層構造が必要な場合には適さない．

k-means 法の注意点について述べる．この手法は，クラスタは超球状の形状で，クラスタ中の対象数はどれも等しいということを暗黙のうちに仮定しているので，この仮定に反する構造の抽出は困難である．文献 [Guha 98] では，図4のように三つの密な領域が存在するデータに k-means 法を適用した結果を例示している．直観に反して，クラスタ中の事例数が等しくなるように，対象数の大きな領域は三つに分割されている．クラスタの対象数に差がある場合には，k-means 法の代わりに7章の確率モデルに基づく手法を検討されたい．また，クラスタが超球形でない場合は，標準偏差で正規化したマハラノビス距離などのスケーリングの変換の採用や，10章を参考にされたい．

最初に適用するクラスタリング手法は一般には以下のようになろう．まず，対象が属性ベクトルで記述されている場合，計算量が k-means は $O(Nk)$ に対し，階層的手法は $O(N^2)$ なので，k-means を用いる．ただし，階層構造が必要な場合には Ward 法を用いる．属性ベクトルではなく，対象間の距離だけが与えられている場合は，群平均法を適用する．ただし，階層構造が必要な場合は最短距離法や最長距離法を用いる．もちろん，あらゆる状況で最良な手法は存在しないので，必ずこの選択が良いというわけではない．だが，これらの手法を適用した結果からデータに関する知見を得て，その知見に基づき，ほかの高度な手法を検討するのがよいだろう．

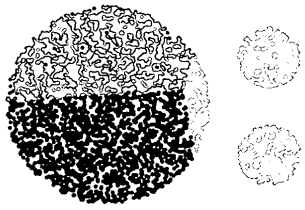


図4 k-means 法に不適な事例 [Guha 98]

```
library(mva)                  # 多変量解析ライブラリ
x <- read.table("datafile") # ファイルの読込み
cl <- kmeans(x, 2, 20)        # クラスタリング
plot(x, col=cl$cluster)       # 散布図の描画
```

図5 R による k-means 法の実行と表示のプログラム例

5. WWW 上の情報

第1部の最後に統計・データマイニング関係のポータルサイトをいくつか紹介する．第2部で紹介する最新手法はないが，基本的な手法を実装したソフトウェアはこれらのサイトで検索できる．

The R Project: http://www.r-project.org/
S 言語のフリーな処理系 R のサイト．多くの OS で動作し，kmeans や hclust というクラスタリング用の関数も実装され，その表示も容易である（図5）．多くの統計手法が実装されており，それらの情報もここで入手できる．

Netlib: http://www.netlib.org/
数値計算手法に関する著名なポータルサイト．

Scientific Applications on LINUX:
http://sal.kachinatech.com/
LINUX 上で動作する科学技術計算ソフトウェアに関するポータルサイト．

StatPages.net: http://www.statpages.net/
統計関係情報のポータルサイト．

Recursive-Partitioning.com:
http://www.recursive-partitioning.com/
研究発表された手法で，ソフトウェアを公開しているものを集めたリンク集．

KDnuggets: http://www.kdnuggets.com/
データマイニングで著名なポータルサイト．

第2部 データマイニングのための クラスタリング手法

6. カテゴリ属性を扱うクラスタリング

数値属性で表される対象を分類するクラスタリング手法が一般的だが，バスケットデータ（各取引ごとに顧客が購入した商品のリスト），文書，Webページなどを分類する目的で，カテゴリ属性を扱う手法の研究が行われている．

§1 基本的手法の拡張

まず，3章の手法を修正して適用することが考えられる．最短距離法などの階層的手法は対象間の距離が与えられれば利用できる．そこで，ハミング距離やJaccard係数 [Jain 88] などのカテゴリ属性用の距離を導入すれば，階層的手法を適用できる．一方，k-means法を修正する手法もある．Huangらのk-mode [Huang 98] は，クラスタの代表として，セントロイドの代わりにモード，すなわち各属性についてクラスタ中で最も頻度の高い属性値を選んだ属性ベクトルを用いる．また，代表と対象の距離には単純一致係数（simple matching），すなわち二つの属性ベクトルで一致しない属性の数を用いる．これらの手法が有効な分野も多くあるが，バスケットデータやWebページの分類では適さない場合もあるため，以下のような研究がある．

§2 ROCK

GuhaらのROCK（RObust Clustering using linKs）[Guha 99] はリンクという概念を用いた手法である．リンクとは，二つの対象に共通に近隣である対象の数である．ただし，二つの対象が近隣であるとは，類似度（Jaccard係数などで測る）がしきい値以上であることである．二つの対象だけでなく，それらの近隣の影響を考慮することで，少数の例外的な対象の影響を受けにくいことが利点である．リンクの大きな対象を同じクラスタに，小さなものを別のクラスタに分類する目的で次の評価関数を最大化する．

$$\sum_{i=1}^{k} n_i \times \sum_{x_q, x_r \in C_i} \frac{\mathrm{link}(x_q, x_r)}{{n_i}^{1+2f(\theta)}}$$

ただし，$\mathrm{link}(x_q, x_r)$ は対象 x_q と x_r の間のリンク数，${n_i}^{1+2f(\theta)}$ は属性値がランダムだった場合のリンク数の期待値を表す．この関数を最大化するように設計されたヒューリスティックによって，階層的手法で分割を探索する．

§3 CACTUS

ROCKは計算量が$O(N^3)$と大きいので，この問題に取り組んだCACTUS（CAtegorical ClusTering Using Summaries）[Ganti 99] をGantiらは提案している．CACTUSは，任意の類似性に基づく近隣関係ではなく，対象集合中の属性値の共起性に基づいた連結関係を用いる．直観的には，共起性の強い属性値を集めて構成する領域で，対象数が多いものをクラスタとするが，形式的には以下のように定義される．i番目の属性値が$a_i \in \mathcal{D}_i$で，j番目の属性値が$a_j \in \mathcal{D}_j$である対象の数を$\sigma(a_i, a_j)$で表す．属性値がランダムな場合の期待値$\mathrm{E}[\sigma(a_i, a_j)]$に対して，$\sigma(a_i, a_j) > \sigma\mathrm{E}[\sigma(a_i, a_j)]$なら$a_i$と$a_j$は強連結であるという．ただし，$\sigma > 1$はパラメータ．この属性値の関係を属性値の集合$S_i \subset \mathcal{D}_i$と$S_j \subset \mathcal{D}_j$に拡張する．すなわち，$S_i$と$S_j$中の属性値の任意の対が強連結であるとき$S_i$と$S_j$は強連結であるという．このときクラスタ$S = S_1 \times \cdots \times S_d$は次の3条件を満たす．(1) すべての$S_i$と$S_j$が強連結，(2) すべての$S_i$は極大，(3) $\sigma(S)$が$\alpha\mathrm{E}[\sigma(S)]$より大きい．ただし，$\sigma(S)$は，属性値ベクトルが$S$に含まれる対象の数で，$\mathrm{E}[\sigma(S)]$はランダムな場合の期待値．さらに，強連結な属性についての要約情報があれば，データ全体の情報がなくてもクラスタを抽出できる性質を利用して，記憶容量を削減する．この要約情報には，違う属性の属性値間の強連結性を保持する属性間要約と，同じ属性の属性値について，ほかの属性との強連結情報の類似性を保持する属性内要約がある．さらに，2次元空間$\mathcal{D}_i \times \mathcal{D}_j$上にあるクラスタを最初に見つけ，次元を一つずつ増やしながら探索することで，効率の良い計算を可能にしている．

§4 STIRR

GibsonらによるSTIRR（Sieving Through Iterated Relational Reinforcement）[Gibson 98] は属性値の共起関係をもとに分類する手法である．通常のクラスタリングと異なり，対象を分類するのではなく，各属性の属性値を二つのクラスタに分割する．図6は，STIRRにおけるデータの表現形式を示している．四角が属性を表し（x_{i1}, x_{i2}, x_{i3}の3属性），四角内の円をノードと呼び，それぞれ各属性がとり得る属性値を示す．対象はこれらのノードを結ぶ線で表す．例えば，矢印で示した破線はすべての属性で2番目の属性値をとる対象を表す．各ノードには，重みが割り当てられており，その重みの符号の正負によって，属性値は分類される．この重みは，適当な初期値から開始して，収束するまで重みを更新する手法で求める．図中のノードvの重みを更新する場合について述べる．ノードvの属性値を含む任意の対象x_τについて，v以外のノードに割り当てられた（図ではx_{i2}とx_{i3}中のノード）重みを「結合」した値を$w(x_\tau)$とする．結合の方法としては，重みの総和など4種類の手法を提案している．この$w(x_\tau)$を，属性値vを含むすべての対象（図中では点線で表されたもの）について求め，そ

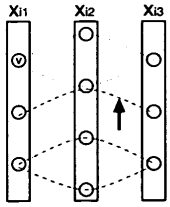


図6 STIRRにおけるデータの表現形式

の総和を新しい v の重み $w(v)$ とする．この操作をすべてのノードについて反復適用する．

§5 グラフスペクトル理論を用いた手法

STIRR はグラフスペクトル理論と関連があり，これを用いた手法がいくつか研究されている．Ding らの Mcut [Ding 01] では，対象をノード，対象間の類似度を辺の重みとするグラフと考える．類似度には，例えば，文書分類の場合には文書ベクトルのコサインなどを用いる．グラフを C_1 と C_2 の二つに分割したとき，切断される辺の重みの総和を $\mathrm{cut}(C_1, C_2)$，グラフ C_1 内のノード間の重みの総和を $W(C_1)$ で表して

$$\mathrm{Mcut} = \frac{\mathrm{cut}(C_1, C_2)}{W(C_1)} + \frac{\mathrm{cut}(C_1, C_2)}{W(C_2)}$$

を最小にする分割を求める．直観的に，分子を小さくすることは外的分離を，分母を大きくすることは内的結合を強めることになる．この Mcut を最小にする分割を求める問題は NP 完全である．しかし，C_1 に分類されることを 0，C_2 なら 1 で表す離散変数を，0 に近ければ C_1 に分類されるやすいという連続変数で近似すれば，グラフスペクトル理論を用いて最適化が容易になる．そして，近似した変数を，再び離散化すれば分割が得られる．もちろん，この分割は Mcut を最小にするわけではないが，よい近似となる．

ほかに，グラフスペクトル理論を用いたクラスタリングの研究としては，画像処理で線分の抽出を行う [Tsuda 96]，リンク情報に基づき Web ページを分類する [He 01]，情報検索に用いた [井上 00]，文書と語彙の二部グラフを用いて文書分類をする [Dhillon 01] などがある．

7. 確率モデルに基づく手法

§1 最尤推定を用いた手法

確率モデルに基づいたクラスタリング手法は，データマイニング，統計，パターン認識の分野で非常によく研究されている．これは，i 番目のクラスタについて，パラメータ θ_i をもつ対象の確率（密度）分布 $f_i(x \mid \theta_i)$ と，クラスタの混合比 $\forall \alpha_i > 0$, $\sum_i^k \alpha_i = 1$ で表される次の混合分布を用いる手法である．

$$\Pr[x \mid \theta_1, \ldots, \theta_k] = \sum_i^k \alpha_i f_i(x \mid \theta_i)$$

$f_i(x \mid \theta_i)$ には，対象の属性が実数値の場合は正規分布，カテゴリ値の場合は多項分布が用いられることが多い．そして，与えられた対象集合 X が，この混合分布に基づき発生する場合の尤度を最大にする最尤推定によりパラメータを求めて，混合比と確率分布の積が最大になるクラスタへ各対象を分類する．最尤推定には，一般に EM アルゴリズム [Dempster 77] が用いられる．これは簡単にいうと，対象のクラスタへの確率的割当てと，各クラスタ分布のパラメータ推定を交互に行う方法である．この手法で，すべてのクラスタと属性で標準偏差が等しく，共分散が単位行列である正規分布を用いた場合と k-means 法では類似した結果が得られる．4章で述べたように，k-means 法ではクラスタ数が等しい超球状のクラスタしか抽出できない．これらの制約を緩和したい場合には，標準偏差や共分散行列への制約を緩めた混合分布を用いた手法を利用するとよい．確率モデルに基づく手法の利点は，理論的背景が明確である点である．一方，対象間の類似性尺度が確率分布により暗黙的に決定される（例えば，正規分布ならユークリッド距離が仮定される）ため，任意の類似性尺度の利用が難しいことが欠点である．また，確率分布が指数族でない場合にはかなり計算が複雑になる．Meilă ら [Meilă 01] は EM アルゴリズムを含めた最適化や初期化の方法をいくつか実験的に比較している．時系列データの分類を，確率分布モデルにマルコフモデルを導入して行う [Cadez 00] や [吉岡 02] などもある．

§2 AutoClass

最尤推定では，クラスタ数や確率分布の選択ができない問題がある．そこで，パラメータの事前分布を導入し事後確率最大（Maximum A-Posteriori: MAP）推定をする手法として Cheeseman らの AutoClass [Cheeseman 96, Hanson 91] がある．以下の Web サイトでソフトウェアも公開されている（http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/）．

Paliouras ら [Paliouras 00] は WWW サーバのセッションログからコミュニティを抽出する問題に，この AutoClass や Kohonen の自己組織化マップ [Kohonen 97] などを適用し比較している．

[後編（2）に続く]

◇ 参 考 文 献 ◇

[Cadez 00] Cadez, I. V., Gaffney, S. and Smyth, P.: A General Probabilistic Framework for Clustering Individuals and Objects, *Proc. 6th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 140-149 (2000)

[Cheeseman 96] Cheeseman, P. and Stutz, J.: Bayesian Classification (AutoClass): Theory and Results, in Fayyad, U. M., Diatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds.: *Advances in Knowledge Discovery and Data Mining*, chapter 6, pp. 153-180, AAAI Press/The MIT Press (1996)

[Cutting 92] Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proc. 15th Annual ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 318-329 (1992)

[Dempster 77] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum Likelihood from Incomplete Data via The EM Algorithm, *Journal of the Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1-38 (1977)

[Dhillon 01] Dhillon, I. S.: Co-clustering documents and words using Bipartite Spectral Graph Partitioning, *Proc. 7th Int'l*

Conf. on Knowledge Discovery and Data Mining, pp. 269-274 (2001)
[Ding 01] Ding, C. H. Q., He, X., Zha, H., Gu, M. and Simon, H. D.: A Min-max Cut Algorithm for Graph Partitioning and Data Clustering, *Proc. IEEE Int'l Conf. on Data Mining,* pp. 107-114 (2001)
[Dubes 79] Dubes, R. and Jain, A. K.: Validity Studies in Clustering Methodologies, *Pattern Recognition,* Vol. 11, pp. 235-254 (1979)
[Everitt 93] Everitt, B. S.: *Cluster Analysis,* Edward Arnold, third edition (1993)
[Fayyad 96] Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in Fayyad, U. M., Diatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining,* chapter 1, pp. 1-34, AAAI Press/The MIT Press (1996)
[Fisher 87] Fisher, D. H.: Knowledge Acquisition via Incremental Conceptual Clustering, *Machine Learning,* Vol. 2, pp. 139-172 (1987)
[Fisher 91] Fisher, D. H. and Pazzani, M. J.: Computational Models of Concept Learning, in Fisher, D. H., Pazzani, M. J. and Langley, P. (eds.): *Concept Formation: Knowledge and Experience in Unsupervised Learning,* chapter 1, pp. 3-43, Morgan Kaufmann (1991)
[福田 01] 福田剛志, 森本康彦, 徳山 豪: データマイニング, データサイエンスシリーズ, 第3巻, 共立出版 (2001)
[Ganti 99] Ganti, V., Gehrke, J. and Ramakrishnan, R.: CACTUS-Clustering Categorical Data Using Summaries, *Proc. 5th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 73-83 (1999)
[Gibson 98] Gibson, D., Kleinberg, J. and Raghavan, P.: Clustering Categorical Data: An Approach Based on Dynamical Systems, *Proc. 24th Very Large Database Conf.,* pp. 311-322 (1998)
[Guha 98] Guha, S., Rastogi, R. and Shim, K.: CURE: An Efficient Clustering Algorithm for Large Databases, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp. 73-80 (1998)
[Guha 99] Guha, S., Rastogi, R. and Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes, *Proc. 15th Int'l Conf. on Data Engineering,* pp. 512-521 (1999)
[Hanson 91] Hanson, R., Stutz, J. and Cheeseman, P.: Bayesian Classification with Correlation and Inheritance, *Proc. 12th Int'l Joint Conf. on Artificial Intelligence,* pp. 692-698 (1991)
[He 01] He, X., Ding, C. H. Q., Zha, H. and Simon, H. D.: Automatic Topic Identification Using Webpage Clustering, *Proc. IEEE Int'l Conf. on Data Mining,* pp. 195-202 (2001)
[Hinneburg 99] Hinneburg, A. and Keim, D. A.: Clustering Methods for Large Databases: From the Past to the Future, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* p. 509 (1999)
[Huang 98] Huang, Z.: Extensions to the k-Means Algorithm for Clustering Large Data with Categorical Values, *Journal of Data Mining and Knowledge Discovery,* Vol. 2, pp. 283-304 (1998)
[井上 00] 井上光平, 浦浜喜一: 共起関係行列に基づくファジークラスタリングとデータ検索への応用, 電子情報通信学会論文誌 (D-II), Vol. J83-D-II, No. 3, pp. 957-966 (2000)
[石井 98] 石井健一郎, 上田修功, 前田英作, 村瀬 洋: わかりやすいパターン認識, オーム社 (1998)
[Jain 88] Jain, A. K. and Dubes, R. C.: *Algorithms for Clustering Data,* Prentice Hall (1988)
[Jain 99] Jain, A. K., Murty, M. N. and Flynn, P. J.: Data Clustering: A Review, *ACM Computing Surveys,* Vol. 31, No. 3 (1999)
[Jain 00] Jain, A. K., Duin, R. P. W. and Mao, J.: Statistical Pattern Recognition: A Review, *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 22, No. 1, pp. 4-37 (2000)
[Keim 99] Keim, D. A. and Hinneburg, A.: Tutorial 3. Clustering Techniques for Large Data Sets − From the Past to the Future, in *Tutorial Notes of The 5th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 141-181 (1999)
[Kohonen 97] Kohonen, T.: *Self-Organizing Maps,* second edition, Springer-Verlag (1997)
[Meilă 01] Meilă, M. and Heckerman, D.: An Experimental Comparison of Model-Based Clustering Methods, *Machine Learning,* Vol. 42, pp. 9-29 (2001)
[Michalski 83] Michalski, R. S. and Stepp, R. E.: Learning from Observataion: Conceptual Clustering, in Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (eds.), *Machine Learning I: An Artificial Intelligence Approach,* chapter 11, pp. 331-363, Morgan Kaufmann (1983)
[Milligan 85] Milligan, G. W. and Cooper, M. C.: An Examination of Procedures for Determining The Number of Clusters in A Data Set, *Psychometrika,* Vol. 50, No. 2, pp. 159-179 (1985)
[宮本 99] 宮本定明: クラスター分析入門ーファジィクラスタリングの理論と応用, 森北出版 (1999)
[大橋 85] 大橋靖雄: 分類手法概論, 計測と制御, Vol. 24, No. 11, pp. 999-1006 (1985)
[Paliouras 00] Paliouras, G., Papatheodorou, C. and Karkaletsis, V.: Clustering the Users of Large Web Sites into Communities, *Proc. 17th Int'l Conf. on Machine Learning,* pp. 719-726 (2000)
[坂野 02] 坂野 鋭, 山田敬嗣: 怪奇！！ 次元の呪いー識別問題, パターン認識, データマイニングの初心者のために (前編), 情報処理, Vol. 43, No. 5, pp. 562-567 (2002)
[樋木 96] 樋木哲夫: 概念クラスタリング, 日本ファジィ学会誌, Vol. 8, No. 3, pp. 463-467 (1996)
[Tsuda 96] Tsuda, K., Minoh, M. and Ikeda, K.: Extracting straight lines by sequential fuzzy clustering, *Pattern Recognition Letters,* Vol. 17, pp. 643-549 (1996)
[吉岡 02] 吉岡 琢, 石井 信: 制約付き混合主成分分析による時系列データのクラスタリング, 第5回情報論的学習理論ワークショップ, pp. 196-201 (2002)

2002年11月8日 受理

著者紹介

神嶌 敏弘 (正会員)

1968年生まれ. 1992年京都大学情報工学科卒業. 1994年同大学院修士課程修了. 同年, 電子技術総合研究所入所. 2001年博士 (情報学). 同年, 電子技術総合研究所は産業技術総合研究所へ再編. 機械学習とその応用の研究に従事. ACM, 情報処理学会会員.