クラスタリング

Clustering

神嶌 敏弘

http://www.kamishima.net/

クラスタリング

クラスタリング (clustering)

クラスター分析 (cluster analysis)
データクラスタリング (data clustering)

- データの集合を**クラスタ**という部分集合に分けること
- 代表的な**教師なし学習**手法

教師なし学習

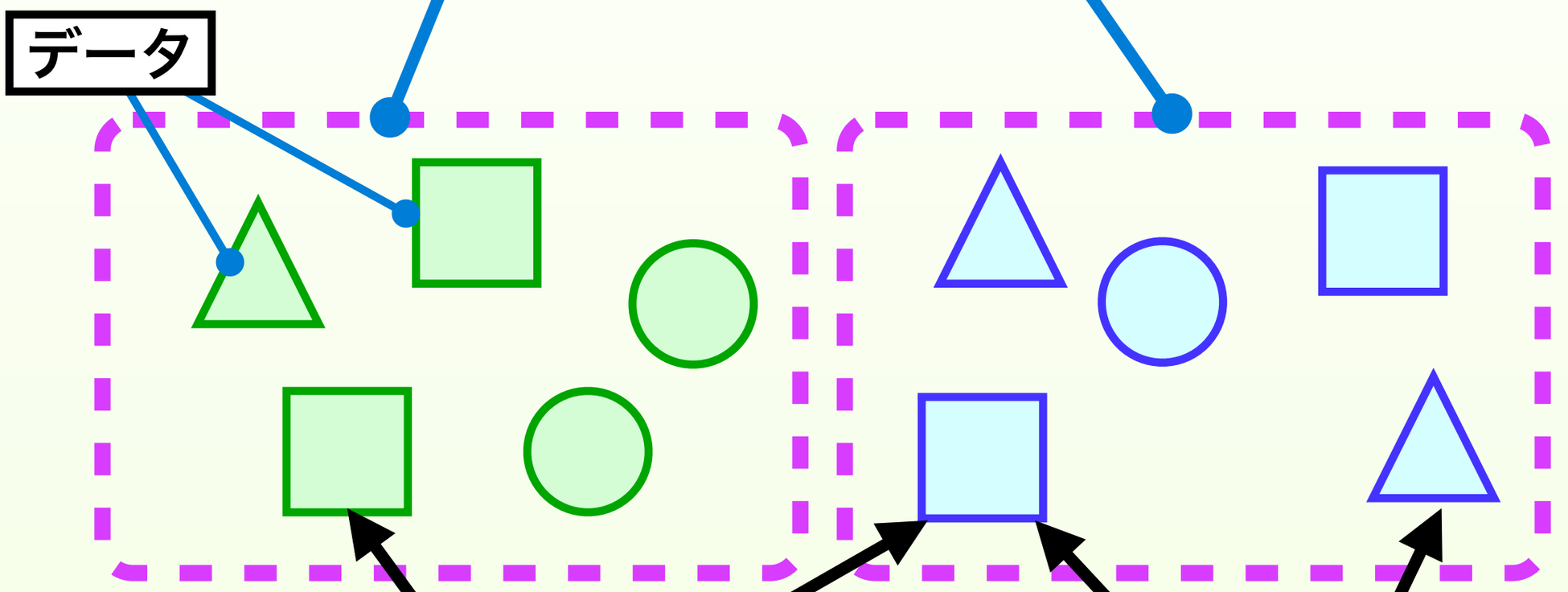
人間が与えた正解は不要で，観測データだけを対象に分析を行う

教師あり学習

クラス分類など，人間が与えた正解を，観測データから予測する規則を学習

クラスタ

クラスタ (cluster)
内的結合と外的分離の性質を持つデータの部分集合

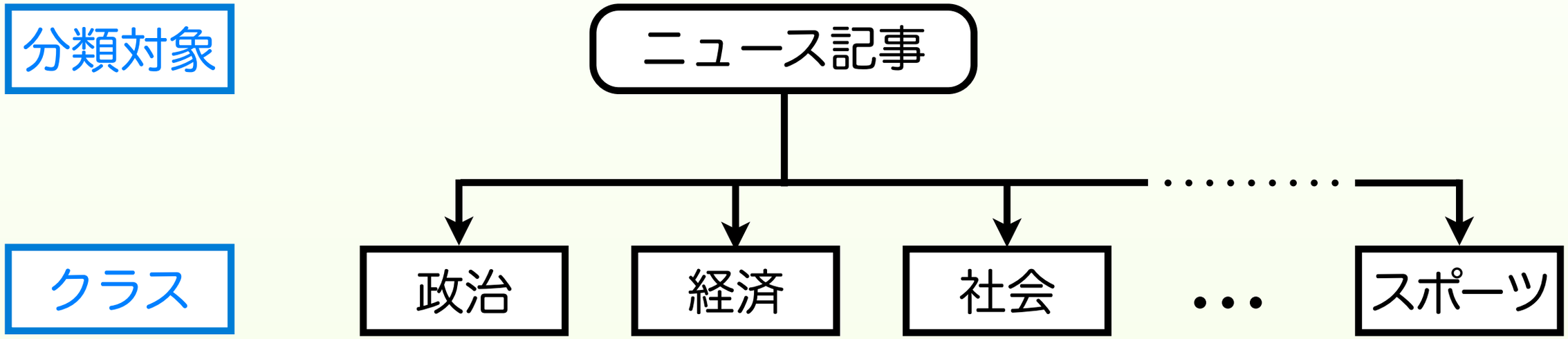


外的分離
external isolation
違うクラスタにある対象は似ていない

内的結合
internal cohesion
同じクラスタ内の対象は互いに似ている

クラスタリングと分類

分類・識別・クラス分類
classification, discrimination



分類対象それぞれを，事前に定めたクラスに割り当てる

クラス

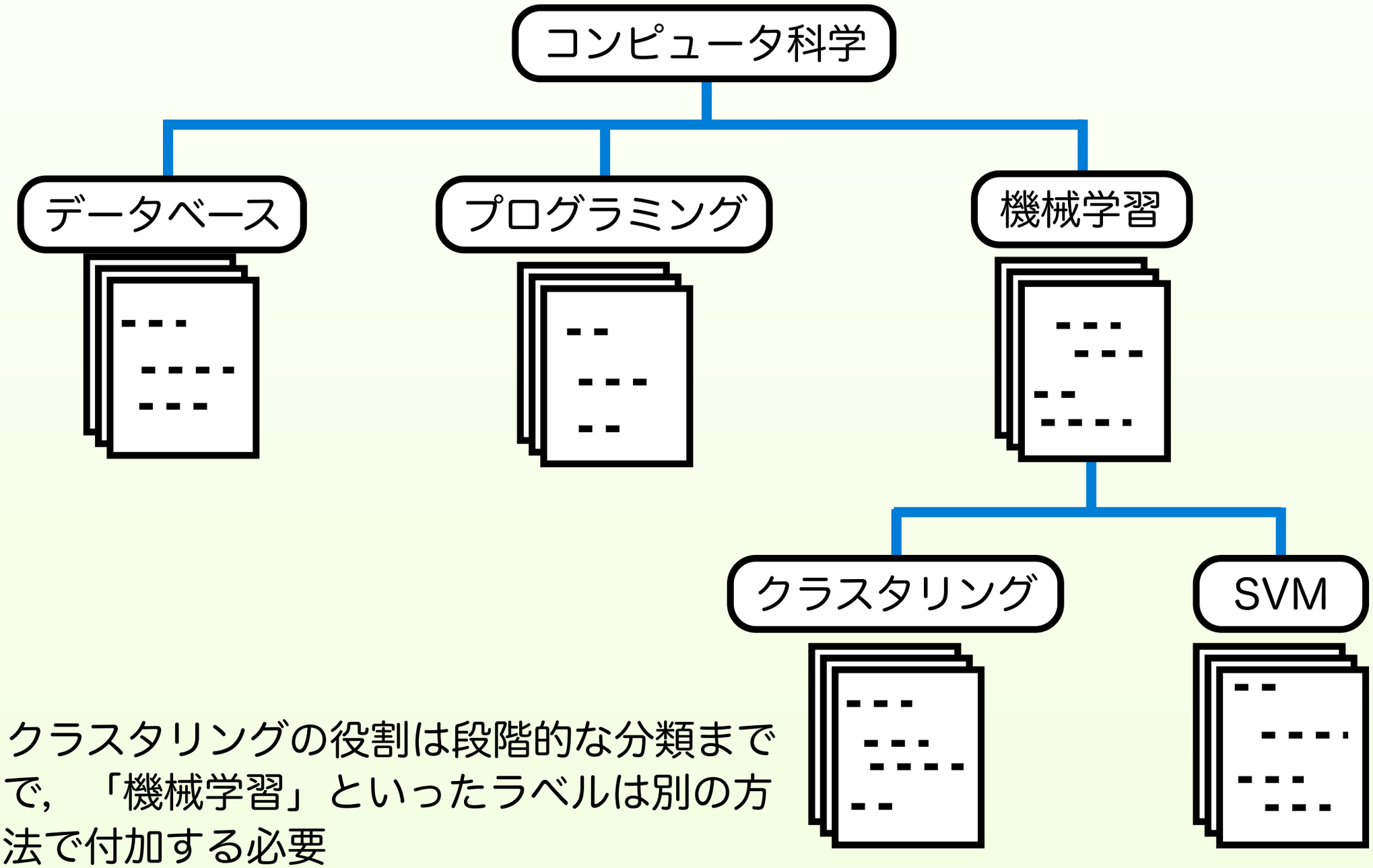
- 人間が事前に決めておくグループ
- 各グループは意味付けされている

クラスタ

- 似たものを集めた結果として出来るグループ
- 各グループの意味は後から解釈する

利用例：文書のクラスタリング

大量の文書を，クラスタリングで分類し，これらの文書の特徴を把握



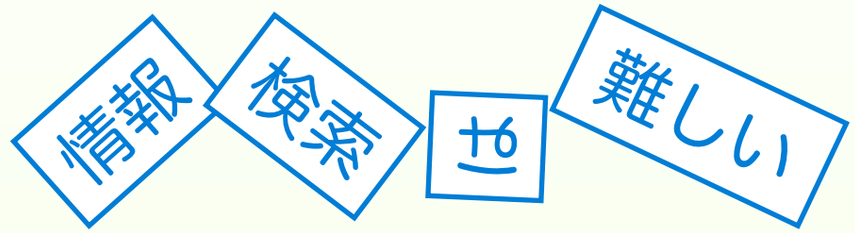
※ クラスタリングの役割は段階的な分類までで，「機械学習」といったラベルは別の方法で付加する必要

BOW表現とベクトル空間モデル

Bag of Words (BOW)表現

単語の順序関係を無視して，文書を単語の集まりとみなす

情報検索は難しい ➡



ベクトル空間モデル (vector space model)

単語をベクトルの各要素に割り当て，数値化した重みで表現

文書ベクトル

猫　犬　猿　寝る　食べる

$$\mathbf{d}_i = [\ w_{i1}\ ,\ w_{i2}\ ,\ w_{i3}\ ,\ \ldots\ ,\ w_{ij}\ ,\ \ldots\ ,\ w_{iT}\ ]$$

T：単語の総数

重みとしては索引語頻度，TF-IDF などが代表的

TF(索引語頻度)とコサイン距離

TF (Term Frequency; 索引語頻度)

文書内でその単語が生じる頻度
「何度も繰り返し言及される概念は重要な概念」

例：親亀の上に子亀を乗せた．子亀の上に孫亀を乗せた

親亀＝1回　子亀＝2回　孫亀＝1回　上＝2回　乗せる＝2回

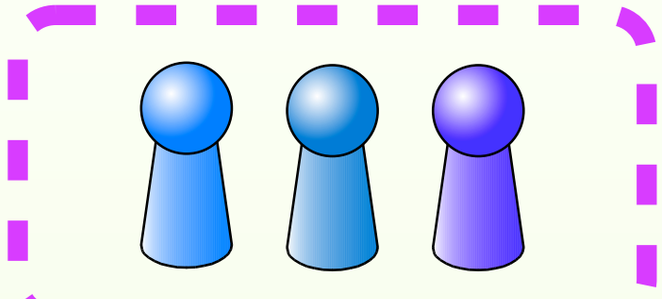
文書間の距離　文書 i と文書 j の非類似度を，文書ベクトル $\mathbf{d}_\mathrm{i}$ と $\mathbf{d}_\mathrm{j}$ の間のコサインで測る

$$d(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|}$$

非類似度行列を凝集型階層的クラスタリング手法に入力

利用例：小売のマーケティング

スーパーのポイント会員の購買記録を使って，顧客を分類
各グループごとに，適切な販売戦略を立てる



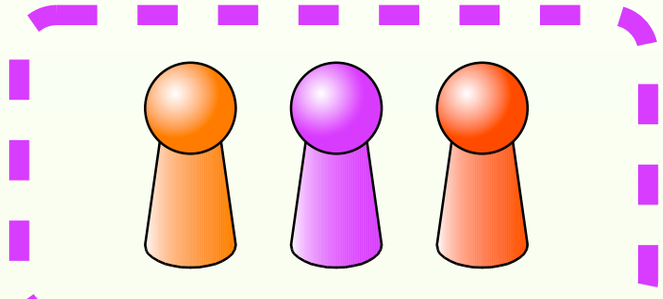
主に生鮮食料品を
購入する顧客

↓

バーゲンなどでまとめ買い

↓

ダイレクトメール



主に中食製品を
購入する顧客

↓

その場で購入するものを決める

↓

セットメニューによる客単価の向上

利用例：小売のマーケティング

スーパーのポイント会員を，その購買傾向が似ているクラスタに分割

i 番目の会員 $\mathbf{x}_i$ を，商品カテゴリ別の，先月の合計購入価格を要素とする **特徴ベクトル** で表現

第1属性　x_{i1}　：先月の弁当・総菜の購入額

第2属性　x_{i2}　：先月の清涼飲料の購入額

⋮

第m属性　x_{im}　：先月の野菜類の購入額

例： 会員3番 $\mathbf{x}_3$ ＝〈3000円, 1200円, …, 0円〉

特徴ベクトルを分割最適化型クラスタリング手法に入力

クラスタリング手法の分類

- 階層的クラスタリング (hierarchical clustering)
  - 凝集型階層的クラスタリング (agglomerative hierarchical clustering)
    データ一つが個々のクラスタの状態から，順次クラスタを併合し，クラスタの階層を生成する
  - 分割型階層的クラスタリング (divisive hierarchical clustering)
    データ集合全体が一つのクラスタの状態から，順次クラスタを分割して，クラスタの階層を生成する

- 分割最適化クラスタリング (partitional optimization clustering)
  別名：partitional clustering, non-hierarchical clustering
  クラスタの良さを表す関数を定義し，その関数を最適化するようなクラスタを見つけ出す

距離 (非類似度)

距離 (distance) / 非類似度 (dissimilarity)

- データやクラスタの似ていなさを示す数値
- 距離の公理を満たしていることが望ましい

距離行列 (distance matrix)

- クラスタ i と j の間の距離を要素 d_{ij} とする行列
- 距離の公理を満たしているなら，対角要素 $d_{ii} = 0$

$$D = [d_{ij}]$$

データやクラスタが，距離行列ではなく，特徴ベクトルで表されているときにはどうする？

距離：実数値特徴ベクトル

特徴ベクトルの要素が実数値の場合 $\mathbf{x}_\mathrm{i}$ と $\mathbf{x}_\mathrm{j}$ の距離 d_ij

K 次元特徴ベクトル： $\mathbf{x}_i = [x_{i1}, x_{i1}, \ldots, x_{iK}]^\top$

ユークリッド距離
Euclidean distance

$$\left[\sum_{k=1}^{K}(x_{ik} - x_{jk})^2\right]^{1/2}$$

一番よく使われる
L_2 距離ともいう

シティブロック距離
city-block distance

$$\sum_{k=1}^{K}|x_{ik} - x_{jk}|$$

はずれ値に対して頑健
L_1 やマンハッタン距離などとも

ミンコフスキー距離
Minkowski distance

$$\left[\sum_{k=1}^{K}|x_{ik} - x_{jk}|^p\right]^{1/p}$$

L_2距離などの一般化
L_p 距離ともいう

マハラノビス距離
Mahalanobis distance

$$\left[(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})\right]^{1/2}$$

$\bar{\mathbf{x}}$ と Σ はサンプルの平均と共分散行列

ベクトルの要素ごとの分散
に大きな差があるとき

コサイン類似度
cosine similarity

$$\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}$$

距離とは逆向きの類似度
ベクトルの角に意味がある文書ベクトルなどに

距離：カテゴリ値特徴ベクトル

特徴ベクトルが {大,中,小} など有限の値のどれかをとるカテゴリ値

距離/非類似度ではなく，類似度なので次式で変換して用いる

$$d_{ij}=1 - s_{ij}$$

単純一致係数 (simple matching coefficient)

$$\frac{1}{K}\sum_{k=1}^{K}\mathrm{I}[x_{ik}=x_{jk}]$$

I[条件] は条件が成立したとき1，そうでないとき0をとる指示関数

Jaccard係数 (Jaccard coefficient)

$$\frac{\sum_{k=1}^{K}\mathrm{I}[x_{ik}=1\wedge x_{jk}=1]}{K-\sum_{k=1}^{K}\mathrm{I}[x_{ik}=0\wedge x_{jk}=0]}$$

0/1 のいずれかの値をとるカテゴリ値で，特に値が 1 であることに注目している場合．例えば，マーケティングで1がある商品の購入を示す場合など

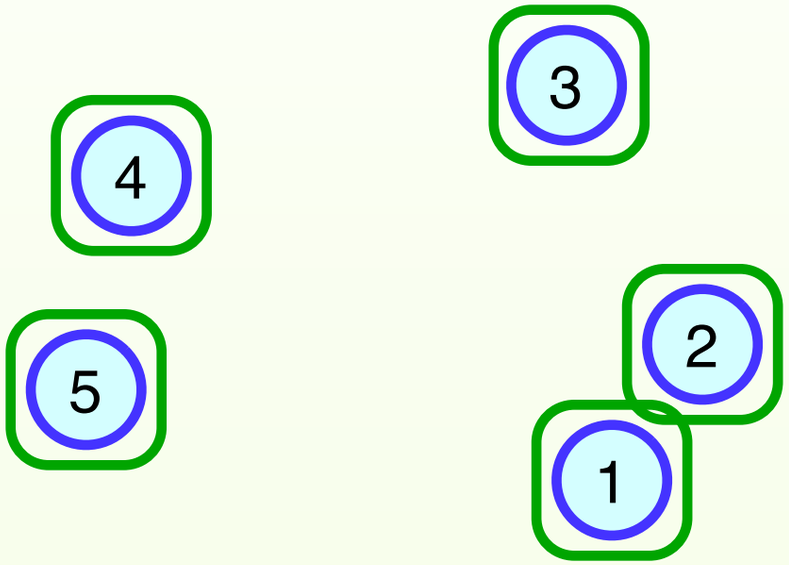
※ このような場合を非対称な二値変数という

凝集型階層的クラスタリング

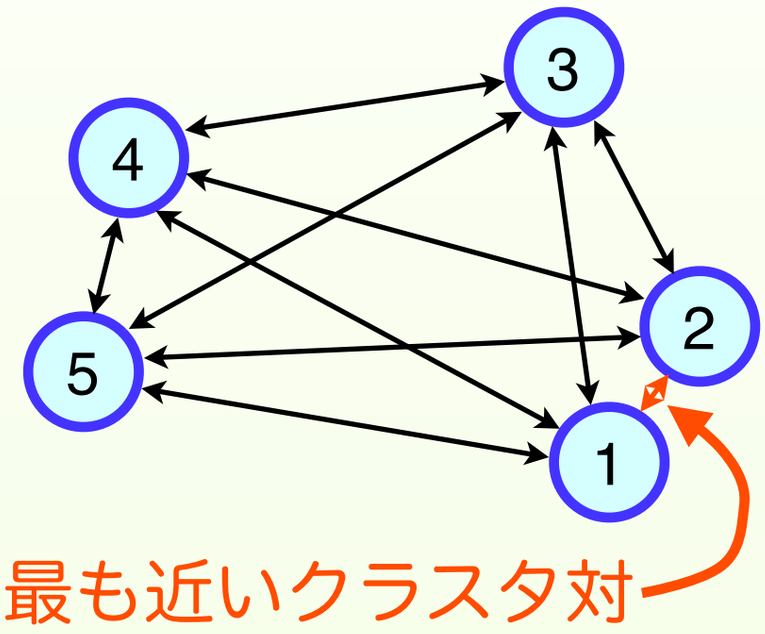
Agglomerative Hierarchical Clustering

凝集型階層的クラスタリング

凝集型階層的クラスタリング
agglomerative hierarchical clustering



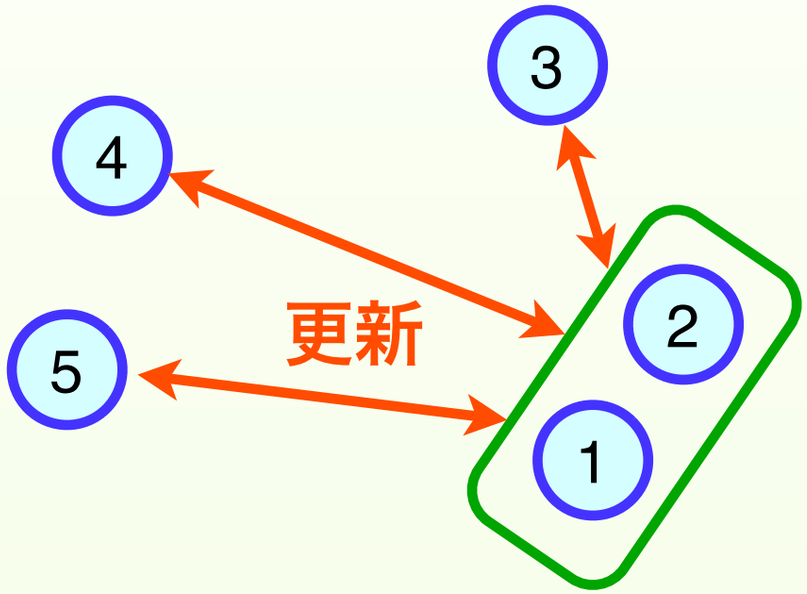
1 個々のデータが，それぞれ孤立したクラスタを形成している状態から開始



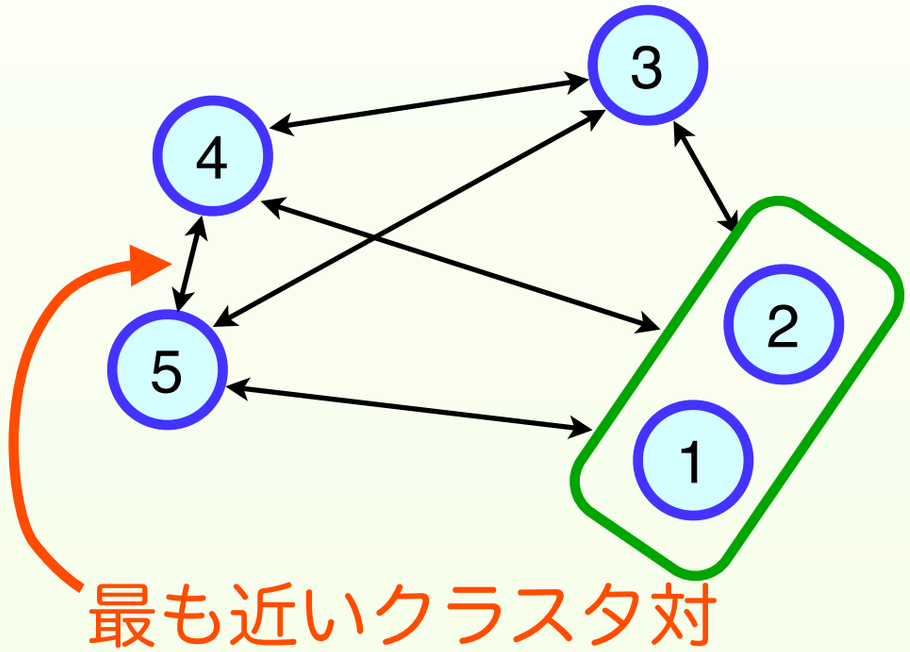
2 全てのクラスタ対の間の距離を計算し，最も近いクラスタ対を見つけ出す

凝集型階層的クラスタリング

凝集型階層的クラスタリング
agglomerative hierarchical clustering



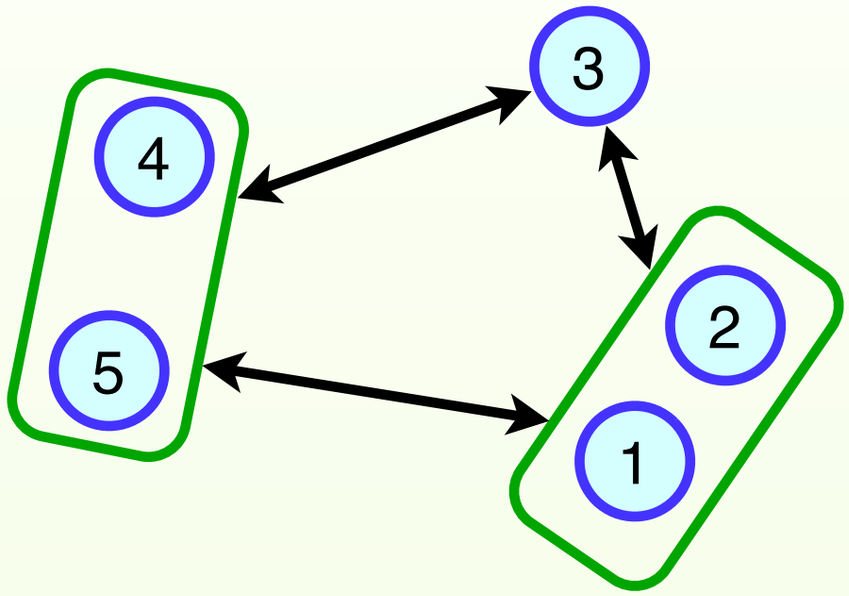
3 最も近いクラスタを併合し，この新しいクラスタと他のクラスタの距離を計算



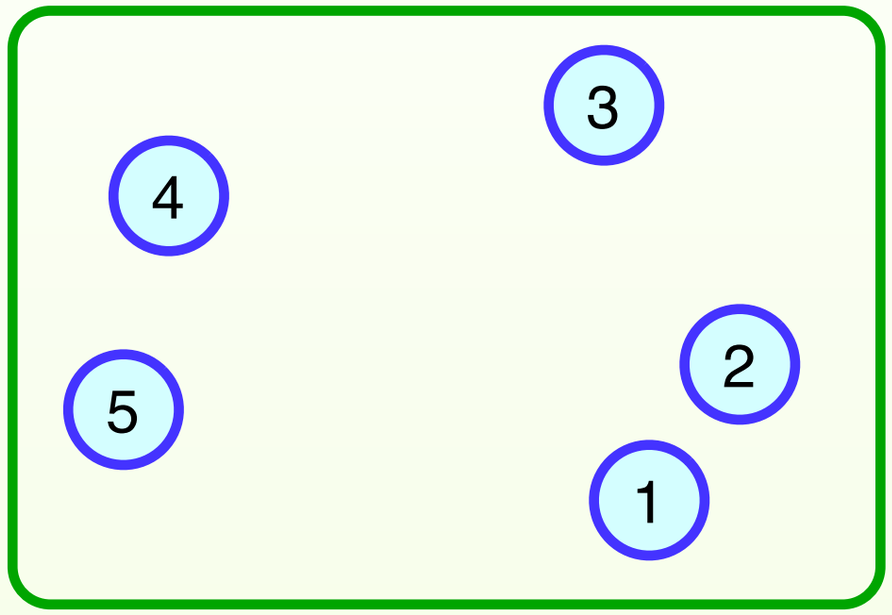
4 新クラスタを含めたクラスタ対の中で一番近い対を見つけ出す

凝集型階層的クラスタリング

凝集型階層的クラスタリング
agglomerative hierarchical clustering



5 同様にクラスタの併合と，距離の再計算を繰り返す

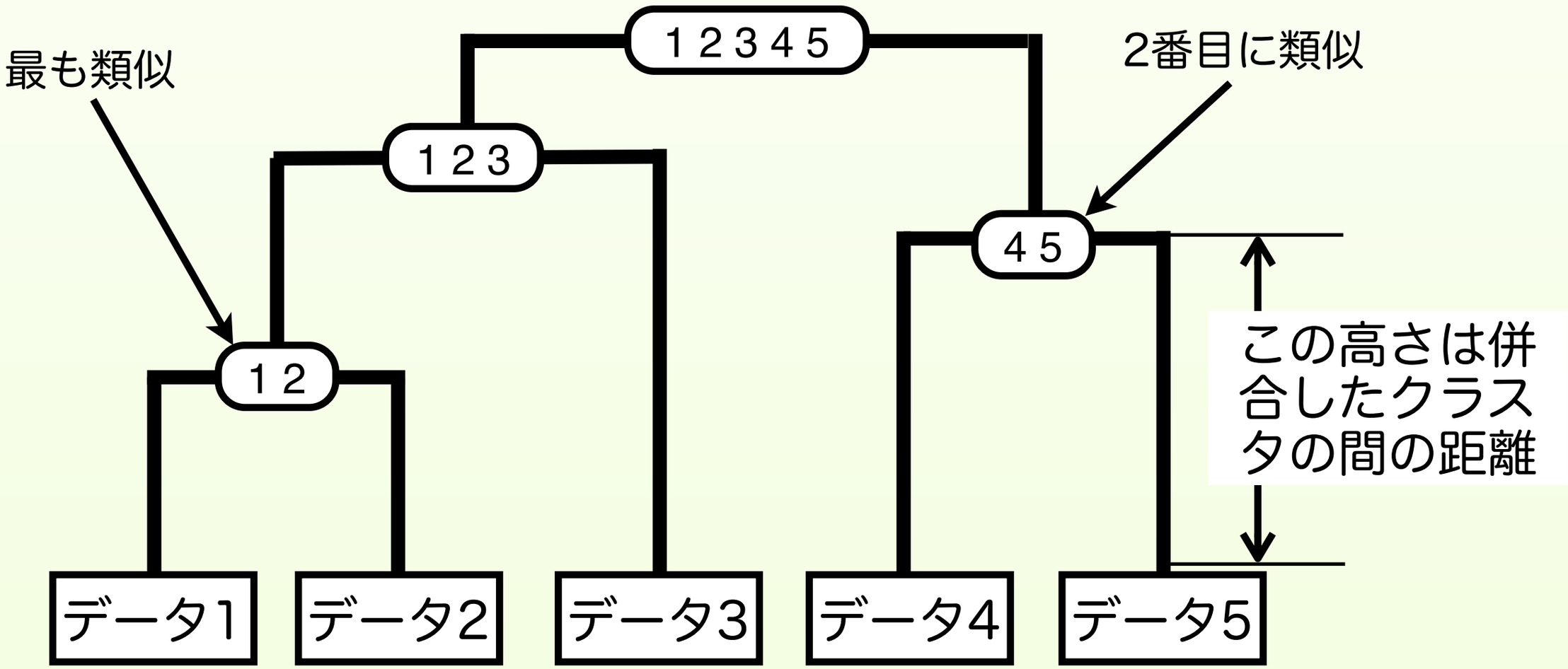


6 全体が一つのクラスタになった時点で終了

デンドログラム

デンドログラム(樹状図, 樹形図; dendrogram)

- クラスタリングの過程を図に示したもの
- クラスタの併合の様子を木で表現



距離/非類似度の更新

クラスタ i と j の併合した新クラスタ n と，他クラスタ k との距離
この距離の更新方法の違いで凝集型階層的クラスタリングを分類

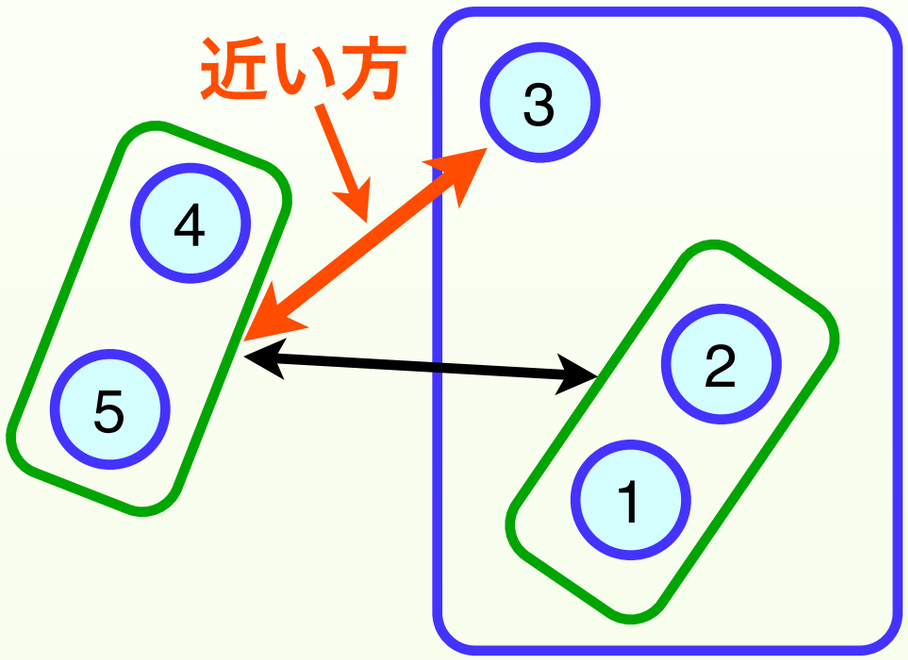
- **単リンク法** (single linkage method)
  最短距離法 (minimum distance method)
- **完全リンク法** (complete linkage method)
  最長距離法 (maximum distance method)
- **群平均法** (group average method)
  UPGMA (unweighted pair-group using arithmetic averages)
- **ウォード法** (Ward’s method)
  最小分散法 (minimum variance mthod)
- **重心法** (セントロイド法; centroid method)
  UPGMC (unweighted pair-group using centroids)
- **重み付き平均法** (weighted average method)
  WPGMA (weighted pair-group using arithmetic averages)
- **メジアン法** (median method)
  WPGMC (unweighted pair-group using centroids)

※ 簡単な更新式が利点であったWPGMAやメジアン法は計算機の発達であまり使われなくなった

※ UPGMA / WPGMA の “using arithmetic averages” は “using average linkages” とする場合も

単リンク法 / 最短距離法

クラスタ C_a={1,2} と C_b={3} を併合し C_c={1,2,3} を生成するとき
既存の C_k={4,5} と新しい C_c={1,2,3} の間の距離の決め方



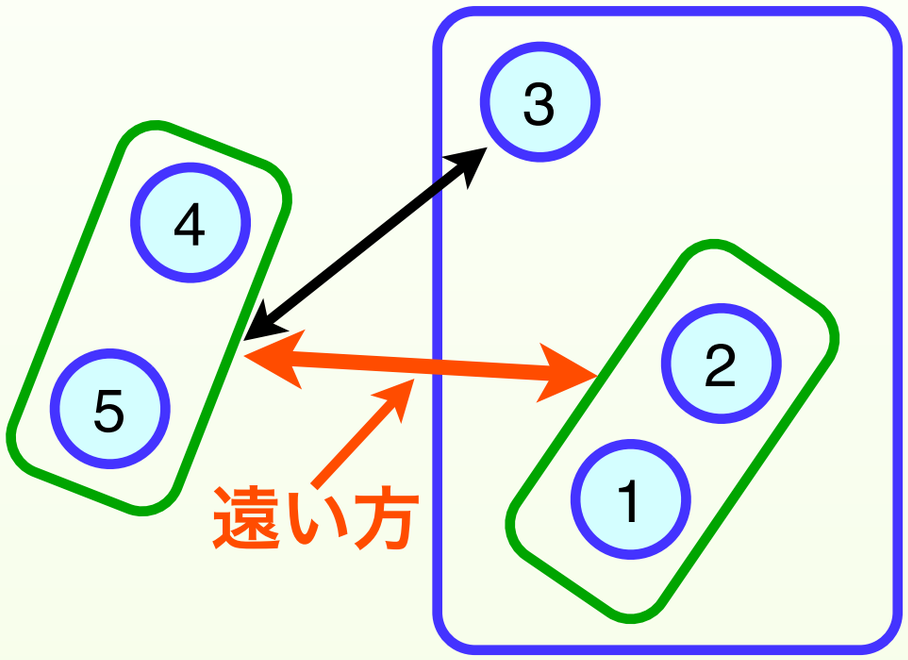
$$d_{kc} = \min\{d_{ka}, d_{kb}\}$$

併合前の近い方のクラスタまでの距離を
併合後のクラスタまでの距離とする

- ある距離以下のデータ対にリンクを作り，一つでもリンクがあるときは同じクラスタにまとめるのと等価（単リンク法の由来）
- 最小全域木と密接な関連がある
- 実用上はずれ値に弱い問題
- 背後に branching random walk があるため，細長いクラスタが出やすい

完全リンク法 / 最長距離法

クラスタ C_a={1,2} と C_b={3} を併合し C_c={1,2,3} を生成するとき
既存の C_k={4,5} と新しい C_c={1,2,3} の間の距離の決め方



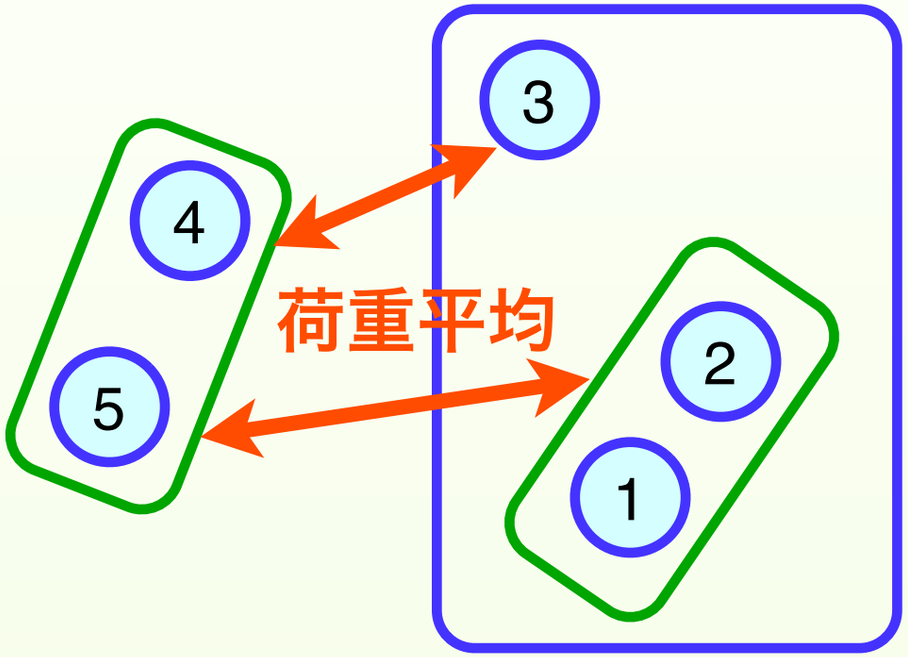
$$d_{kc} = \max\{d_{ka}, d_{kb}\}$$

併合前の遠い方のクラスタまでの距離を
併合後のクラスタまでの距離とする

- ある距離以下のデータ対にリンクを作り，完全グラフになっている部分(極大クリーク)を同じクラスタにまとめるのと等価（完全リンク法の由来）
- クラスタの大きさが揃いやすい
- 実用上はずれ値に弱い
- 背後に超球内の均一分布があるため，球状のクラスタが出やすい

群平均法 / UPGMA

クラスタ C_a={1,2} と C_b={3} を併合し C_c={1,2,3} を生成するとき
既存の C_k={4,5} と新しい C_c={1,2,3} の間の距離の決め方



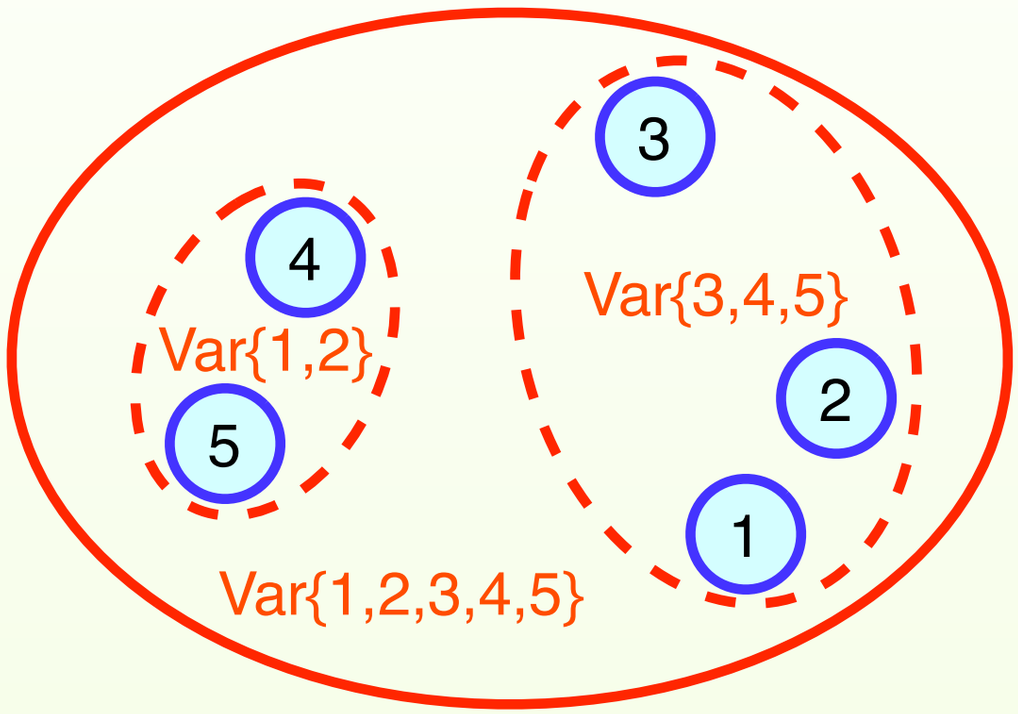
$$d_{kc} = \frac{|C_a|}{|C_c|} d_{ka} + \frac{|C_b|}{|C_c|} d_{kb}$$

元のクラスタへの距離を，元のクラスタの大きさに比例した重みをつけてとった算術平均が，併合後のクラスタとの距離

- 新クラスタからデータを一つ，他クラスタからも一つ選んだデータ対の，全ての対にわたる距離の算術平均を更新後の距離とするのと等価
- 単リンク法と完全リンク法の中間的なクラスタを導く
- はずれ値に強く，実用的に便利

ウォード法 / 最小分散法

クラスタ C_a={1,2} と C_b={3} を併合し C_c={1,2,3} を生成するとき
既存の C_k={4,5} と新しい C_c={1,2,3} の間の距離の決め方



$$d_{kc} = \mathrm{Var}(C_k \cup C_c) - \Big(\mathrm{Var}(C_k) + \mathrm{Var}(C_c)\Big)$$

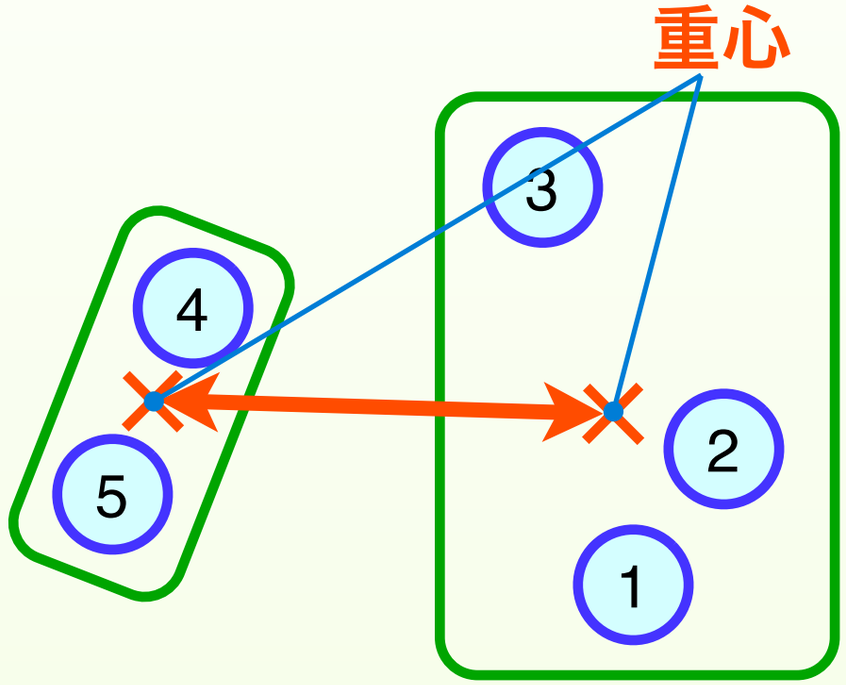
$\mathrm{Var}(C)$：クラスタ C 内のデータの分散

併合後のクラスタの分散と，併合前のクラスタそれぞれの分散の和との差が最小になるクラスタの対を併合する

- はずれ値に強く，実用的に便利
- ユークリッド距離の2乗で距離を測る
- 等方分散のガウス分布が背後にあるため，球状のクラスタが出やすい

重心法 / UPGMC

クラスタ C_a={1,2} と C_b={3} を併合し C_c={1,2,3} を生成するとき
既存の C_k={4,5} と新しい C_c={1,2,3} の間の距離の決め方



$$d_{kc} = \frac{|C_a|}{|C_c|} d_{ka} + \frac{|C_b|}{|C_c|} d_{kb} - \frac{|C_a||C_b|}{|C_c|^2} d_{ab}$$

それぞれのクラスタの重心の間の，ユークリッド距離の2乗

- ユークリッド距離の2乗で距離を測る

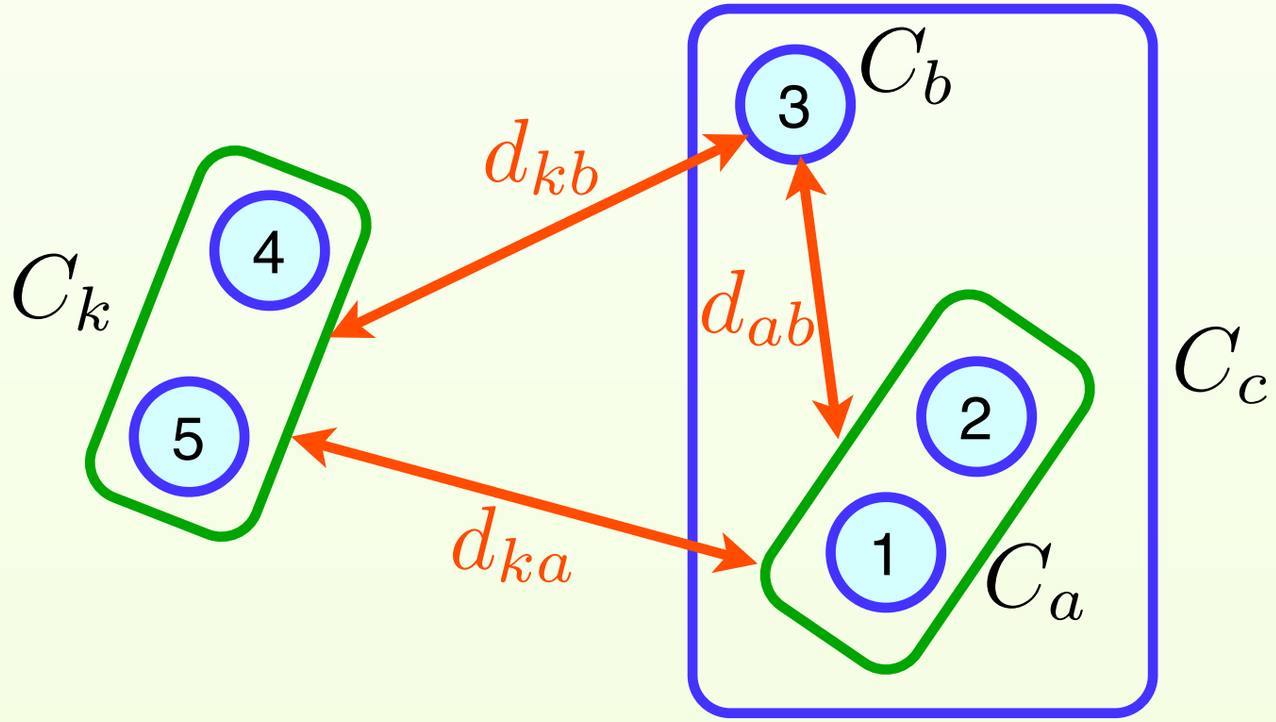
Lance-Williamsの更新式

[Lance+ 67]

Lance-Williamsの更新式
Lance-Williams updating formula

$$d_{kc} = \alpha_a d_{ka} + \alpha_b d_{kb} + \beta d_{ab} + \gamma |d_{ka} - d_{kb}|$$

凝集型階層的クラスタリングの距離の更新式を統一的に表現できる式
距離の更新が定数時間 O(1) で可能になる



※ Lance-Williams recurrence formula ともいう

Lance-Williamsの更新式

$$d_{kc} = \alpha_a d_{ka} + \alpha_b d_{kb} + \beta d_{ab} + \gamma |d_{ka} - d_{kb}|$$

係数 α_a, α_b, β, γ の一覧

	α_a	α_b	β	γ
単リンク	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
完全リンク	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
群平均	$\frac{\|C_a\|}{\|C_c\|}$	$\frac{\|C_b\|}{\|C_c\|}$	0	0
重み付き平均	$\frac{1}{2}$	$\frac{1}{2}$	0	0
重心	$\frac{\|C_a\|}{\|C_c\|}$	$\frac{\|C_b\|}{\|C_c\|}$	$-\frac{\|C_a\|\|C_b\|}{\|C_c\|^2}$	0
メジアン	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
ウォード	$\frac{\|C_k\|+\|C_a\|}{\|C_k\|+\|C_c\|}$	$\frac{\|C_k\|+\|C_b\|}{\|C_k\|+\|C_c\|}$	$-\frac{\|C_k\|}{\|C_k\|+\|C_c\|}$	0

空間の拡散 / 濃縮

[Dubien+ 79]

新たに出来たクラスタが，加速度的に，他のクラスタを併合しやすくなる性質を **空間濃縮**，逆に併合しにくくなる性質を **空間拡散**

Lance-Williams の更新式で，次の条件下で β を変えるのが可変法
β を 1 に近づけると空間濃縮が，小さくすると空間拡散が生じる

$$\alpha_a + \alpha_b + \beta = 1 \qquad \alpha_a = \alpha_b \qquad \beta < 1 \qquad \gamma = 0$$

空間濃縮	空間保存	空間拡散
単リンク	群平均，重み付き平均 重心，メジアン	完全リンク ウォード

大 ⟵ クラスタの大きさの散らばり ⟶ 少

チェイニング：単リンク法は，鎖状にデータが存在するときそれらが順次大きなクラスタに取り込まれる現象

※ 空間の濃縮 / 拡散 (space dilation/extraction) は，縮小 / 拡大とも

単調性とデンドログラムの反転

[Batagelj 81]

クラスタ C_a と C_b を併合してできた C_c と，別のクラスタ C_k との距離について次の条件が成立するとき **単調 (monotone)** という

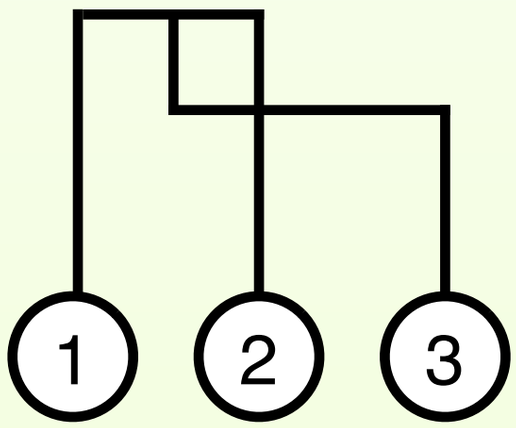
$$d_{kc} \geq d_{ab}$$

距離が単調である，Lance-Williamsの更新式の係数の必要十分条件

$$\gamma \geq -\min\{\alpha_a, \alpha_b\},\ \alpha_a + \alpha_b \geq 0,\ \alpha_a + \alpha_b + \beta \geq 1$$

重心法とメジアン法は単調ではないが，他の単リンク法などは単調

デンドログラムの反転(reversal)/交差(crossover)



単調でない場合は，距離が減少することがあるため，デンドログラムが交差することがある

例：クラスタ 1 と 2 を併合後に，クラスタ 3 と併合すると距離が減少する場合のデンドログラム

アルゴリズム

Lance-Williamsの更新式のように距離の更新が定数時間（O(1)）でできるなら，ヒープを用いた優先度付きキューを使って，一般の凝集型階層的クラスタリングの計算量は $O(n^2 \log n)$ となる

1. n 個のデータについて，それから一番近いデータを得るためのヒープを生成 [$O(n^2) = O(n) \times n$]
2. (n - 1) 回併合を繰り返す
    1. 最も近いクラスタ対を見つける [$O(n) = n \times O(1)$]
    2. n個のクラスタとの間の距離の更新 [$O(n) = n \times O(1)$]
    3. ヒープの更新 [$O(n \log n) = n \times (2 \log n + \log n)$]

※ ヒープの計算量は初期化 O(n)，挿入・削除 log(n)，先頭参照 O(1) とする
※ [ ] 内は，該当ステップの計算量

可約性を用いたアルゴリズム

[Murtagh 83]

距離の更新が定数時間　O(1)　でき，距離に可約性の性質がある凝集型階層的クラスタリングの計算量は $O(n^2)$ となる

可約性　Ca と Cb を併合した Cc と，他の Ck の距離に関する性質

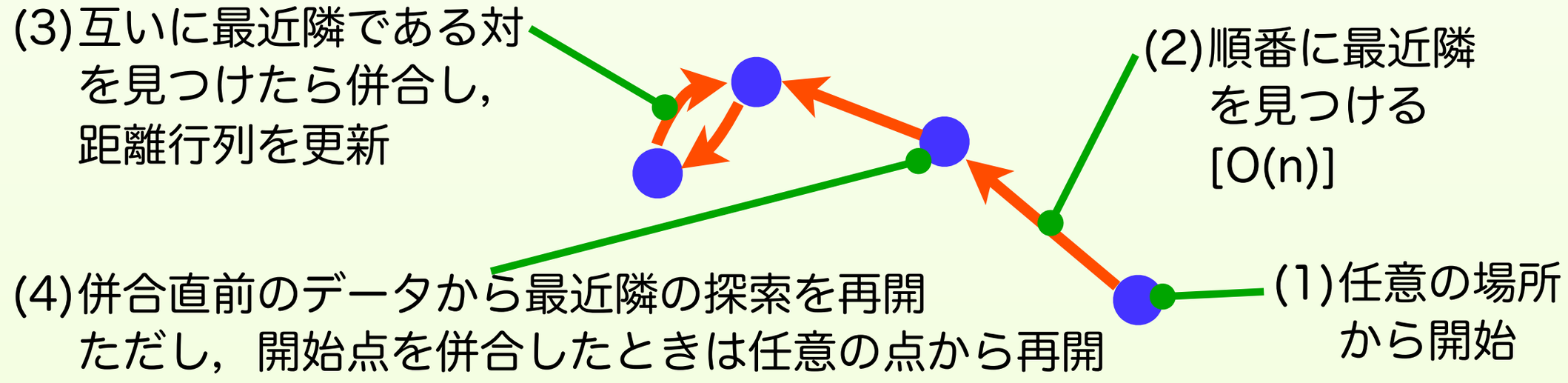
$$d_{ab} \leq \min\{d_{ka}, d_{kb}\} \Rightarrow \min\{d_{ka}, d_{kb}\} \leq d_{kc}$$

可約なとき，併合後のクラスタが元のクラスタより近づくことはない

重心法とメジアン法以外の五つの方法は可約性をもつ

最近隣グラフによるアルゴリズム

- 最近隣グラフは非循環で，距離が単調に減少する
- 可約性により併合直前のクラスタ以外の近隣グラフは変化しない



最小全域木(MST)と単リンク法

[Murtagh 83]

最小全域木 (MST; minimum spanning tree)
全てのノードを連結するグラフで，辺の重みが最小のもの

- 単リンク法の各段階で併合される最も近いクラスタ対は，必ず最小全域木上の辺で連結されている
- ユークリッド距離で低次元ならば $O(n^2)$ より小さな計算量で最小全域木は生成可能なので，下記のアルゴリズムは $\Omega(n \log n)$ $O(n^2)$

1. 最小全域木を生成
2. n - 1 回反復
    1. 木の中の辺で，距離が最短のものを見つける
    2. 辺の両端のノードに対応するクラスタを併合

分割最適化型クラスタリング

partitional optimization clustering

分割最適化クラスタリング

分割最適化クラスタリング
partitional optimization clustering
partitional clustering, non-hierarchical clustering
クラスタの良さを定義する目的関数を最適にする分割を探索する方法

第2種のスターリング数
n 個のデータを k 個のクラスタに分割する場合の数

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\left(\begin{array}{c} k \\ i \end{array} \right) i^n$$

データ数 n に対して解の候補は指数的に増加するので
一般に大域最適解は計算できない

↓

代わりに局所最適解を求めて近似する

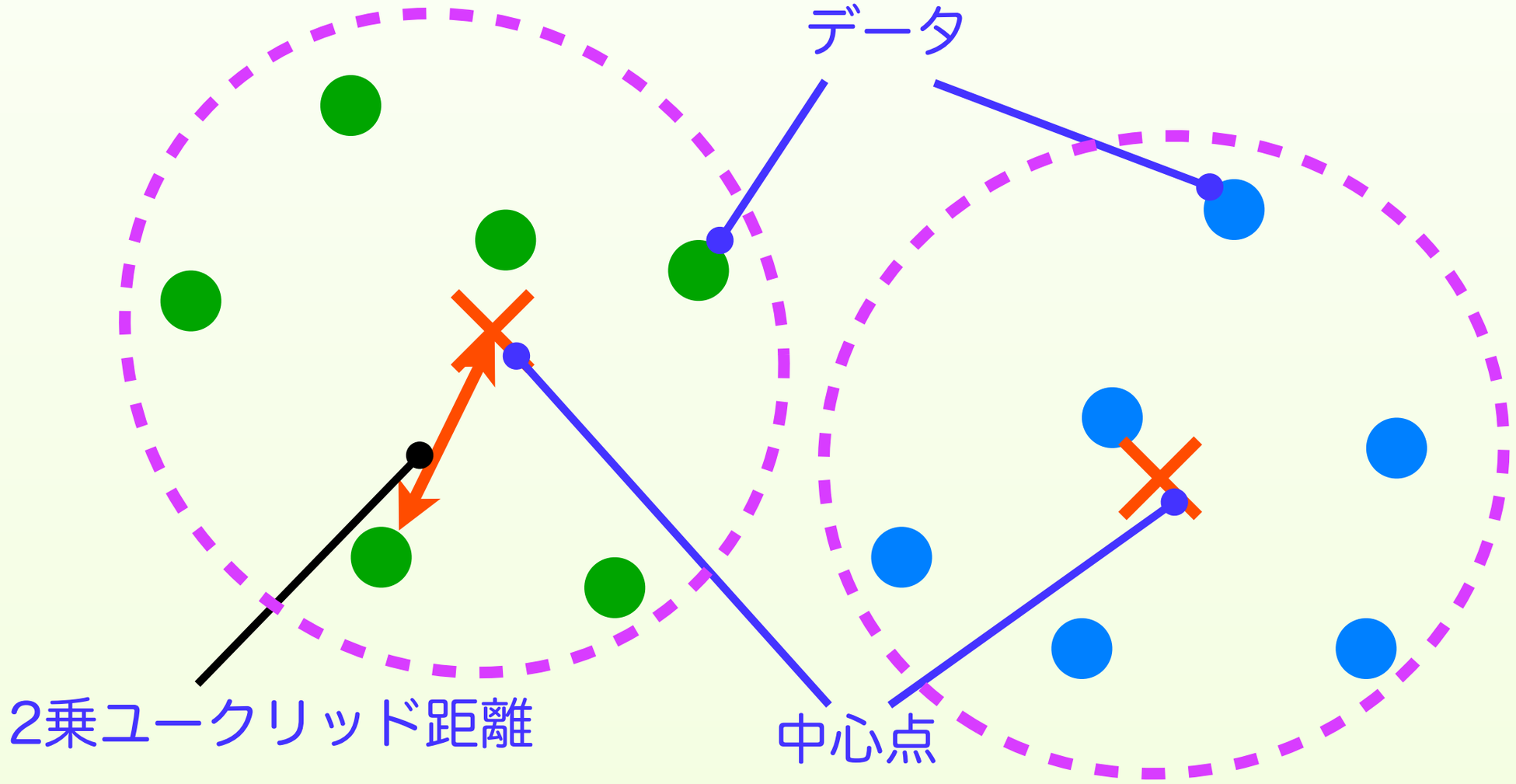
※ 目的関数に劣モジュラ性があり2分割する場合など例外的に多項式時間で計算可

k-means法

[Forgy 65, MacQueen 67]

k-means法　代表的な分割最適化手法

中心点（セントロイド; centroid）とクラスタ内のデータの間の2乗ユークリッド距離の総和が最小になるようにする



k-means法

目的関数

クラスタ内の点についての総和　　クラスタの中心点

$$f(\{C_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\bar{\mathbf{x}}_k - \mathbf{x}_i)^2$$

全クラスタについての和　　中心点とデータ間の2乗距離

- 球状のクラスタが得られやすい
- 同じ大きさのクラスタが得られやすい
- 初期クラスタを変えると，クラスタリング結果も変わる

k-means法

適当な初期クラスタに，二つのステップを反復的に，収束するまで適用

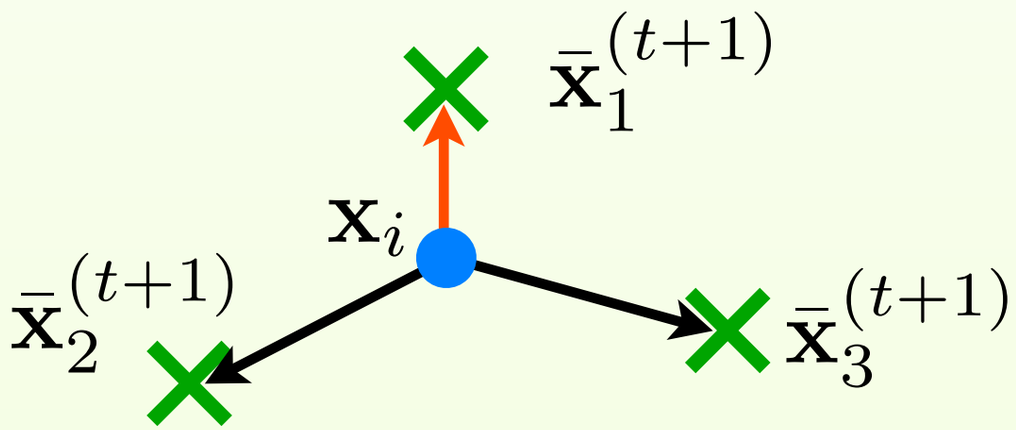
① 中心の更新

現在のクラスタに割り当てられている対象の中心を計算

$$\bar{\mathbf{x}}_k^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x}_i \in C_k^{(t)}} \mathbf{x_i}$$

② 対象の再割当て

全ての点について，最も近い更新後の中心をもつクラスタに再割当てして新クラスタ$C_k^{(t+1)}$を得る



- クラスタへの割り当てに変化がなくなればアルゴリズムは終了
- 各反復で目的関数 $f(\{C_k\})$ は単調減少する

k-means法

初期クラスタの選択方法

k-means法の結果は，局所最適解なので初期クラスタに依存する

ランダムに選ぶ

- k 個の点をランダムに選び k-means法 を適用することを何度か適用し，目的関数が最小になる結果を選択する
- 実装が容易だが，理論的に良いクラスタが得られる可能性は低い

互いに離れた点を選ぶ

- 任意の点を初期中心点として最初に選び，その後はすでに選ばれた点からの平均距離が最も遠い点を次の初期中心点として選ぶ．これを k 回繰り返す
- 理論的にも良いクラスタが得られる可能性は高い
- 単純な実装をすると計算量は $O(N^2)$ となり，k-means法自体よりも大きくなってしまうため，サンプリングなどで近似する必要がある

EMアルゴリズムによる方法

[Dempster+ 77]

混合分布を導入し，EMアルゴリズムで最尤推定でパラメータを求める

混合分布

データ全体の分布　　混合比

$$f(\mathbf{x}_i; \{\alpha_k\}, \{\theta_k\}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{x}_i; \theta_i)$$

全てのクラスタについての和　　要素分布

- 要素分布をいろいろ変えることで，いろいろなクラスタを得ることができる
- 要素分布には正規分布，多項分布，隠れマルコフモデルなどが利用される
- データは $\alpha_k f_k(\mathbf{x}_i; \theta_i)$ に比例する割合で，各クラスタに分類される

k-means法 ：EMアルゴリズムに次の制限を加えたものと見なせる

- 要素分布は，分散が全クラスタ・全属性で同じ σ の正規分布
- 混合比は 0 または 1 に制限

クラスタリングの利用

クラスタの解釈

- ニューヨーク・タイムス紙 1990 年 8 月の約 5,000 件の記事を分類した結果 [Cutting 92]

教育, 国内, イラク, 芸術, スポーツ, 石油, ドイツ統合, 裁判

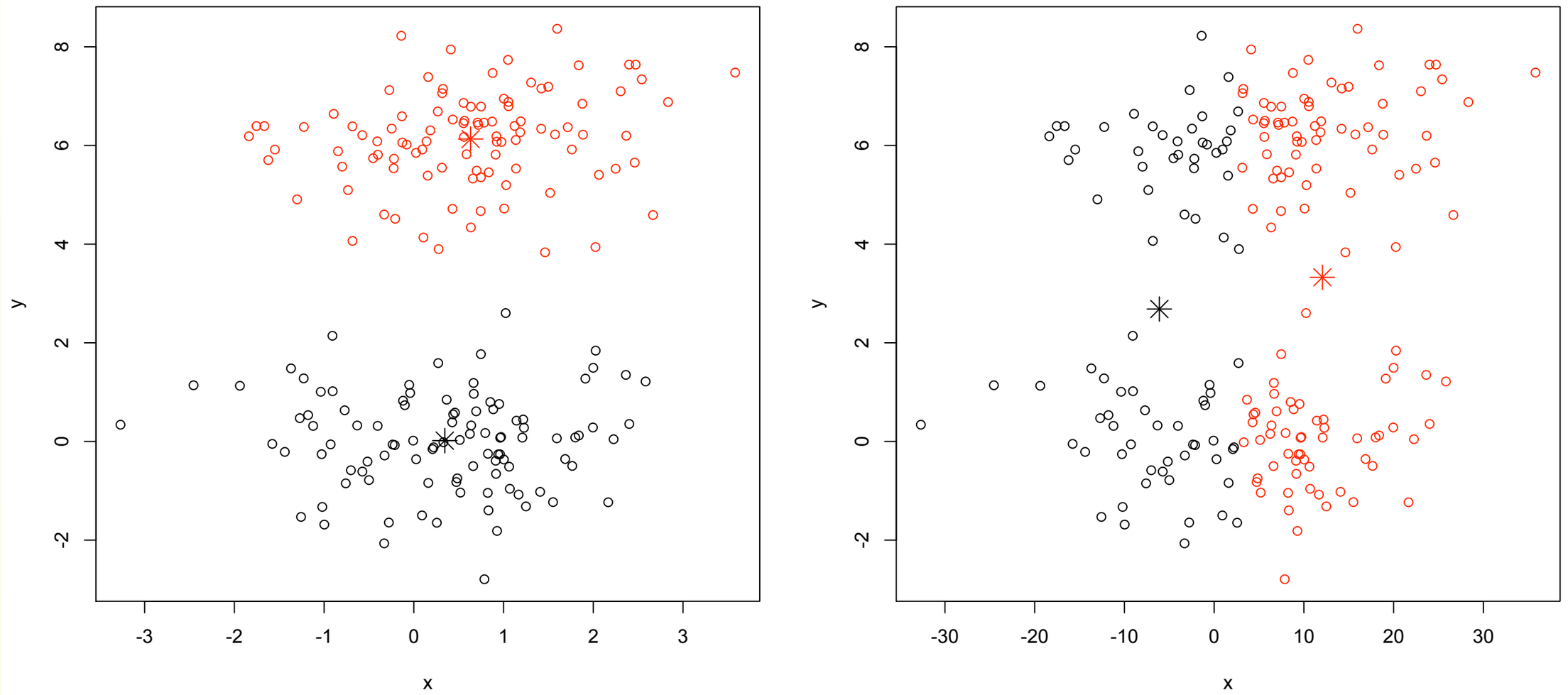
- これは，確かにこの文書集合の「正しい」分類
- 湾岸戦争関連のイラクと石油をまとめても，やはり「正しい」分類
- どちらにも，その分類を正当化する視点が存在する

- 分類結果は絶対的でも，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えない
- 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

クラスタリングは探索的(exploratory)な解析手法で，分類結果はある主観や視点に基づいている．よって，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

クラスタリング結果への影響

データの正規化の影響

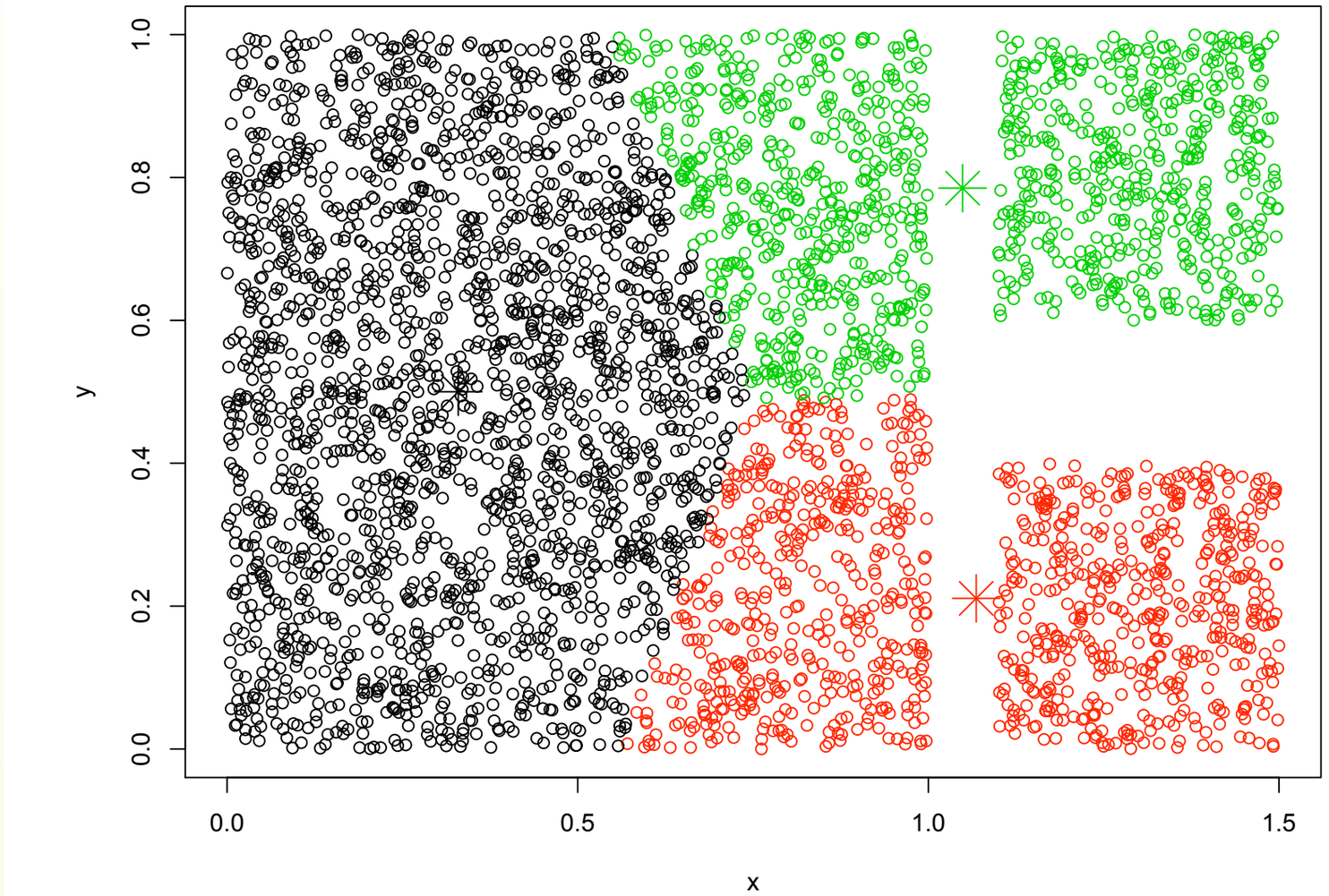


x 軸を10倍したら k-means法の結果が変わってしまった！

※各軸ごとにサンプルの標準偏差で割っておく対処法も

クラスタリング結果への影響

クラスタリングの性質の影響



k-means法は同じ大きさのクラスタに分割する性質がある

※クラスタの大きさが異なるものも抽出できる方法を使う

クラスタリングの正当性の検証

[Dubes+ 79]

クラスタリングの利用目的によって「正当な」クラスタは変わる

データの概要を把握するために探索的に用いる場合

- 究極的には分析者が納得すればいい
- 元データと矛盾していない結果や，均一なデータを無理矢理分割していないといった，入力データの性質を反映しているかを評価
  - ➡ 結果自体の良さを検証する内的妥当性尺度 (internal validity index)

何か分類したいグループがあって，それを自動的に抽出したい場合

- こういうデータ集合なら，このように分割されて欲しいといった分類結果の事例を用意できる
- クラスタリング手法の選択やパラメータの調整によって，分類結果の事例に近い結果を導くものを探す
  - ➡ 分割例との近さを測る外的妥当性尺度 (external validity index)

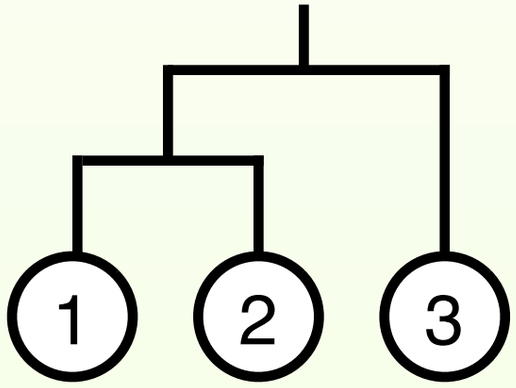
内的妥当性尺度

[Farris 69]

入力データの性質が，分類結果に反映されているかを検証

cophentic相関係数 (CPCC)

元のデータ間の距離と，デンドログラム上のパス長で測った距離の，全データ対にわたるPearson相関係数．この値が小さいなら，デンドログラムには元データの構造が反映されていない．



デンドログラム中のデータ間距離を，パス長で測る

例：1-2間の距離は2，1-3間の距離は3

均一なデータを，無理に分割した結果になっていないかを検証

- 獲得されたクラスタ間で，入力された特徴に差があるかを検定
  - 二つのクラスタ間で，カテゴリ値の頻度に差があるかを χ 二乗検定
  - 二つのクラスタ間で，実数値の平均に差があるかを t 検定

外的妥当性尺度

得られた分類結果と，望ましい分類結果の類似性を評価

- 同じデータ集合 X に対する二つのクラスタリング結果 π と ρ
- データ集合中の全ての対 $x_1, x_2 \in X$ ($M=N(N-1)/2$ 個) について
  - a_{11}：π と ρ のどちらでも同じクラスタにある対の数
  - a_{01}：π では違うクラスタにだが，ρ では同じクラスタにある対の数
  - a_{10}：ρ では違うクラスタにだが，π では同じクラスタにある対の数
  - a_{00}：π と ρ のどちらでも違うクラスタにある対の数

Rand尺度 (Rand index) [Rand 71]

$$\frac{a_{11}+a_{00}}{M}$$

同じクラスタになるかどうかの判定の正解率

正規化相互情報量 (normalized mutual information)

$$\frac{I(\pi,\rho)}{\sqrt{H(\pi)H(\rho)}}$$

H(π) は分割πのどのクラスタに入るかを確率変数としたエントロピー，I(π,ρ)は相互情報量

いったい何を使えばいいのか？

最初は基本的な手法でデータの構造を探る

階層構造が必要？

はい

群平均 / ウォード法

- はずれ値に比較的頑健
- 極端な形状のクラスタは抽出されにくい
- 群平均法は空間保存

いいえ

k-means法

- 計算量が O($N\,k$)
- 極端な形状のクラスタは抽出されにくい

様子がつかめてきたら他の手法でより詳細な解析を

- 線状や超球状のクラスタなら，それぞれ単リンクや完全リンクを利用
- クラスタの大きさに差があったり，形状が楕円だったら，適切な分布を設定したEMアルゴリズムを利用

まとめ

- **クラスタリングとは？**
  与えられたデータ集合を，外的分離と内的結合の性質をもつようなクラスタに分類する方法
- **凝集型階層的クラスタリング**
  データ一つずつがクラスタの状態から始めて，逐次的にクラスタを併合してクラスタの階層構造を獲得する
- **分割最適化クラスタリング**
  クラスタの良さを定義する目的関数を最適にする分割を探索する
- **利用上の注意**
  - 得られたクラスタは絶対的・客観的なものではなく，あくまで一つのある視点から見た結果であることに留意
  - 階層構造が必要なら群平均法かウォード法，不要なら k-means法をまず適用

発展的な手法

半教師ありクラスタリング

本来のクラスタリングは教師なし学習で「正解」はない

↓

適切な分割が想定されていて，その導出のためにクラスタリングを利用する場合もある

↓

適切な分割の例題を使えばいいのでは？

半教師ありクラスタリング

mustリンク：結ばれたデータの対は同じクラスタに分類される

cannotリンク：結ばれたデータの対は違うクラスタに分類される

これらの教示情報を満たすようなクラスタリング結果を求める

COP-k-means

[Wagstaff+ 01]

- mustリンク・cannotリンクの概念を提唱
- k-means法で，これらの制約を満たすようにする
  - k-means法ではクラスタへの割り当てと，中心の計算を反復
  - クラスタへの割り当てでは，最も近い中心にデータ点を割り当てるが，ここで制約を満たす点の中で一番近いクラスタに割り当て
- cannotリンク制約を満たす分割があるかどうかは，k彩色問題と関連があり，NP完全であることが知られる
  - 欲張り探索である COP-k-means では，クラスタへの割り当て順によっては，解が存在しても，その発見に失敗することがある
- mustリンクには推移性がある，すなわち，mustリンクで繋がった一連のデータ点は同じクラスタに分類される

大規模データの処理

データマイニングでは大規模データの処理が目標

➡ メモリや計算速度の限界

問題の特徴を使った効率的なアルゴリズム

- **CLIQUE**: ルールの単調性を用いた

データを小さくまとめる

- サンプリングする (結果の分散が大きくなる)
- データを小さくまとめる技術 (data squashing [DuMouchel 99]) を利用する
  - **BIRCH**: CF-tree と呼ぶ要約表現

BIRCH

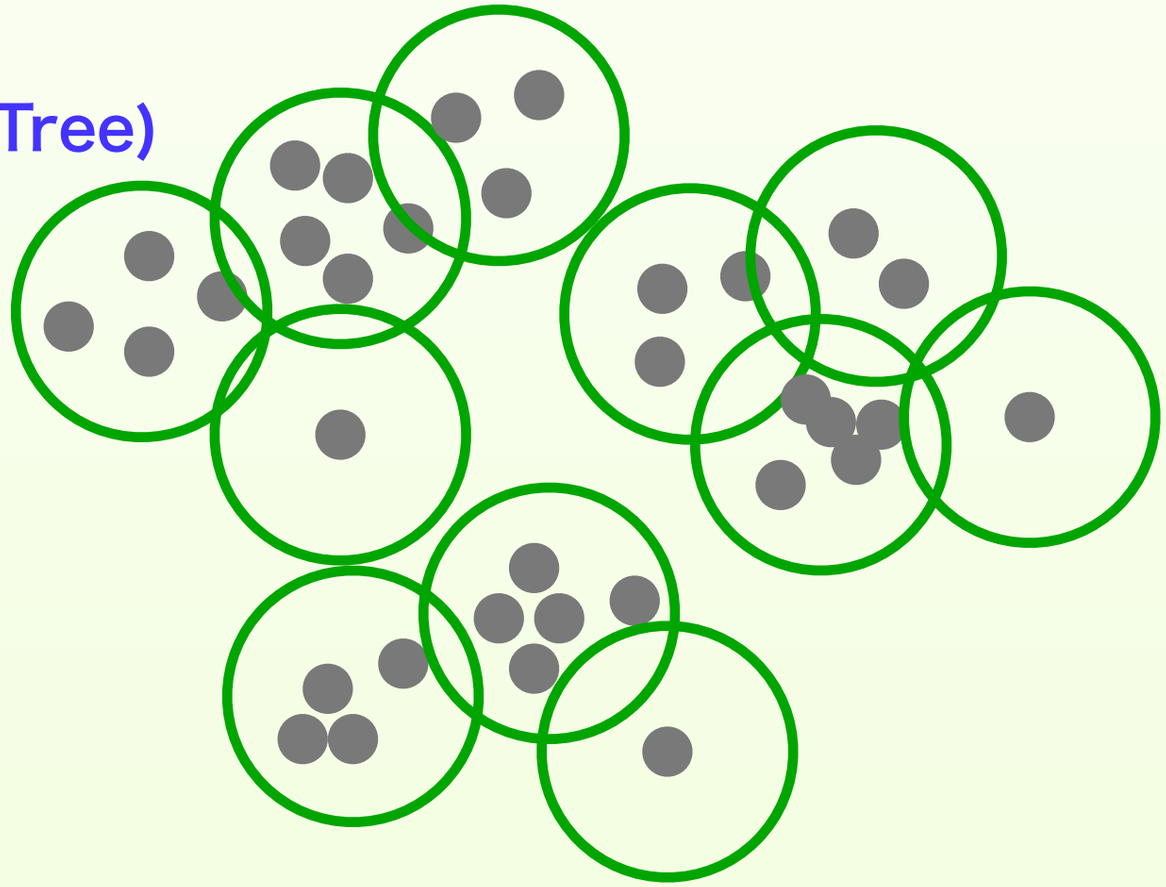
[Zhang 97, Zhang 97]

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

- メモリの利用効率を明示的に考慮したクラスタリング手法
- データ数 N に対し，計算量は O(N) に抑制
- CF-tree というデータの要約表現を利用する

CF-Tree (Clustering Feature-Tree)

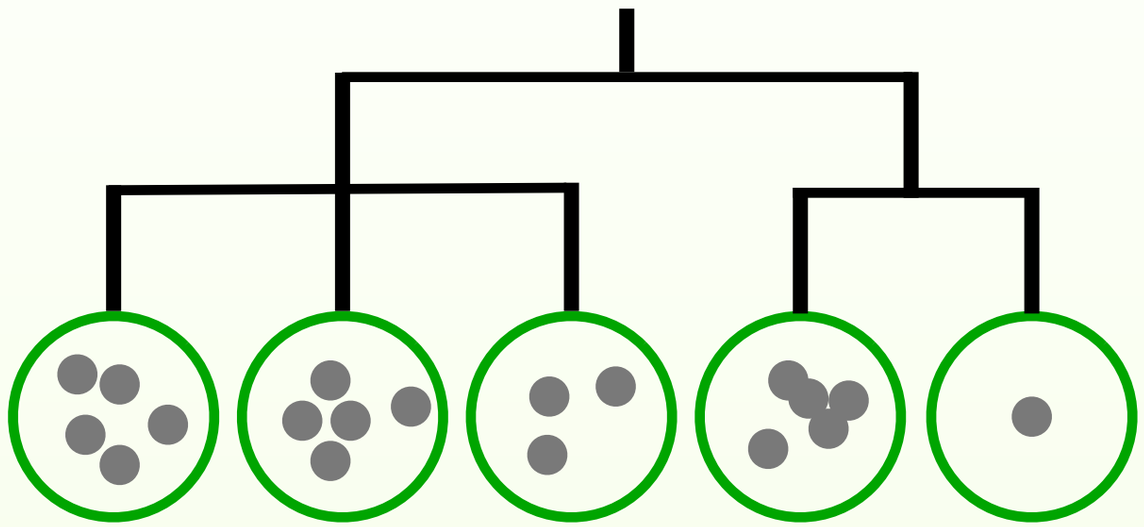
- クラスタリングする前に，直径が一定内の超球中にあるデータ点はひとまとめにする
- まとめたデータは必ず同じクラスタに分類される



BIRCH (CF-Tree)

CF-Tree

- **葉ノード**：直径が一定以内の超球中にあるデータ点の集合



B+-tree：CF-Tree で採用しているデータの格納手法

- データに近い葉ノードを高速に検索可能
- 新データは，近い葉ノードがあればそこへ，なければ新しい葉ノードを作る．効率的に更新可能
- データ集合にアクセスするのは1回だけ

CF (Clustering Feature)

- 葉ノードのデータを少量のメモリで保持

BIRCH (Clustering Feature)

CF (Clustering Feature)　葉ノードのデータを少量のメモリで保持

- 個々のデータは保持せず，データの要約統計量だけを保持

C_k：葉ノードに保持されるデータの集合

データ数　　データの2乗和

$$\left(n_k \, , \ \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \, , \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \cdot \mathbf{x}_i \right)$$

データの総和

C_iとC_jの間の距離

$$\sqrt{\frac{1}{n_i n_j} \sum_{\mathbf{x}_s \in C_i} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_s \cdot \mathbf{x}_t}$$

C_iとC_jのCFだけで計算可能

この距離を使って階層的クラスタリングを適用

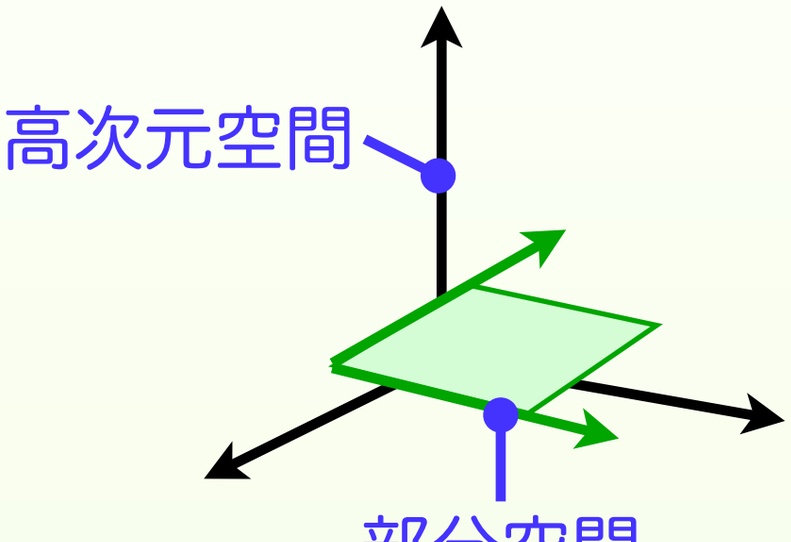
高次元データへの対応

次元の呪い　　高次元ではデータ間の類似性が等しくなる

対処法は……

特徴選択　有用な特徴だけを選択

次元削減　有用な部分空間の発見



前処理としてこれらの手法を適用するのではなく

部分空間クラスタリング
subspace clustering
特徴選択・次元縮約と学習を同時に実行

CLIQUE

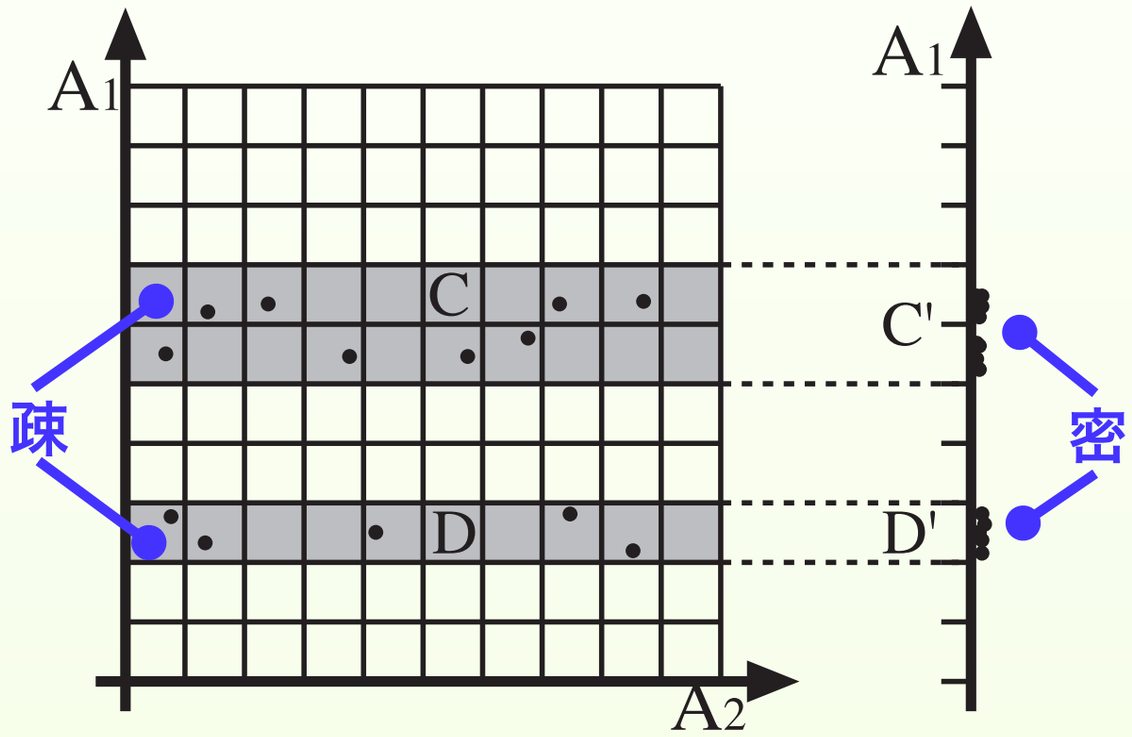
[Agrawal+ 98]

CLIQUE (CLustering In QUEst)

特徴選択を利用するクラスタリング手法

データを格子状に区切って扱う

- 特徴 A_1 と A_2 の両方を考慮すると，データ点は疎
- 特徴 A_1 に射影すれば密



データが密になるような特徴の組み合わせと範囲がクラスタ

Apriori アルゴリズムと同様の単調性を用いた効率的探索

特徴 A_1 と A_2 の両方で密 ← 特徴 A_1 と A_2 のそれぞれで密

(高次元になればなるほど疎になる)

ORCLUS

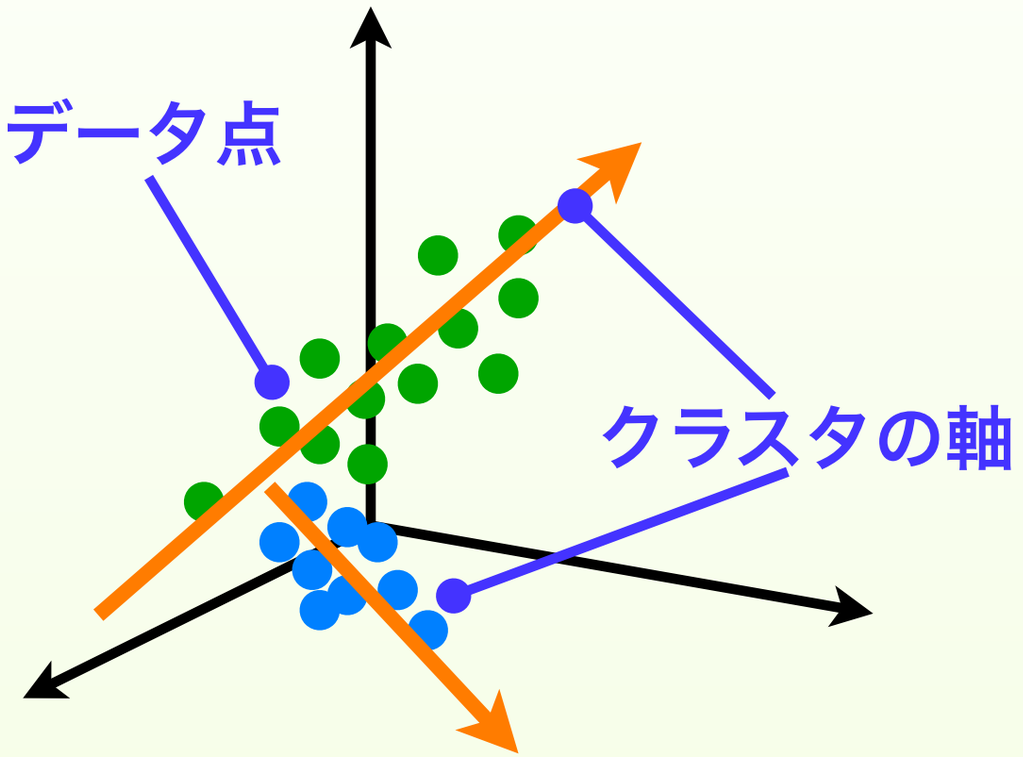
[Aggarwal+ 00]

ORCLUS (arbitrarily ORiented projected CLUSter generation)

次元削減 (特徴変換)を利用するクラスタリング手法

クラスタごとに異なる部分空間

- データ点がその軸に沿って分布するような軸を見つける
- 固有値分解して，小さな固有値に対応する部分空間を採用
  (主成分分析の逆)

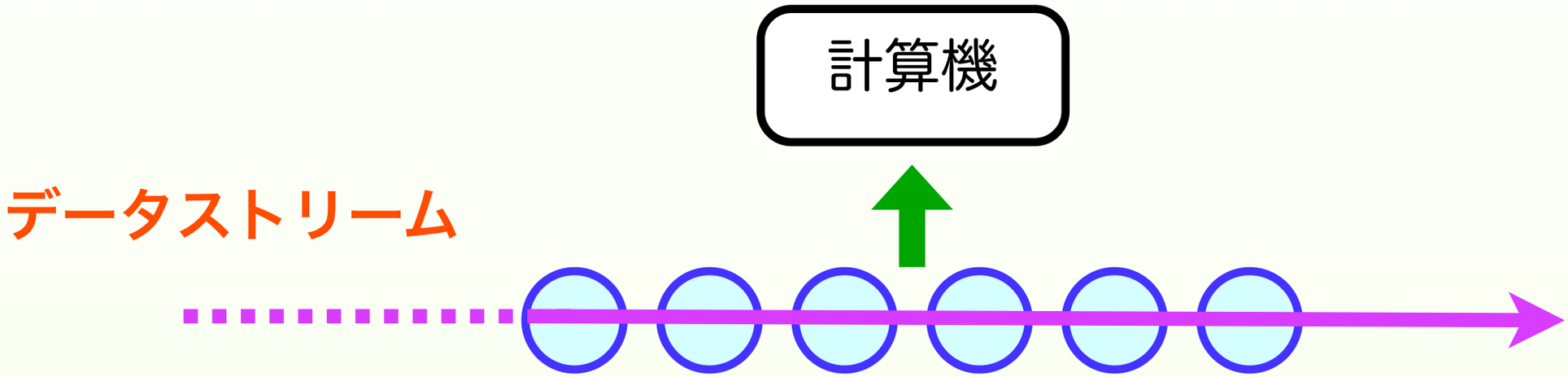


部分空間とその周囲で密なデータ集合がクラスタ

高次元空間から部分空間を，直接見つけるのは困難

低次元部分空間 & 少数データ ➡ 高次元部分空間 & 多数データ

データストリーム



時系列データと似ているが……

- **膨大な量 & 無限に続く** ➡ 全てのデータの蓄積は無理
- **高速** ➡ 次のデータが到着するまでに処理を終える
- **時間的に変化** ➡ 古いデータの影響の除外が必要

例：金融・流通での取引データ，電話・ネットの通信記録，センサーネットワーク

CluStream

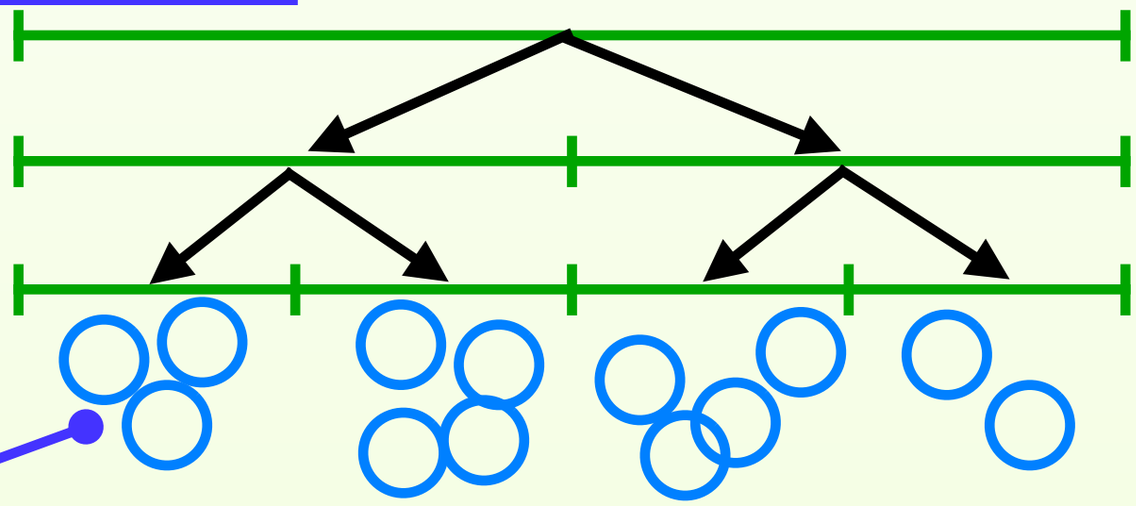
[Aggarwal+ 03]

CluStream

- データストリームのクラスタリング
- データストリームの要約情報の生成と，それを用いたクラスタリングの2段階
- 任意の期間のクラスタリング結果を計算可能

ピラミッド型時間枠　膨大な情報を少量のストレージに保存

再帰的に詳細化される時間枠



マイクロクラスタ：各時間枠内のデータをクラスタリングした結果

BIRCHのCFを用いて表現すると，非常に少量のデータで保持可能

参考文献

Bibliography I


C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu.
A framework for clustering evolving data streams.
In *Proc. of the 29th Very Large Database Conf.*, pp. 81–92, 2003.


C. C. Aggarwal and P. S. Yu.
Finding generalized projected clusters in high dimensional spaces.
In *Proc of The ACM SIGMOD Int'l Conf. on Management of Data*, pp. 70–81, 2000.


R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan.
Automatic subspace clustering of high dimensional data for data mining application.
In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pp. 94–105, 1998.


V. Batagelj.
Note on ultrametric hierarchical clustering algorithms.
Psychometrika, Vol. 46, pp. 351–352, 1981.


D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey.
Scatter/gather: A cluster-based approach to browsing large document collections.
In *Proc. of the 15th Annual ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 318–329, 1992.


A. P. Dempster, N. M. Laird, and D. B. Rubin.
Maximum likelihood from incomplete data via the em algorithm.
Journal of The Royal Statistical Society (B), Vol. 39, No. 1, pp. 1–38, 1977.

Bibliography II


R. Dubes and A. K. Jain.
Validity studies in clustering methodologies.
Pattern Recognition, Vol. 11, pp. 235–254, 1979.


J. L. DuBien and W. D. Warde.
A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms.
The Canadian Journal of Statistics, Vol. 7, No. 1, pp. 29–38, 1979.


W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon.
Squashing flat files flatter.
In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 6–15, 1999.


J. S. Farris.
On the cophenetic correlation coefficient.
Systematic Zoology, Vol. 18, pp. 279–285, 1969.


E. W. Forgy.
Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications.
Biometrics, Vol. 21, pp. 768–780, 1965.


A. K. Jain, M. N. Murty, and P. J. Flynn.
Data clustering: A review.
ACM Computing Surveys, Vol. 31, No. 3, 1999.

Bibliography III

神嶌敏弘.
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！ —.
人工知能学会誌, Vol. 18, No. 1, pp. 59–65, 2003.

神嶌敏弘.
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 —.
人工知能学会誌, Vol. 18, No. 2, pp. 170–176, 2003.

G. N. Lance and W. T. Williams.
A general theory of classificatory sorting strategies.
The Computer Journal, Vol. 9, pp. 373–380, 1967.

F. Murtagh.
A survey of recent advances in hierarchical clustering algorithms.
The Computer Journal, Vol. 26, No. 4, 1983.

W. M. Rand.
Objective criteria for the evaluation of clustering methods.
Journal of The American Statistical Association, Vol. 66, pp. 846–850, 1971.

齋藤堯幸, 宿久洋.
関連性データの解析法 — 多次元尺度構成法とクラスター分析法.
共立出版, 2006.

Bibliography IV


K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl.
Constrained k-means clustering with background knowledge.
In *Proc. of the 18th Int'l Conf. on Machine Learning*, pp. 577–584, 2001.


R. Xu and D. Wunsch.
Clustering.
Wiley-IEEE Press, 2008.


T. Zhang, R. Ramakrishnan, and M. Livny.
Birch: An efficient data clustering method for very large databases.
In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pp. 103–114, 1996.


T. Zhang, R. Ramakrishnan, and M. Livny.
Birch: A new data clustering algorithm and its applications.
Data Mining and Knowledge Discovery, Vol. 1, pp. 141–182, 1997.