解 説

データマイニング分野のクラスタリング手法 (2)

—大規模データへの挑戦と次元の呪いの克服—

A Survey of Recent Clustering Methods for Data Mining (Part 2)

— Challenges to Conquer Giga Data Sets and the Curse of Dimensionality —

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.

前編(1)（Vol. 18, No. 1に掲載）に引き続きデータマイニング分野のクラスタリング手法を紹介する．本稿で用いる記号を表1に記した．

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
$\mathbb{A}_1,...,\mathbb{A}_d$	: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）

8. 大規模データへの挑戦

データマイニングでは非常に大規模のデータを処理する必要があるが，既存の手法には二つの問題点がある．一つは，データのすべてを主記憶上には保持できないこと．もう一つは，計算量に関する問題で，階層的手法では $O(N^2)$，k-means では $O(Nk)$ であるが，$O(N)$ であることが望ましい．本章ではこれらの問題に対処した手法を紹介する．

§1 CLARANS

近年の大規模化についての研究はNgらのCLARANS (Clustering Large Applications based on RANdomized Search) [Ng 94] に始まる．CLARANSは，PAM (Partition Around Medoids) とCLARA (Clustering LARge Applications) の拡張手法なので，まずこれらについて述べる．PAMは k-means と類似しているが，クラスタをセントロイドではなくmedoid（クラスタ中の対象の一つ）で代表する点が異なる．medoid以外の対象は，最も近いmedoidが代表するクラスタに分類され，medoidまでの距離の二乗和（ポテンシャル）を最小にする分割を求める．アルゴリズムは k-means と同様に，k 個の初期medoidをランダムに選択する．その後は，対象の割当てと，ポテンシャルを最小にするようなmedoidの選択を交互に行って，medoidが更新されなくなれば停止する．PAMの計算量は $O(N^2k)$ であり，非実用的なので，ランダムサンプルを用いたCLARAが提案された．N 個の対象から $40+2k$ 個をランダムサンプリングしPAMを適用してmedoidを求める．最後に，ランダムサンプリングされなかった対象を，最も近いmedoidに割り当てることで分割を生成する．この手続きを何度か実行してポテンシャルが最小になる分割を選択する．CLARAの計算量は $O(k^3+Nk)$ と小さくなるが，精度は低下する．CLARANSはこの点を改良した手法である．CLARAのように最初にサンプリングするのではなく，medoidの更新時にランダムサンプリングをして，そのサンプルだけを新たなmedoidの候補として評価する点が異なる．論文では，経験的にサンプルサイズを $k(N-k)/80$ と250の大きいほうとしている．実験により，同じ時間でCLARANSはCLARAよりもポテンシャルの小さな分割を発見できることを示している．

§2 BIRCH

ZhangらのBIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [Zhang 96, Zhang 97] は，クラスタリングにおいて，記憶容量を明示的に考慮し，計算量を $O(N)$ に抑えた点で注目すべき研究である．記憶容量と計算量を削減できる鍵は，最初にデータ全体を走査してClustering Feature Tree (CF-tree) というデータの要約情報を生成し，以後の操作をこのCF-treeだけに限定する点にある．このCF-tree は N に対して十分小さいので，最初にデータを走査するための計算量 $O(N)$ が全体の計算量になり，さらに主記憶上

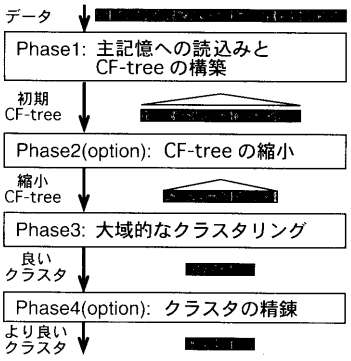


図1　BIRCH アルゴリズム [Zhang 97]

に保持可能である．

CF-tree は，Clustering Feature（CF）という部分クラスタの要約情報を付随させたバランス木である．この木は，分岐係数（B は非終端，L は終端）としきい値 T をパラメータとする．各非終端ノードは，最大 B 個の子ノードへのリンクを，子ノードが表す部分クラスタの CF とともに保持し，終端ノードは最大 L 個の CF と，終端ノード間の双方向リンクを保持する．パラメータ B と L は，主記憶のページの大きさを考慮して定める．T は，終端ノードの CF が表す部分クラスタ C の直径

$$\sqrt{\frac{\sum_{x_i, x_j \in C,\, x_i \neq x_j} (x_i - x_j)\cdot(x_i - x_j)}{n(n-1)}}$$

の最大値を定める．CF-tree は，直径 T 未満の対象集合（部分クラスタ）の要約情報だけを終端ノードに保持するため，N に対して非常に小さくなる．要約情報 CF は第2要素がベクトルで，ほかはスカラーの三組である．

$$\mathrm{CF} = \left(n, \sum_{x_i \in C} x_i, \sum_{x_i \in C} x_i \cdot x_i \right)$$

この CF には，二つの部分クラスタ C_i と C_j の CF から，部分クラスタ $C_i \cup C_j$ の CF が計算できる性質がある．さらに，Zhang らは CF のみから計算可能な部分クラスタ間の距離 $D0 \sim D4$ も示している．次式は，その一つである．

$$D2(C_i, C_j) = \sqrt{\frac{\sum_{x_s \in C_i} \sum_{x_t \in C_j} x_s \cdot x_t}{n_i n_j}}$$

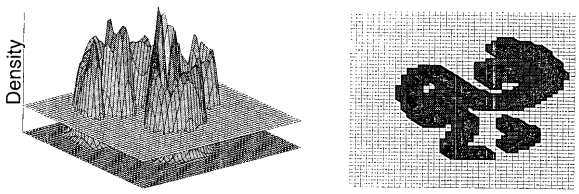
CF のこれらの性質により，BIRCH はデータ全体を参照することなく，CF-tree を更新できる．クラスタリングでは，終端ノードの部分クラスタを，あたかも，一つの対象のようにまとめて扱う．ただし，同じ部分クラスタ中の対象が違うクラスタには分類されないが，同じノードの部分クラスタは違うクラスタに分類される可能性がある．

次に，図1のアルゴリズムについて述べる．Phase1 ～ 4 のうち，1 と 3 は必須で，2 と 4 はオプションである．Phase1 は BIRCH の中心的な段階で，データを読み込んで CF-tree を生成する．生成手順は B+-tree というバランス木の生成手順と同様である．すなわち，各対象は，CF-tree のルートから終端ノードに向かって一番近い（$D0 \sim D4$ の一つで測る）部分クラスタを表すリンクをたどる．もし，終端ノードの中で一番近い CF に新しい対象を加えても，その直径が T 未満であれば，その CF にその対象を追加する．そうでなければ，終端ノードに CF を追加する．ただし，追加後に CF の数が L より大きくなる場合は，その終端ノードを二つに分け，親の非終端ノードの子ノードを増やす．さらに，その非終端ノードの子ノード数が B より大きくなる場合はさらに上位の非終端ノードの子ノード数を増やし，ルートノードの子ノード数が制限を超える場合は木を一段深くする．ほかに，限られた主記憶容量でできるだけ詳細な CF-tree を生成するために，小さな T で開始して，容量が不足したときに T を増やして CF-tree を再構築する手法や，木のバランスが悪くなった場合に再構築する手法なども示している．Phase2 では，前の段階で生成された初期 CF-tree を縮小して，次の Phase のクラスタリングをより容易にする．具体的には，少数の対象しか含まない孤立した CF をはずれ値として除去したり，非常に近い CF をまとめるといった操作をする．Phase3 では，既存のクラスタリング手法を用いてクラスタを生成する．論文では，$D0 \sim D4$ の距離（$D2$ と $D4$ を推奨）を用いて凝集型階層的手法を適用している．最後の Phase4 は，クラスタを精錬する段階である．Phase3 で求めたクラスタのセントロイドを求め，すべての対象を最も近いセントロイドに再分類する．この段階を実行すると，データを2度走査することになるので，データベースの JOIN 演算が必要な場合など，I/O コストが非常に大きい場合には省略できる．

§3　格子を用いた手法

データを要約するために，対象が存在する空間を，各次元ごとに L 個の区間に離散化して，全部で L^d 個の格子に分割し，同じ格子内の対象をまとめて扱う手法がある．L を小さくすれば高速になるが，精度の低いクラスタしか生成できなくなるので，L を適切に設定できれば有効な手法となる．これらの手法は，$L^d \ll N$ であれば，最初にデータを走査する計算量 $O(N)$ が全体の計算量になる．このような手法をいくつか紹介する．

Hinneburg らの DENCLUE（DENsity-based CLUsterEring）[Hinneburg 98] は対象の密度に基づくクラスタを求める．直観的に説明すると，対象の密度関数（図2



（a）密度関数と密度＝ξ 平面　（b）しきい値 ξ 以上の領域

図2　DENCLUE アルゴリズム [Hinneburg 98]

(a)) を求め，密度がξより大きな領域を取り出すと図2 (b) のように，いくつかの連結領域が抽出される．そして，各対象から始めて密度の高い方向に山登り法で進み，同じ連結領域に到達する対象がクラスタを構成する．ここで，密度関数を求める必要があるが，格子に基づく近傍を定義し，近傍の対象だけで密度関数を求めることで計算を高速にする．形式的には，一辺が2σの格子の中で対象を含む格子の集合をG_p，格子中の対象数が$\xi/(2d)$より多い格子の集合をG_{sp}とする．格子中の対象のセントロイド間の距離が4σであるものを連結しているといい，ある格子$g \in G_{sp}$とそれに連結したG_p中の格子全体の集合をG_rとする．ここで$|G_{sp}| \ll |G_p| \ll N$なので計算量は少なくて済む．対象$x \in g$について，格子$g_1 \in G_r$のセントロイドとxとの距離が4σ以下であるようなすべての格子g_1に含まれる対象の集合を近傍near(x) とする．このとき対象xでの密度関数は

$$\sum_{x \in \mathrm{near}(x_1)} \exp\left(-\frac{(D(x, x_1))^2}{2\sigma^2}\right)$$

で求められる．BIRCHと異なり任意形状のクラスタが抽出できるが，その代わりにξやσを調整する必要がある．

Xuらの DBCLASD (Distribution Based Clustering of LArge Spatial Databases) [Xu 98] は，対象が高密度に均一分布している隣接した格子をクラスタCとして抽出する．クラスタ内の各対象$x \in C$について，クラスタ内で最も近い対象までの距離$\min_{x_i \in C,\, x \neq x_i} D(x, x_i)$を考える．クラスタ内で対象が均一分布しているときの，これらの距離の集合の分布は既知なので，このことを用いてクラスタ内の均一性を検定できる．アルゴリズムは，seedとなる格子から始めて，均一性が維持されるような隣接する格子を順次クラスタに加えることで，クラスタを拡張する．

その他，格子を用いる手法には，Waveletを用いたSheikholeslamiらのWaveCluster [Sheikholeslami 98] や，検索条件に適合する領域を抽出するための索引付け手法であるWangらのSTING (STatistical INformation Grid-based method) [Wang 97] などがある．

§4 VFKM

CF-treeや格子のような要約情報ではなく，サンプリングを用いるCLARANSのような手法には，DomingosらのVFKM [Domingos 01] もある．これは，CLARANSではサンプリング数をヒューリスティックにより定めたが，損失の理論的な限界を考慮することで適切なサンプル数を決定する点が異なる．この手法ではk-means法を，事例数を増やしながら何度か適用する．k-meansを，同じ初期状態から適用したとして，無限個の対象があった場合の分割をC^∞，その中のクラスタのセントロイドの位置をc_i^∞とする．同様にN個の対象があった場合の分割をC^N，セントロイドの位置をc_i^Nとする．このとき，

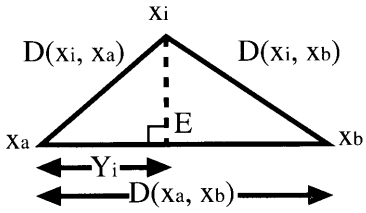


図3 直線$x_a x_b$への射影 [Faloutsos 95]

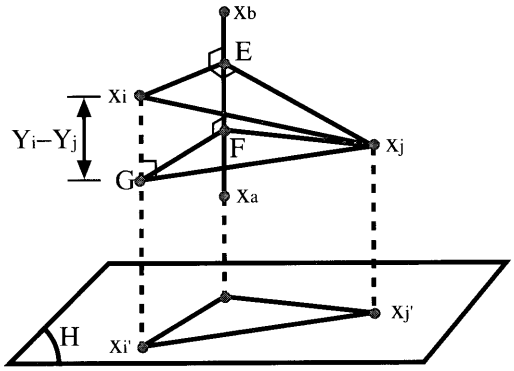


図4 直線$x_a x_b$に垂直な平面Hへの射影 [Faloutsos 95]

損失をセントロイド間の距離の総和

$$L(C^N, C^\infty) = \sum_i^k D(c_i^N, c_i^\infty)$$

で測る．この損失がε以下である確率がδ以下，すなわち，

$$\Pr[L(C^N, C^\infty) \leq \varepsilon] \leq \delta$$

を満たすようにサンプル数を定める．そのため，まず小さな対象集合にk-meansを適用し，その結果に基づき損失の上限を推定する．目標を達成できれば停止するが，そうでなければ，クラスタリングの結果に基づいてサンプリングする対象数Nを増やし，目標を満たすまでk-meansを繰り返し適用する．全体の計算時間は，最後のk-meansに要した時間の3倍を超えないとしており，対象の分布によっては非常に少ない対象を参照するだけでクラスタを生成できる．

9. 「次元の呪い」の克服

§1 FastMap

4章で述べたように，クラスタリングは高次元空間では球面集中現象により意図した結果が得られない．この問題に対処する一つの方法は，多次元尺度法や主成分分析によって次元を縮退させることである．しかしこれらの手法の計算量$O(N^2)$なので，Nが大きい場合には実行できない．そこで，FaloutsosらはFastMap [Faloutsos 95] と呼ぶ，ヒューリスティックに基づく計算量$O(N)$の手法を提案している．縮退後の第1属性の値の計算方法を図3に示す．対象集合の中からできるだけ離れた2個の対象x_aとx_bを見つける．これらを通る直線$x_a x_b$に，ほかの対象x_iを射影することで，x_iの縮

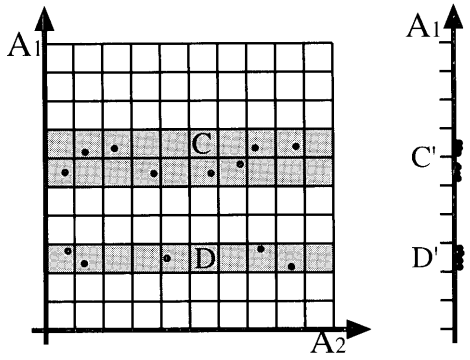


図5　CLIQUE のクラスタ [Agrawal 98]

退後の属性値 Y_i を求める．具体的には，図3のように，x_i から直線 x_ax_b への垂線の足を E とし，x_a と E の間の距離

$$Y_i=\frac{(D(x_a,x_i))^2+(D(x_a,x_b))^2-(D(x_b,x_i))^2}{2D(x_a,x_i)}$$

を x_i の縮退後の第1属性値とする．縮退後の第2属性の値は，図4のような，直線 x_ax_b に垂直な $d-1$ 次元の超平面 H を用いて求める．すなわち，対象を H に射影すれば，x_a と x_b が重なるので対象数が一つ減って $N-1$ に，また次元数も $d-1$ になるだけで，上記の方法が再び適用できる．ここで，対象の H 上での属性値は不要で，H 上での対象間の距離がわかれば十分である点が重要である．対象 x_i と x_j の H 上の距離 $D'(x_i,x_j)$ は，直角三角形 x_ix_jG を考えると

$$(D'(x_i,x_j))^2=(D(x_i,x_j))^2-(Y_i-Y_j)^2$$

の関係から容易に求められる．多次元尺度法のストレスを評価尺度とした実験では，同じ実時間では多次元尺度法を上回る精度を FastMap は達成できると述べている．ほかに，複数の軸を同時に処理する安らの HyperMap [安 02] や，Fisher 判別分析で次元を縮退させる末永らの研究 [末永 02] などもある．

§2　CLIQUE

次元の縮退とクラスタリングを同時に行う手法もいくつか提案されている．相関ルールの高速な発見手法 Apriori [Agrawal 94] で著名な Agrawal は Apriori で用いた性質をクラスタリングに適用した CLIQUE（CLustering In QUEst）[Agrawal 98] を提案している．CLIQUE は，図5のように空間を間隔 ξ で格子状に分割し，それぞれをユニットと呼ぶ．ユニット中の対象の密度がしきい値 τ より大きなものを密であるという．軸に平行な超矩形領域を考え，それらの交わりを選言標準形で表す．例えば，第 i 属性を A_i と記すと，

$$((4\le A_1<8)\wedge(2\le A_2<5))\vee$$
$$((2\le A_1<6)\wedge(4\le A_2<8))$$

のように領域を表す．領域中のユニットがすべて密であり，さらに，この領域が極大であるとき，この領域をクラスタとする．ここで図5左の2次元の空間を考えると，どのユニットも少数の対象しか含んでいない．よって，この2次元空間中ではクラスタは存在しない．しかし，属性 A_2 を無視し，属性 A_1 へ射影した部分空間（図の右）を考えると，密な領域 C' や D' が見いだせる．CLIQUE はこのような任意の部分空間中のクラスタを発見する．ここで，効率的な探索を行うために，Apriori と同様の単調性，すなわち，d 次元空間中のクラスタは，その任意の $d-1$ 次元部分空間でもクラスタであるという性質を利用する．アルゴリズムは，1次元空間での密なユニットを検出してクラスタを見つけ，次元数を一つ増やしてはクラスタを検出する手続きを繰り返す．密なユニットの情報だけを格納すればよいので必要な主記憶容量は少なくて済む．また，計算量は，$\tilde{d}$ を密なユニットが存在する最も高い次元，K を定数とすると，$O(K^{\tilde{d}}+N\tilde{d})$ である．$\tilde{d}$ に対して指数オーダであるが，実際には，$\tilde{d}$ が大きくなることはまれである．

§3　ORCLUS

Aggarwal らの ORCLUS（arbitrarily ORiented projected CLUSter generation）[Aggarwal 00] は，軸に平行射影だけではなく任意の方向の射影を扱う．この射影軸を見つけるために，共分散行列の特異値分解を用いる．この分解は主成分分析でも用いられ，できるだけ情報を保存する目的で固有値の大きな軸を選択する．ORCLUS では逆に，対象がまとまって射影される軸を見つけるために，固有値の小さな射影軸を選択する．アルゴリズムは，d 次元からはじめて，反復的に前の次元の β 倍の空間への射影を見つける．反復のたびに射影軸の類似したクラスタは併合され，その数は α 倍に減る．さらに，8章の BIRCH の CF を拡張した要約情報を用いて記憶容量を節約している．計算量は，最初の反復で探索するクラスタの数を k_0 とすると $O(k_0^3+k_0Nd+k_0^2d^3)$．k_0 が小さいと計算量は減少するが，精度は低下する．

10. 任意形状のクラスタを抽出する手法

k-means 法などの多くのクラスタリング手法は，クラスタの分布形状が超球や超楕円体であることを仮定している．そのため，図6の高密度部分で構成されるような複雑な形状のクラスタは抽出できない．本章では，このような形状のクラスタの抽出が可能な手法を紹介する．

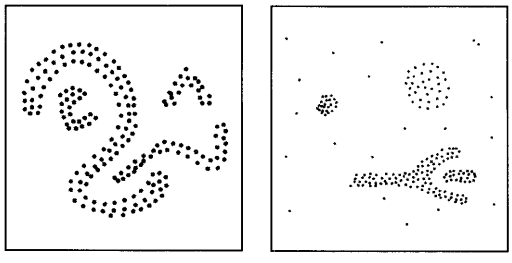


図6　複雑な形状のクラスタの例 [Ester 96]

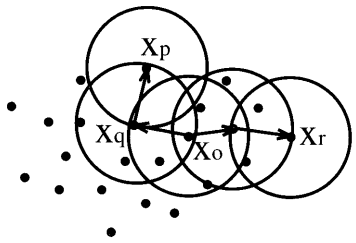


図7 DBSCAN [Ester 96]

事前に指摘しておくが，これらの手法はパラメータなどによっては，恣意的なクラスタが得られる場合がある．よって，クラスタの妥当性を視覚的に確認できる2または3次元のデータへの適用に限定することをすすめる．

§1 DBSCAN

任意形状のクラスタを抽出する近年の代表的な研究にEsterらのDBSCAN [Ester 96] がある．この手法のWebページが以下にあるので参考にされたい．

http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/

密度に基づく手法と著者は述べているが，対象の接続関係を利用した手法である．DBSCANは，距離のしきい値Epsと対象数のしきい値MinPtsという二つのパラメータを用いる．これらのパラメータで決定される図7のような対象の接続関係を定義し，接続している対象を同じクラスタに分類する．この接続関係をdirectly density-reachable（DDR）と呼び，次の条件を満たすとき対象x_pからx_qへDDRであるという．

(1) $x_q \in N_{\mathrm{Eps}}(x_p)$

(2) $|N_{\mathrm{Eps}}(x_p)| \geq \mathrm{MinPts}$

ただし，$N_{\mathrm{Eps}}(x_p)=\{x_q \in X \mid D(x_p, x_q) \leq \mathrm{Eps}\}$

この関係は対称ではなく，図7のx_qからx_pはDDRであるが，その逆は2番目の条件を満たさないのでDDRではないことに注意されたい．そして，図の$x_o \rightarrow x_r$や，$x_o \rightarrow x_q \rightarrow x_p$のようにDDRな関係で到達できる対象の集合で極大のものをクラスタとして抽出する．アルゴリズムは，任意のseedとなる対象から，DDR関係にある対象を順次，同じクラスタに分類する．そのクラスタ中の対象からDDRな対象を新たに見つけることができなければ，極大であるので，そのクラスタ中の対象を取り除き，新たなseedを選択する．図6左のような複雑な形状のクラスタも抽出が可能で，さらに，図の右のようにノイズが存在する場合でもMinPtsを適切に設定できれば，対象の密度の高い部分だけをクラスタとして抽出できる．R*-treeのような索引付け手法を用いれば，計算量は$O(N\log N)$である．DBSCANは対象が点だけに限定されているが，より一般的な距離や対象を利用できるGDBSCAN [Sander 98] や，Epsを変化させた場合のクラスタをまとめて抽出できるOPTICS [Ankerst 00] といった拡張もなされている．

§2 CURE

GuhaらのCURE [Guha 98] は，k-means法のようにセントロイド一つでクラスタを代表させる代わりに，複数の代表点を用いる手法を提案している．クラスタ中で互いに離れている対象をいくつか選択して代表点とする．そして，代表点をクラスタの中心方向にα倍だけ収縮した点を考え，これらの点との距離に基づき，代表点以外の対象が帰属するクラスタを決定する．アルゴリズムは凝集型階層的手法で，二つのクラスタの併合と代表点の更新を，クラスタ数がk個になるまで反復する．DBSCANはチェイニングのような現象により意図しない併合が起きる場合があるが，CUREではこの問題は回避できる．ただし，計算量は大きく，一般に$O(N^2\log N)$で，2次など低次元では$O(N^2)$である．

§3 グラフを用いた手法

その他，グラフを用いたクラスタリングも任意の形状のクラスタが抽出できる．先駆的なZahnの研究 [Zahn 71] は，minimum spanning tree（すべてのノードを接続するグラフで，その辺の長さ総和がが最小のもの）を求め，長い辺を切断することでクラスタを生成する．他には，Gabrielグラフやrelative neighborhoodグラフを用いたUrquhartの手法 [Urquhart 82]，k近隣を用いるMizoguchiらの手法 [Mizoguchi 80]，Karypisらの手法 [Karypis 99]，Harelらの手法 [Harel 01] がある．

その他，フラクタル次元を用いる手法 [Barbará 00] などもある．

11. 制約・教師信号を用いたクラスタリング

§1 クラスタ例からの学習

クラスタリングは，距離関数やポテンシャルの形で暗黙的に示された視点に基づいて「正しい」分割を獲得する．しかし，より明示的にこの視点を定めたい需要もある．このための枠組みとして，神嶌のクラスタ例からの

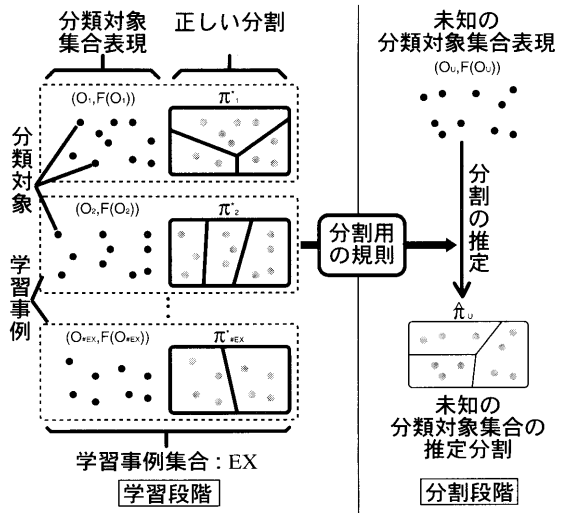


図8 クラスタ例からの学習

学習（Learning from Cluster Examples; LCE）[Kamishima 03a, 神嶌 03b] がある．LCE は，図 8 のように学習段階と分割段階がある．学習段階では，訓練事例集合から，適切な分割を獲得するための分割用の規則を獲得する．分割段階では，分割用の規則を用いて未知の分類対象集合の適切な分割を推定する．訓練事例は，分割される対象集合と，それに対する適切な分割の組である．すなわち，LCE は視点を訓練事例として示す枠組みである．LCE の解法の一つである LCE-MAP（LCE-Maximum A Posteriori）アプローチを以下に示す．対象集合を以下の 3 種類の属性ベクトルで記述する．

$A(x_i)$: **分類対象属性**　N 個の対象それぞれに一つずつあり，その対象の特徴を表現する．

$A(p_{ij})$: **分類対象対属性**　任意の対象 x_i と x_j の対 p_{ij} の特徴を表す．すべての対に一つずつあるので，$N(N-1)/2$ 個のベクトルがある．

$A(\pi)$: **分割属性**　分割後の分類対象集合全体の特徴を表す．

学習段階では，これらの属性の結合確率関数を学習する．

$$\Pr\left[\pi=\pi^*, A(\pi), \{A(x)\}, \{A(p)\}\right]$$

ただし，$\pi = \pi^*$ はある分割 π が適切な分割 π^* と一致するという事象である．この関数を $\Pr[A(\pi) \mid \pi = \pi^*]$ と $\Pr[\pi = \pi^* \mid \{A(x)\},\{A(p)\}]$ の積に分解して求める．分割段階では，事後確率最大原理に基づき，未知の対象集合についてこの関数を最大にするものを推定分割とする．

§2　COP-KMEANS

Wagstaff らは，獲得された分割に対して，同じクラスタに分類されるべき対象の対や，されてはならない対をユーザが指定し，これらの制約を満たすように再分割を繰り返すことで，ユーザが望む分割を対話的に得る手法を提案している．概念クラスタリングの COBWEB を対象にした手法[Wagstaff 00] と，k-means を対象にした COP-KMEANS [Wagstaff 01] があるが，後者を紹介する．k-means 法では，対象を割り当てる段階で，対象を一つずつクラスタに割り当てる．このとき，COP-KMEANS では，割当て済みの対象に関して，ユーザに与えられた制約が満たされなければ，そのクラスタへの割当てを中止し，その次に近いクラスタへの割当てを試みる．もし割当て可能なクラスタがなければ失敗を返すが，たとえ解が存在する場合でも失敗を返すことがあるという問題点がある．なお WWW 上でデモのページが

`http://www.cs.cornell.edu/home/wkiri/cop-kmeans/`

で公開されている．

§3　ラベルあり・なし混在データからの学習

クラスタリングというより，ラベルの集合が事前に与えられている点でクラス分類と呼ぶべき問題であるが，ラベルあり・なし混在データのクラス分類問題がある．これは，少数の一部の対象には教師によりラベルが与えられており，他の対象にはラベルがない場合に，ラベルなしデータを分類したり，両方のデータを用いてクラス分類器を獲得する問題である．この問題については，[Bensaid 96, Goldman 00, Nigam 98, 上田 01] などの研究がある．

12. ま　と　め

以上，第 1 部では，クラスタリングを利用する場合に参考になることがらをまとめ，第 2 部では，データマイニング分野の最近の研究を紹介した．著者は，未解決問題が多く残されており，また，需要も高い研究分野として以下のようなものがあると考えている．

- 時間とともに対象集合の内容が変化するデータストリーム[Barbará 02] の分類
- 単なる属性ベクトルではない，より複雑な時系列[Keogh 98, Ramoni 02, 山田 02] や構造をもつ対象[Taskar 01] の分類
- グラフスペクトルの利用[Ding 01] など，新たな最適化手法
- 11 章で述べたような，教師信号や制約を利用する手法 [Cohn 00]

本稿が，クラスタリングを利用・研究する方々に役立てばさいわいである．

謝　辞

本稿に関し貴重なコメントをいただいた森本康彦様（日本アイ・ビー・エム），藤井 敦先生（筑波大学），栗田多喜夫様（産業技術総合研究所）に感謝します．

◇ 参 考 文 献 ◇

[Aggarwal 00] Aggarwal, C. C. and Yu, P. S.: Finding Generalized Projected Clusters in High Dimensional Spaces, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp. 70-81（2000）

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. 20th Very Large Database Conf.,* pp. 487-499（1994）

[Agrawal 98] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P.: Automatic Subspace Clustering of High Dimensional Data for Data Mining Applicatiosn, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp. 94-105（1998）

[安 02] 安 際元, 古瀬一隆, 陳 漢雄, 石川雅弘, 于　旭, 大保信夫: HyperMap: 高次元空間における写像アルゴリズムとその次元縮小，クラスタリングへの応用，第 13 回データ工学ワークショップ（DEWS2002）, C2-10（2002）

[Ankerst 00] Ankerst, M., Breunig, M. M. and Sander, H.-P. K. J.: OPTICS: Ordering Points To Identify the Clustering Structure, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp. 49-60（2000）

[Barbará 00] Barbará, D. and Chen, P.: Using the Fractal Dimension to Cluster Datasets, *Proc. 6th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 260-264（2000）

[Barbará 02] Barbará, D.: Requirements for Clustering Data Streams, *SIGKDD Explorations,* Vol. 3, No. 2, pp. 23-27 (2002)

[Bensaid 96] Bensaid, A. M., Hall, L. O., Bezdek, J. C. and Clarke, L. P.: Partially Supervised Clustering for Image Segmentation, *Pattern Recognition,* Vol. 29, No. 5, pp. 859-871 (1996)

[Cohn 00] Cohn, D., Caruana, R. and McCallum, A.: Semi-supervised Clustering with User Feedback, http://www-2.cs.cmu.edu/~mccallum/papers/semiup-aaai2000s.ps.gz (2000)

[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)

[Domingos 01] Domingos, P. and Hulten, G.: A General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering, *Proc. 18th Int'l Conf. on Machine Learning,* pp.106-113 (2001)

[Ester 96] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proc. 2nd Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 226-231 (1996)

[Faloutsos 95] Faloutsos, C. and Lin, K.-I.: FastMap: A Fast Algorithm for Indexing,Data-Mining and Visualization of Traditional and Multimedia Datasets, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp. 163-174 (1995)

[Goldman 00] Goldman, S. and Zhou, Y.: Enhancing Supervised Learning with Unlabeled Data, *Proc. 17th Int'l Conf. on Machine Learning,* pp. 327-334 (2000)

[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)

[Harel 01] Harel, D. and Koren, Y.: Clustering Spatial Data Using Random Walks, *Proc. 7th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 281-286 (2001)

[Hinneburg 98] Hinneburg, A.and Keim, D. A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise, *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 58-65 (1998)

[Kamishima 03a] Kamishima, T. and Motoyoshi, F.: Learning From Cluster Examples, *Machine Learning* (2003), (in press)

[神嶌 03b] 神嶌敏弘, 赤穂昭太郎, 元吉文男: クラスタ例からの学習—クラスタ属性の利用, 人工知能学会誌論文誌, Vol. 18, No. 2, pp. 86-95 (2003)

[Karypis 99] Karypis, G., E.-H. Han, and Kumar, V.: Chameleon: Hierarchical Clustering Using Dynamic Modeling, *IEEE Computer,* Vol. 32, No. 8, pp. 68-75 (1999)

[Keogh 98] Keogh, E. J. and Pazzani, M. J.: An Enhanced Representation of Time Series Which Allows Fast And Accurate Classification, Clustering and Relevance Feedback, *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining,* pp. 239-243 (1998)

[Mizoguchi 80] Mizoguchi, R. and Shimura, M.: A Nonparametric Algorithm for Detecting Clusters Using Hierarchical Structure, *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 2, No. 4, pp. 292-300 (1980)

[Ng 94] Ng, R. T. and Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining, *Proc. 20th Very Large Database Conf.,* pp. 144-155 (1994)

[Nigam 98] Nigam, K., McCallum, A., Thrun, S. and Mitchell, T.: Learning to Classify Text from Labeled and Unlabeled Documents, *Proc. 15th National Conf. on Artificial Intelligence,* pp.792-799 (1998)

[Ramoni 02] Ramoni, M., Sebastiani, P. and Cohen, P.: Bayesian Clustering by Dynamics, *Machine Learning,* Vol. 47, pp. 91-121 (2002)

[Sander 98] Sander, J., Ester, M., Kriegel, H.-P. and Xu, X.: Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications, *Journal of Data Mining and Knowledge Discovery,* Vol. 2, pp. 169-194 (1998)

[Sheikholeslami 98] Sheikholeslami, G., Chatterjee, S. and Zhang, A.: WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, *Proc. 24th Very Large Database Conf.,* pp. 428-439 (1998)

[末永 02] 末永高志, 佐藤 新, 坂野 鋭: クラスタ構造に着目した特徴空間の可視化—クラスタ判別法—, 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 5, pp. 785-795 (2002)

[Taskar 01] Taskar, B., Segal, E. and Koller, D.: Probabilistic Classification and Clustering in Relational Data, *Proc. 17th Int'l Joint Conf. on Artificial Intelligence,* pp. 870-876 (2001)

[上田 01] 上田修功: 拡張結合混合モデル, 第4回情報論的学習理論ワークショップ, pp. 89-94 (2001)

[Urquhart 82] Urquhart, R.: Graph Theoretical Clustering Based on Limited Neghbourhood Sets, *Pattern Recognition,* Vol. 15, No. 3, pp. 173-187 (1982)

[Wagstaff 00] Wagstaff, K. and Cardie, C.: Clustering with Instance-level Constraints, *Proc. 17th Int'l Conf. on Machine Learning,* pp. 1103-1110 (2000)

[Wagstaff 01] Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained K-means Clustering with Background Knowledge, *Proc. 18th Int'l Conf. on Machine Learning,* pp. 577-584 (2001)

[Wang 97] Wang, W., Yang, J. and Muntz, R.: STING: A Statistical Information Grid Approach to Spatial Data Mining, in *Proc. of the 23rd Very Large Database Conf.,* pp. 186-195 (1997)

[Xu 98] Xu, X., Ester, M., Kriegel, H.-P. and Sander, J.: A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases, *Proc. 14th Int'l Conf. on Data Engineering,* pp. 324-331 (1998)

[山田 02] 山田 悠, 中本和岐, 鈴木英之進: 動的時間伸縮法に基づく時系列データの高速クラスタリング, 第1回情報科学技術フォーラム講演論文集, LG-3 (2002)

[Zahn 71] Zahn, C. T.: Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters, *IEEE Trans. on Computers,* Vol. 20, No. 1, pp. 68-86 (1971)

[Zhang 96] Zhang, T., Ramakrishnan, R. and Livny, M.: BIRCH: An Efficient Data Clustering Method for Very Large Databases, *Proc. ACM SIGMOD Int'l Conf. on Management of Data,* pp.103-114 (1996)

[Zhang 97] Zhang, T., Ramakrishnan, R. and Livny, M.: BIRCH: A New Data Clustering Algorithm and Its Applications, *Journal of Data Mining and Knowledge Discovery,* Vol. 1, pp. 141-182 (1997)

2002年12月24日 受理

著者紹介

神嶌 敏弘（正会員）は，前掲（Vol.18 , No.1 , p.65）参照.