発表資料

金子邦彦

2004. 8

Outline

Clustering に関する先行研究の紹介

- farthest-point clustering (1985)
  - 精度 : 近似比2を保証.　O($n\log k$) のアルゴリズム
- クラスタリングの精度は上限がある(1988)
  - 精度 :　最悪の場合の近似比→2は保証できる.
    2を大きく超えて改善することは困難
- CLARANS　(1994)
  - 精度 :　近似比2は保証できない
  - 高速
- その他

クラスタリング問題

入力

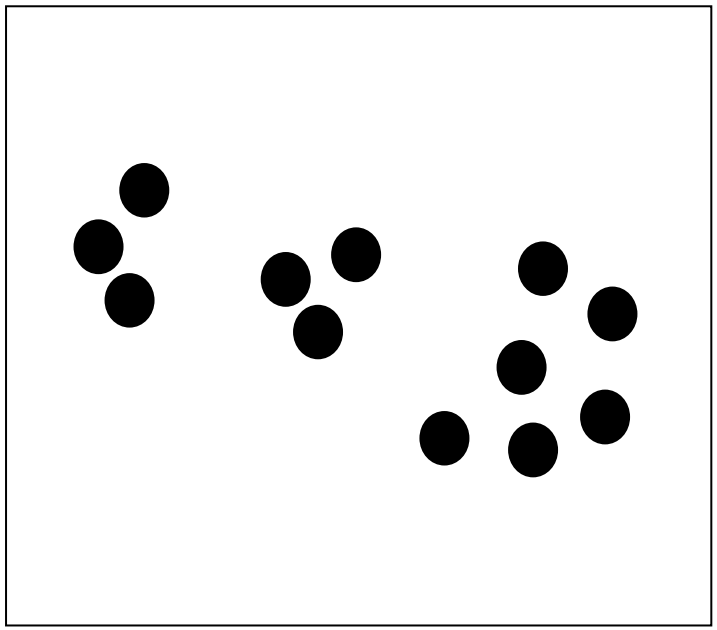
点集合 S

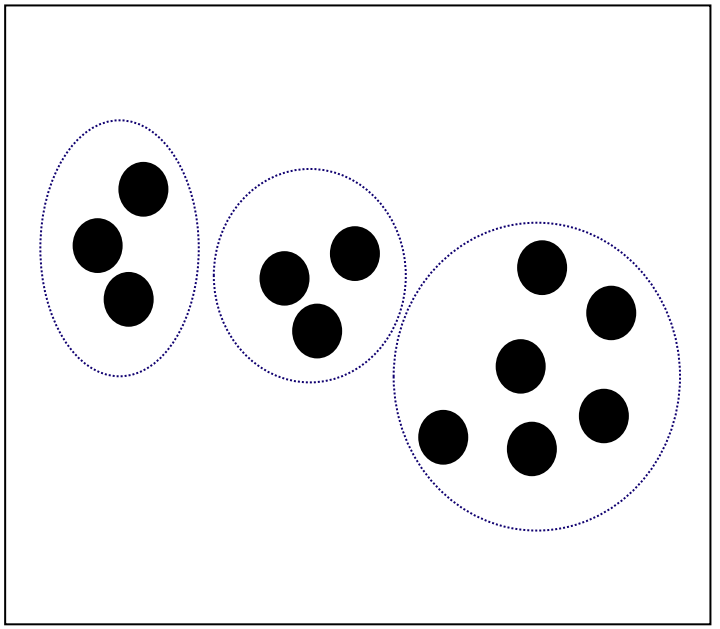
整数 k

出力

S の「良い」分割

$S_1,\ldots,S_k$

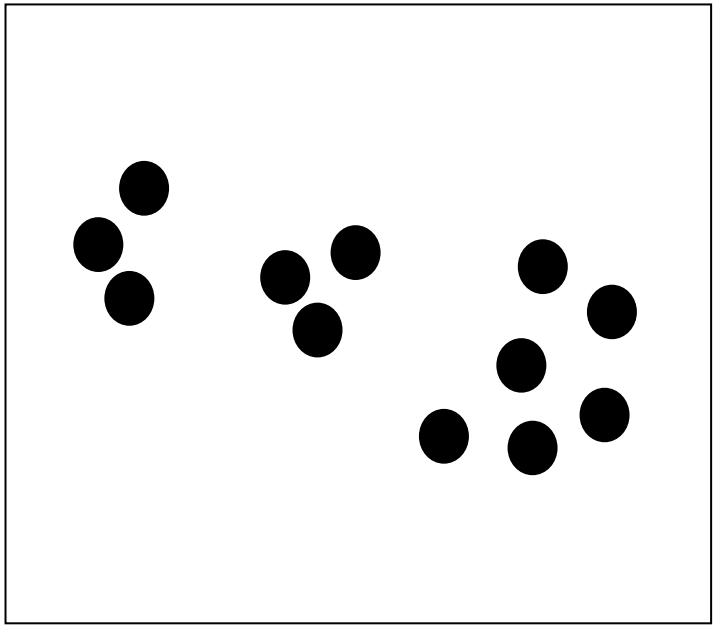




最遠点アルゴリズム
farthest-point clustering

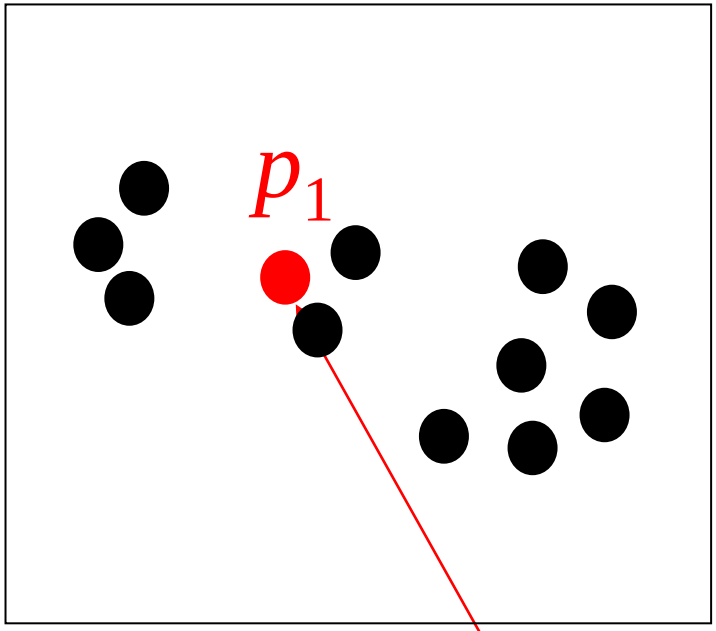
[1] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. Journal of Theoretical Computer Science, No. 38, pp.293-306, 1985.

farthest-point clustering



- 入力
  - 点集合 S
  - 整数 $k = 3$

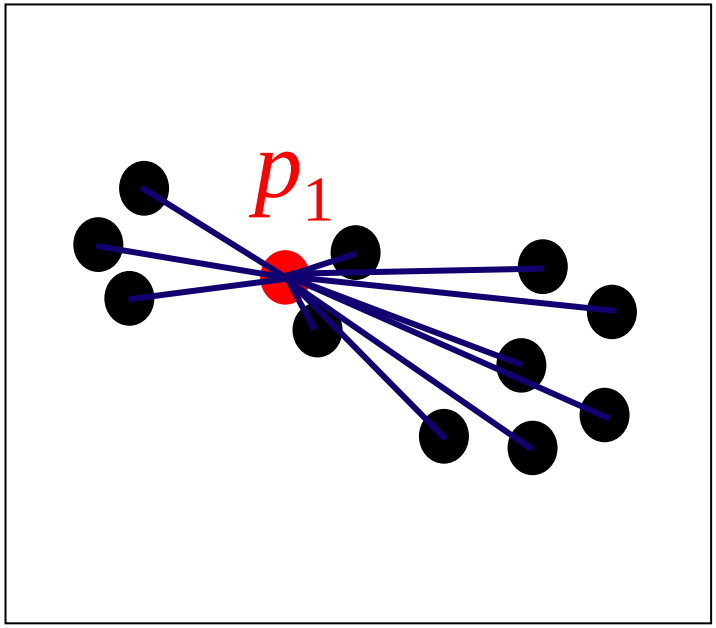
farthest-point clustering



1点 p_1 を，S から選ぶ

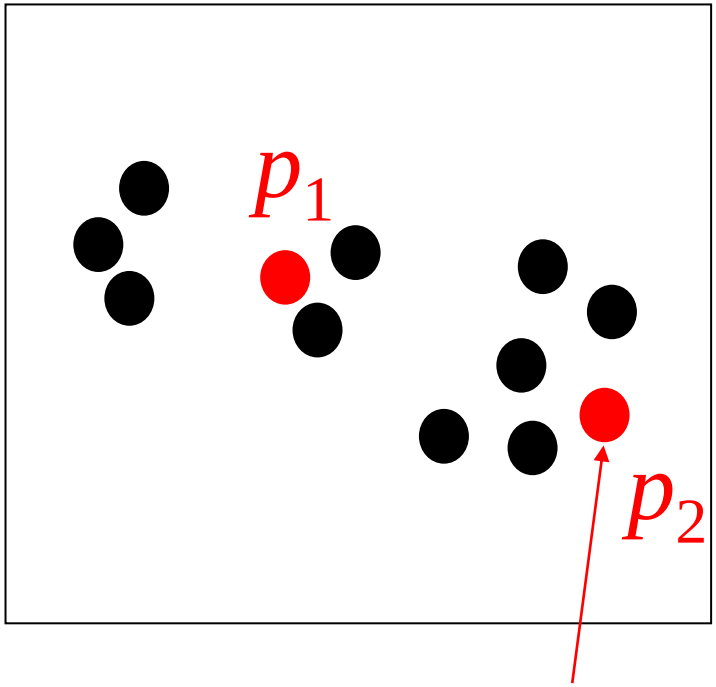
（選び方は後述）

farthest-point clustering



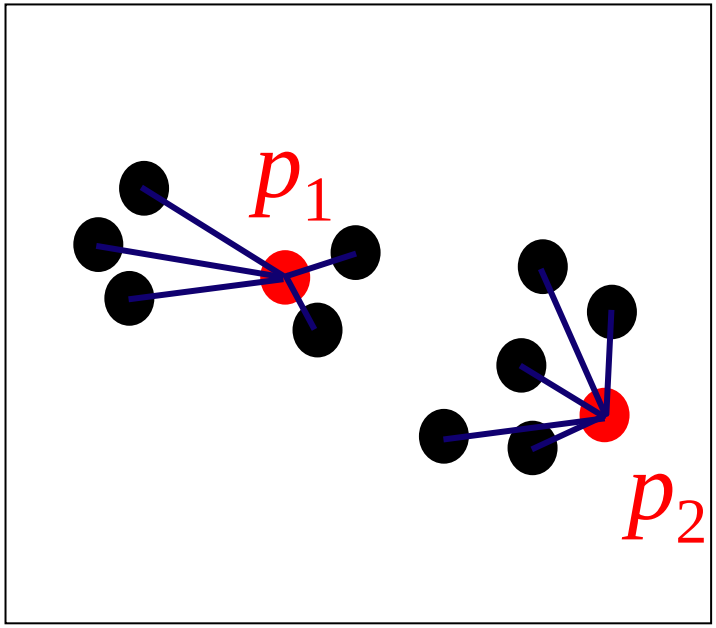
各点とPとの距離を求める: O(n)
(nは点の数)

farthest-point clustering



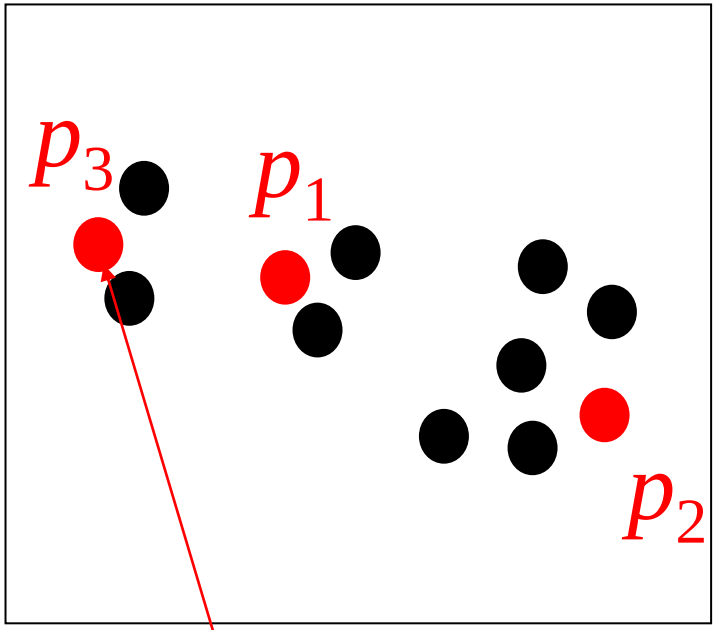
P (=$\{p_1\}$)からの距離が最も遠い点 p_2 を，
S から選び，P に加える

farthest-point clustering



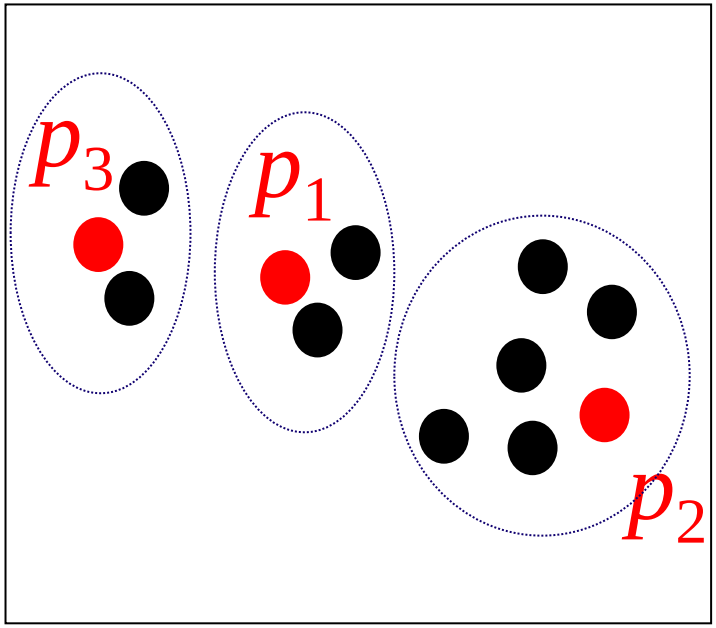
各点と P との距離を求める: O(n)
(n は点の数)

farthest-point clustering



P (=$\{p_1, p_2\}$)からの距離が最も遠い点 p_3 を，
S から選び，P に加える

farthest-point clustering



３点　$P = \{p_1, p_2, p_3\}$を代表点とする
クラスタを作る（これで終了）

farthest-point clustering

1. 任意の1点 p_1 を，S から選ぶ

   $P \leftarrow \{p_1\}$ ・・・ p_1 は S_1 の代表点

2. P からの距離が最も遠い点 p_i を，S から選び，P に加える

   $P \leftarrow P \cup \{p_i\}$ ・・・ p_i は S_i の代表点

   点 p_i と点集合P の距離：

   p_i から最も「近い」P の点と，p_i の相違度

   $\min\{D(p_1, p_i), D(p_2, p_i),...,D(p_{i-1}, p_i)\}$

3. P のサイズが k に達するまで 1, 2 を繰り返す

4. p_1 から p_k に対応した k 個のクラスタを作る

「最初の１点の選択」に関する3手法

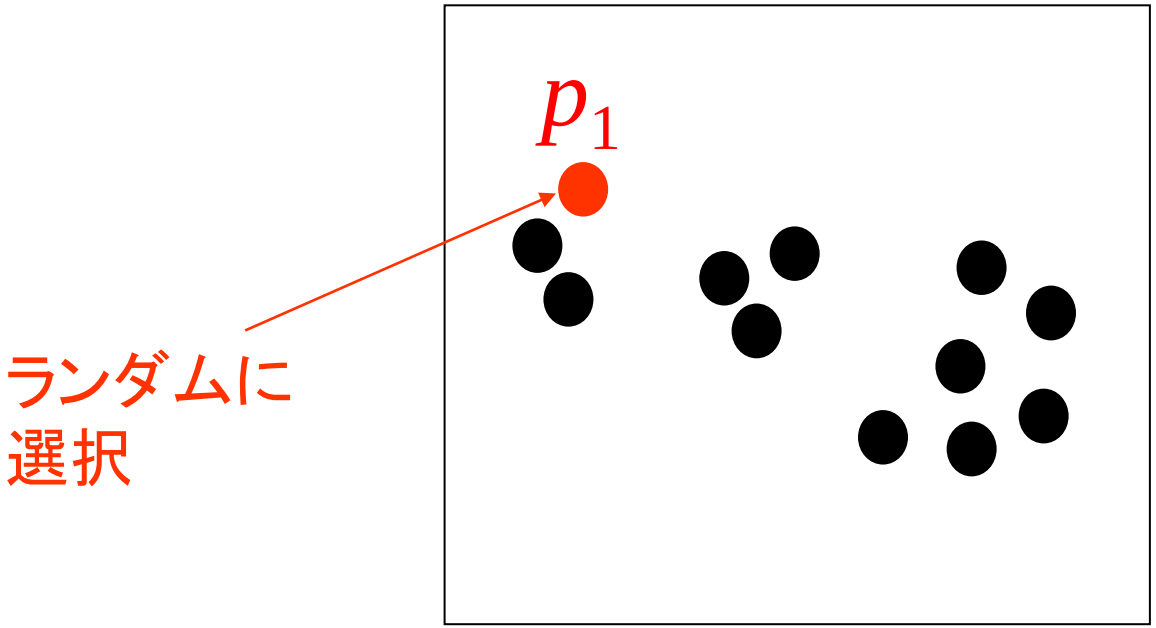
- ランダム
  - 最初の１点は，Ｓからランダムに選択
- 1-center
  - 最初の１点は，「その点から、S内の他の点までの相違度」の最大値が最小になるように選択
- Plus

  S内の全ての点を試す

  - S内の点を1つずつ選び，クラスタリングを行う
  - これを，S内の全ての点について繰り返す
  - 最も「良い」クラスタリング結果を選ぶ

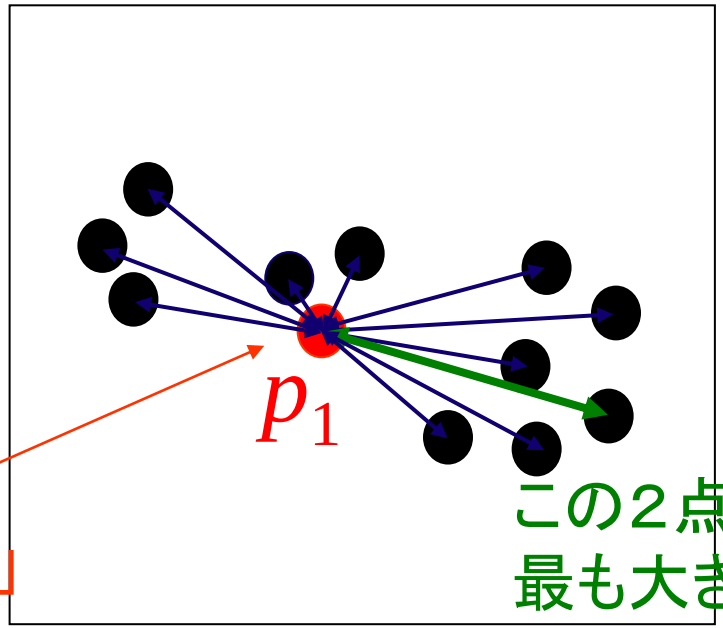
最初の１点の選択

- ランダム
  - 最初の１点は，Ｓからランダムに選択



最初の１点の選択

- 1-center
  - 最初の１点は，「その点から、S内の他の点までの相違度」の最大値が最小になるように選択

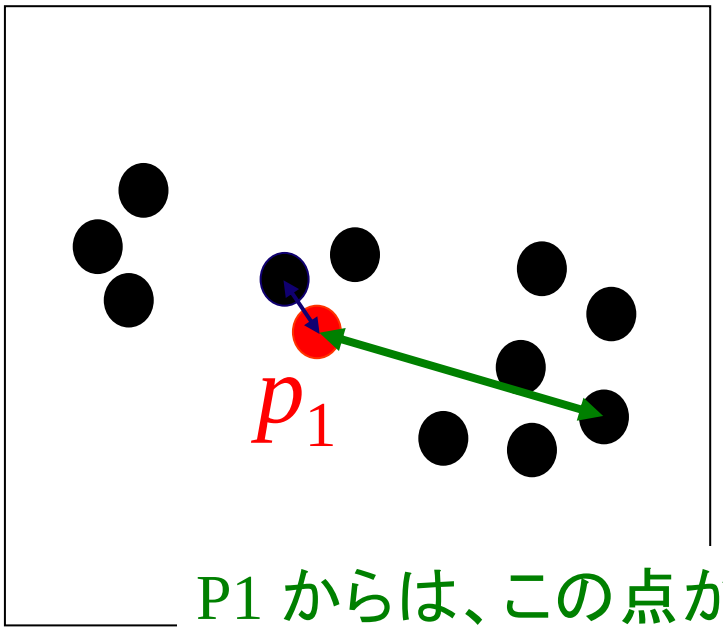


この点を選ぶと
「相違度の最大値」
が最も小さくなる

この２点間が相違度が
最も大きい

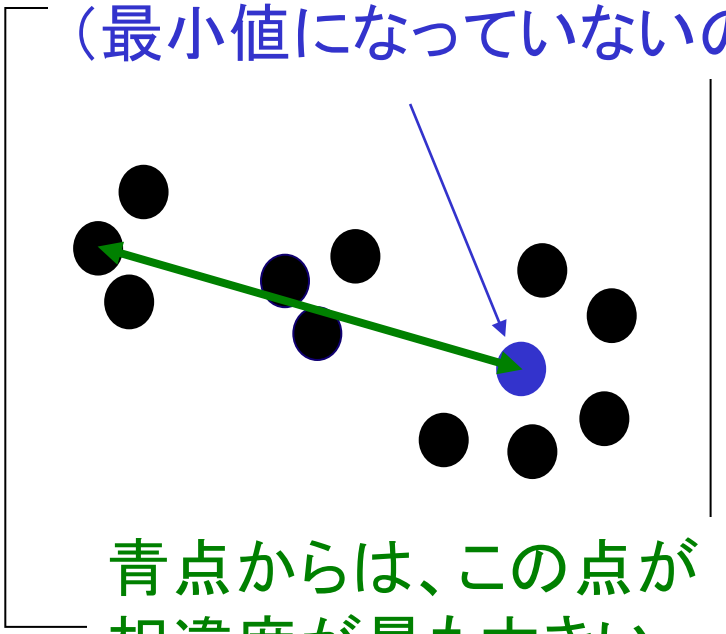
最初の１点の選択

- 1-center
  - 最初の１点は，「その点から、S内の他の点までの相違度」の最大値が最小になるように選択



P1 からは、この点が
相違度が最も大きい

この点を選ぶことは無い
（最小値になっていないので）



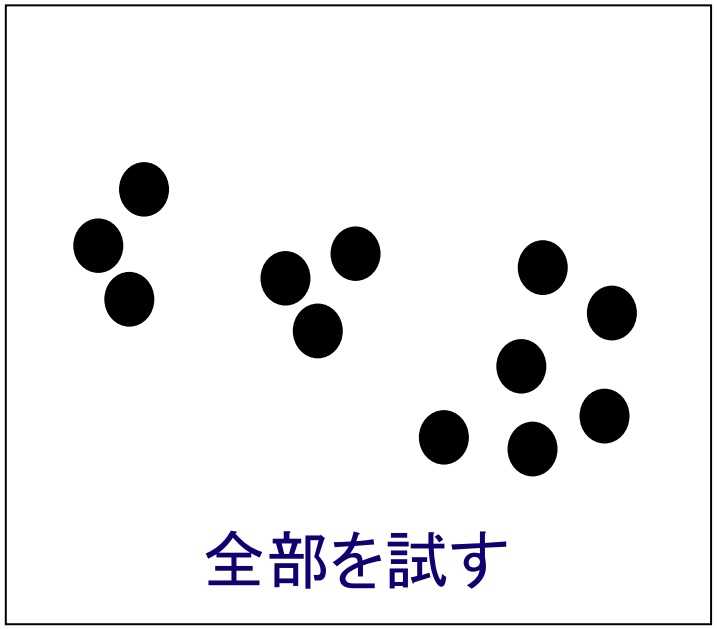
青点からは、この点が
相違度が最も大きい

最初の1点の選択

- Plus

  S内の全ての点を試す

  – S内の点を1つずつ選び，クラスタリングを行う
  – これを，S内の全ての点について繰り返す
  – 最も「良い」クラスタリング結果を選ぶ



アルゴリズムそのもの
の評価のために役立つ

Outline

Clustering に関する先行研究の紹介

- farthest-point clustering (1985)
  - 精度 : 近似比2を保証.　O($n$log$k$) のアルゴリズム
- クラスタリングの精度は上限がある(1988)
  - 精度 :　最悪の場合の近似比→2は保証できる.
    2を大きく超えて改善することは困難
- CLARANS　(1994)
  - 精度 :　近似比2は保証できない
  - 高速
- その他

クラスタリングの尺度
代表的な3種類の紹介

距離空間にマップできない
データもクラスタリング可能

farthest-point Clustering
精度：最悪でも，近似比
(approximation rate) が2

近似比が，精度の尺度

どの尺度を選んでも，最悪のケースでの
近似比(approximation rate) を，2を超え
て改善することは困難

クラスタリング問題

- 入力
  - 点集合 S
    - 点 $p \in S$ の間に相違度(dissimilarity) D が定義されている
  - 整数 k
    - 出力されるクラスタ個数
- 出力
  - S の「良い」分割 $S_1,\ldots,S_k$

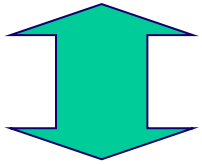
後述

- k=2: 2個のクラスタに区別

  教師あり or 教師なし

  – ベイズ関数
  – サポートベクターマシン　(Support Vector Machine)

    など



- 一般のk (k≧1):　多数のクラスタに区別

  教師あり or　教師なし

一般のk で教師無し　→ 各種のクラスタリングアルゴリズム

クラスタリングの尺度

- Minmax radius clustering:
  - 各クラスタを含む球の半径の最大を最小化

$$mimimize \max_{1 \le i \le k} radius(S_i)$$

- Minmax diameter clustering:
  - 各クラスタの直径（最も遠い2点間の相違度）の最大を最小化

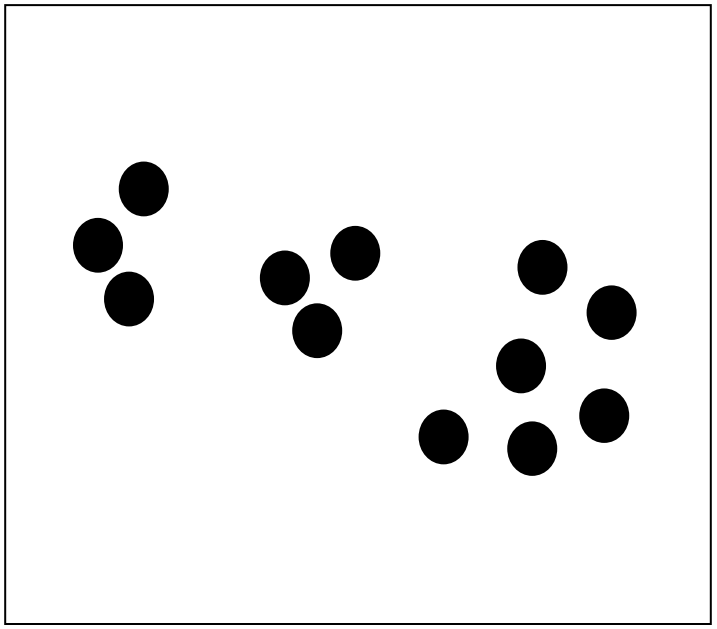
$$mimimize \max_{1 \le i \le k} \max_{v,w} D(p_v, p_w), p_v \in S_i, p_w \in S_i$$

- Variance-based clustering:
  - クラスタ内分散の和を最小化

$$mimimize \sum_{i=1}^{k} \sum_{p_v, p_w \in S_i} D^2(p_v, p_w)$$

Minmax radius clustering

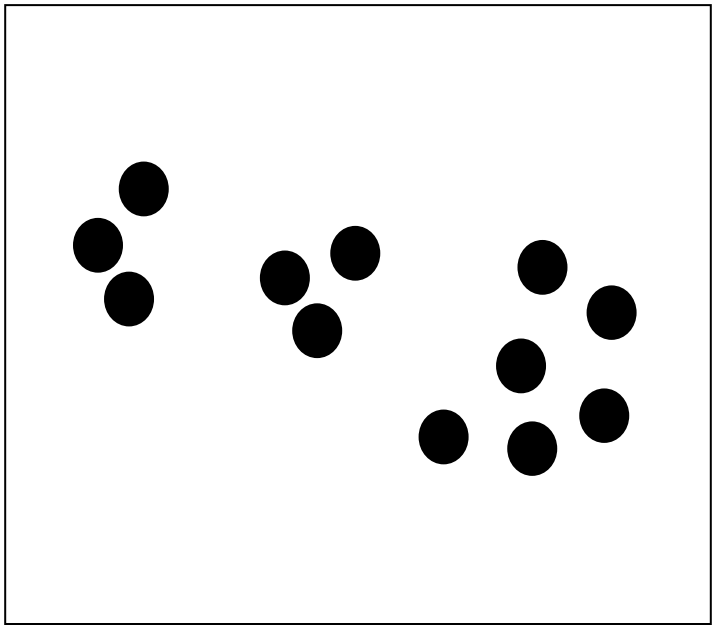
各クラスタを含む球の半径の最大を最小化



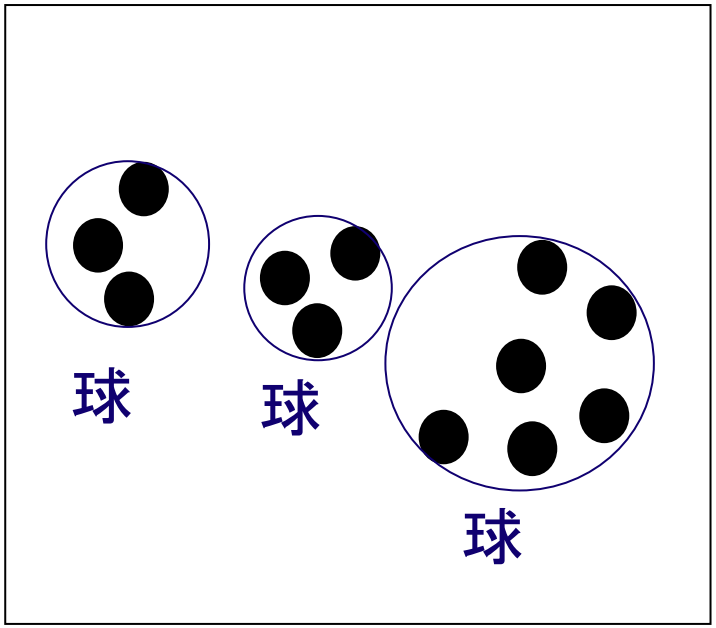
2次元空間の点集合 S

Minmax radius clustering

各クラスタを含む球の半径の最大を最小化



k = 3

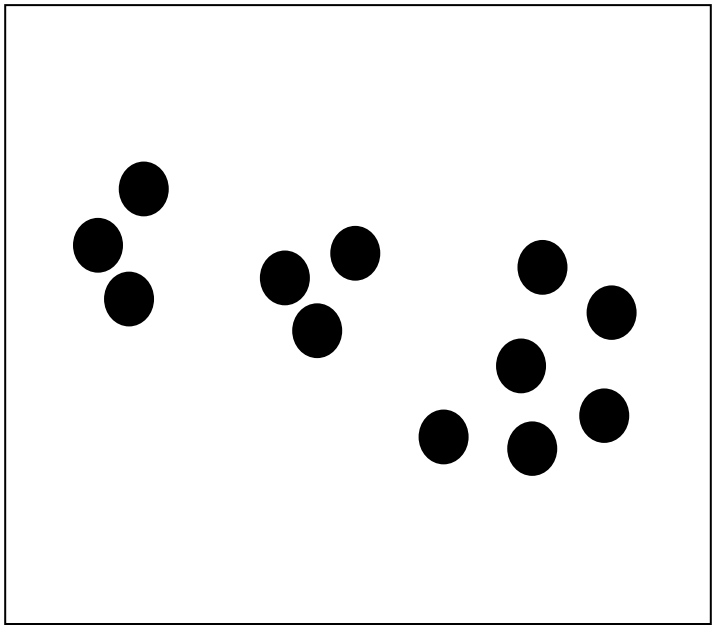


2次元空間の点集合 S

一般の相違度の場合，
点が距離空間内に無いため、
球の中心を定めることが困難
球の中心は，点から選ぶ

Minmax diameter clustering

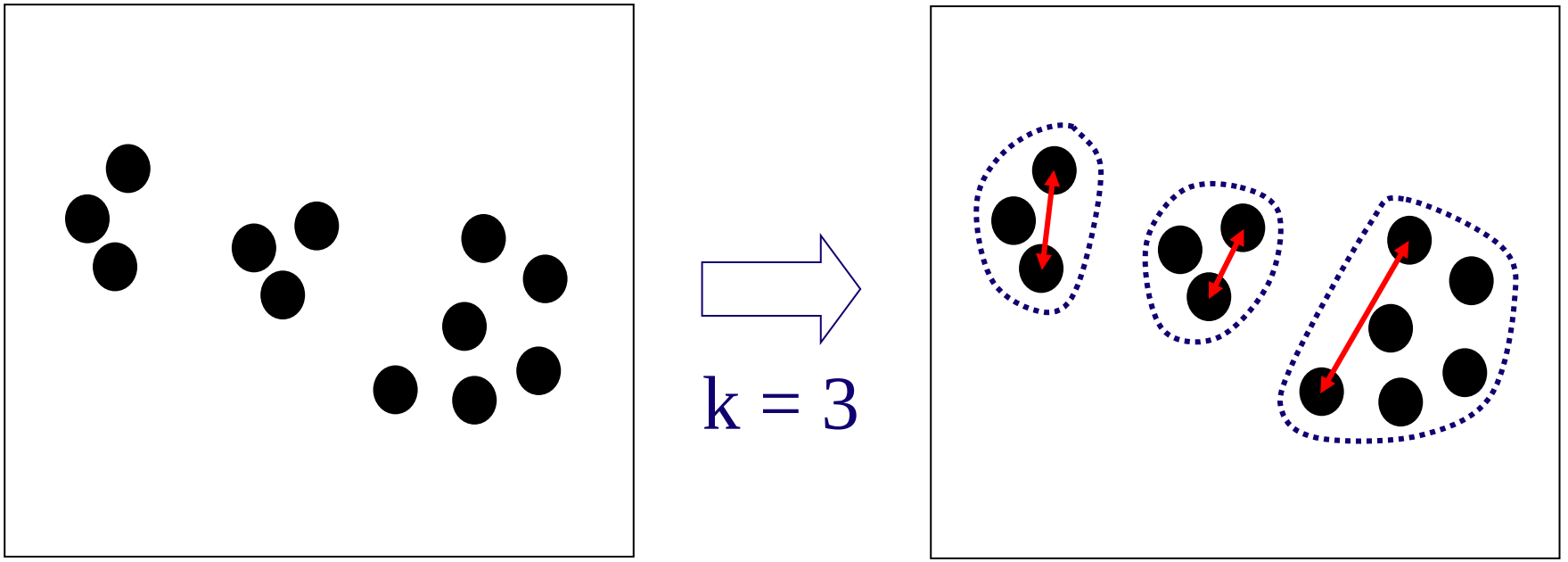
各クラスタの直径（最も遠い2点間の相違度）の最大を最小化



2次元空間の点集合 S

Minmax diameter clustering

各クラスタの直径（最も遠い2点間の相違度）の最大を最小化

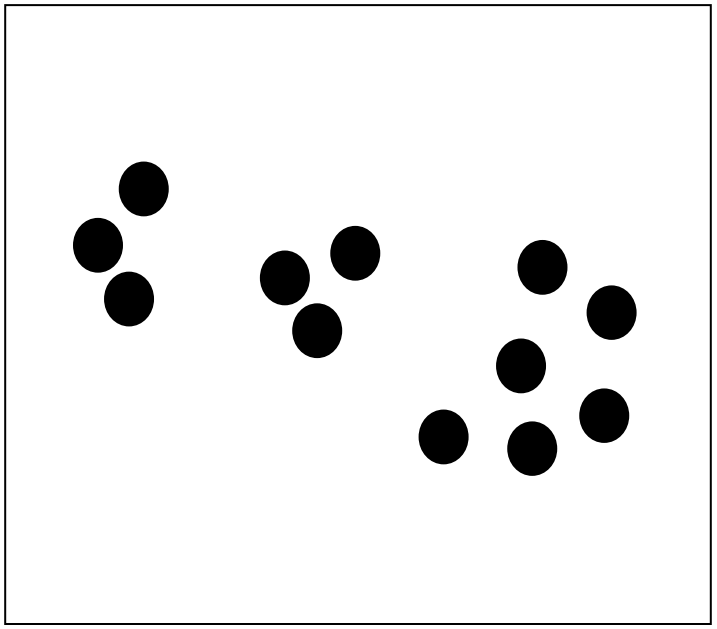


2次元空間の点集合 S

各点間の相違度の最大

Variance-based clustering

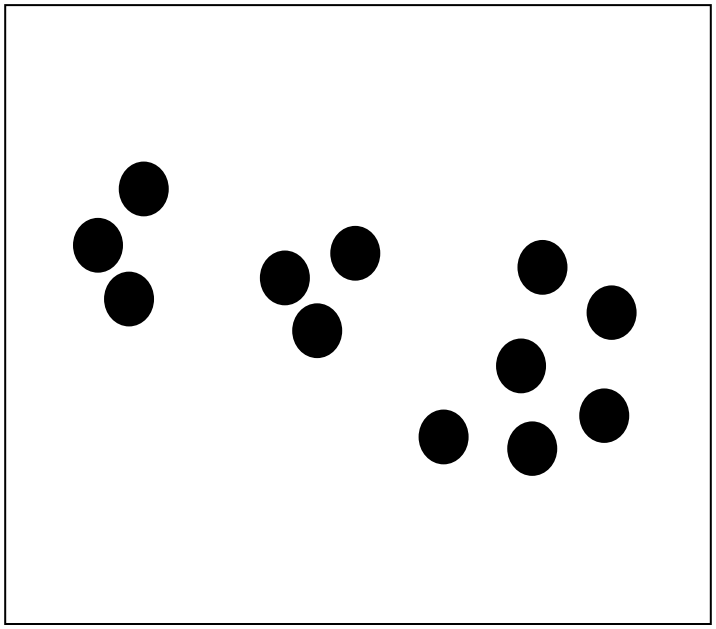
クラスタ内分散の和を最小化



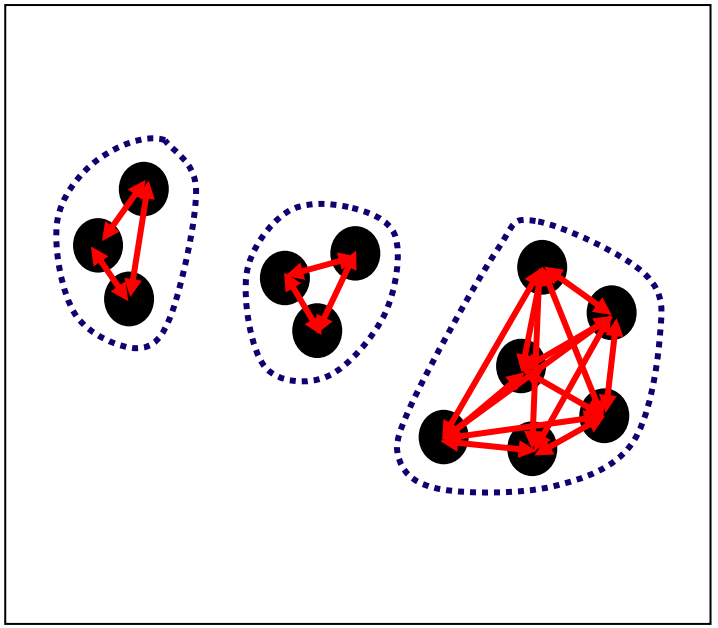
2次元空間の点集合 S

Minmax radius clustering

クラスタ内分散の和を最小化



k = 3



2次元空間の点集合 S

クラスタ内分散の和
（分散＝各点間の
相違度の2乗和）

相違度の種類

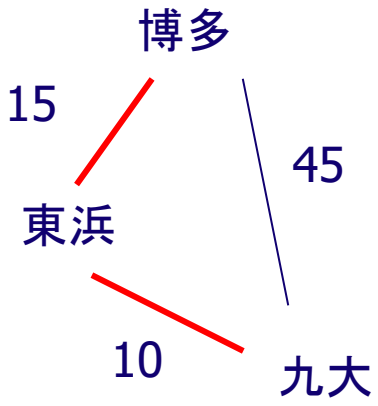
- 距離 (metric)
  - 対称な距離
    - ユークリッド距離 (Euclidean distance)
    - Weighted Euclidean distance
    - マンハッタン距離
    - 編集距離 （editing distance)
    - alignment score　(*) score matrix が三角不等式 $s(a,b)+s(b,c) \geqq s(a,c)$ を満足する場合 BLOSSUM50 等
    - キーワード距離 (keyword distance)
    - ヒストグラム間の各種の距離（EMD距離, マハラノビス距離など）
  - 非対象な距離
    - 相関ルールの確信度の逆数の対数など
- 距離の性質を持たない相違度

  距離の性質　＝　三角不等式 (triangular inequality)
  - 重み付きグラフ （次ページで説明）

重み付きグラフでの相違度

- グラフの 2 node 間の 相違度
  ＝　2 node 間の shortest path
- Dijkstra’s algorithm

道路網の例
（この例では、三角不等式が成立しない）



クラスタリングの尺度

Minmax radius clustering	球の半径	「距離」のみ使える (一般の相違度について使えるように拡張: k-center)	PAM, CLARA, CLARANS 等
Minmax diameter clustering	クラスタの直径（最も遠い２点間の相違度）	一般の相違度について使える	farthest-point Clustering等
Variance-based clustering	クラスタ内分散の和	一般の相違度について使える	BIRCH 等

さきほど説明した

farthest-point Clustering では、

精度、計算量複雑さはどうか？

farthest-point clustering の性質

- 計算量複雑さ:
  - $O(nk)$ [1]
    - k が1つ増えるたびに,「k 番目の代表点 p_k と, 他の点の距離を求める処理」が増える
  - $O(n \log k)$ に改善できる, これ以上改善できない[2]
    - Box Decomposition 法で、クラスタサイズを見積もる

n: 点集合 S の要素数
k: 出力されるクラスタ個数

[2] Feder, T. and Greene, D,
Optimal algorithms for approximate clustering,
In Cole, R. ed., Proc. 20th Annual ACM Symposium on the Theory of Computing, pp. 434-444, 1988

farthest-point clustering の性質

- 精度： 近似比 (approximation rate) は 2

> 得られたクラスタの「各クラスタの直径の最大（最大直径）」は，最適なクラスタリングの最大直径の2倍以下

であることが証明できる

- 実際には、2よりもずっと良い（もっと1に近い数）
- 最悪でも2になることが証明できる

farthest-point clustering は，最適なクラスタリングの2倍以下の最大直径を持つクラスタリングを出力する（approximation rate は 2）

問い

approximation rate を、2よりも改善できるか？

farthest-point clustering は，最適なクラスタリングの2倍以下の最大直径を持つクラスタリングを出力する（approximation rate は 2）

問い

approximation rate を、2よりも改善できるか？

できない[2]

[2] Feder, T. and Greene, D,
Optimal algorithms for approximate clustering,
In Cole, R. ed., Proc. 20th Annual ACM Symposium on the Theory of Computing, pp. 434-444, 1988

近似比(approximation rate) の限界

Minmax radius clustering	球の半径	近似比2は可能	・1.822 以上には改善できない[2] ・k-center 問題では、2以上には改善できない
Minmax diameter clustering	クラスタの直径（最も遠い2点間の相違度）	近似比2は可能	1.969 以上には改善できない[2]
Variance-based clustering	クラスタ内分散の和		「定数」倍以内であること自体が保証できない

Minmax radius clustering, Minmax diameter clustering に関して，k-center 問題を解く2倍未満の approximation rate の解を必ず出力する多項式時間アルゴリズムは存在しない[2].

などの証明

[2] Feder, T. and Greene, D,
Optimal algorithms for approximate clustering,
In Cole, R. ed., Proc. 20th Annual ACM Symposium on the Theory of Computing, pp. 434-444, 1988

Outline

Clustering に関する先行研究の紹介

- farthest-point clustering (1985)
  - 精度 : 近似比2を保証.　O($n\log k$) のアルゴリズム
- クラスタリングの精度は上限がある(1988)
  - 精度 :　最悪の場合の近似比→2は保証できる.
    2を大きく超えて改善することは困難
- CLARANS　(1994)
  - 精度 :　近似比2は保証できない
  - 高速
- その他

問題

- 精度、計算量複雑さの追及

近似比は2を大きく超えて改善できない [2]

- 近似比は2を保証しながら、性能の追及

Hochbaumらの方法[3] などが散見される状況

- 近似比の保証にはこだわらないで、性能の追及

CLARANS[5] 等

[3] Dorit S. Hochbaum and David B. Shmoys,
A best possible heuristic for the k-center problem,
Mathematics of Operations Research,
No. 10, pp. 180-184, 1985.

Clustering Techniques based on partitioning the data

- k-means アルゴリズム
- R-tree, R*-tree
- k-medoid アルゴリズム
  - PAM[5]
  - CLARA
  - CLARANS[6] (Clustering Large Applications based on RANdomized Search)

[5] Kaufman, L. and Rousseeuw, P. J. Finding Groups in Data : An Introduction to Cluster Analysis, John Wiley and Sons, 1990.
[6] Raymond T. Ng and Jiawei Han. Efficient and Effective Clustering Methods for Spatial Data Mining. VLDB 1994, pp. 144-155, 1994

k-Means Clustering

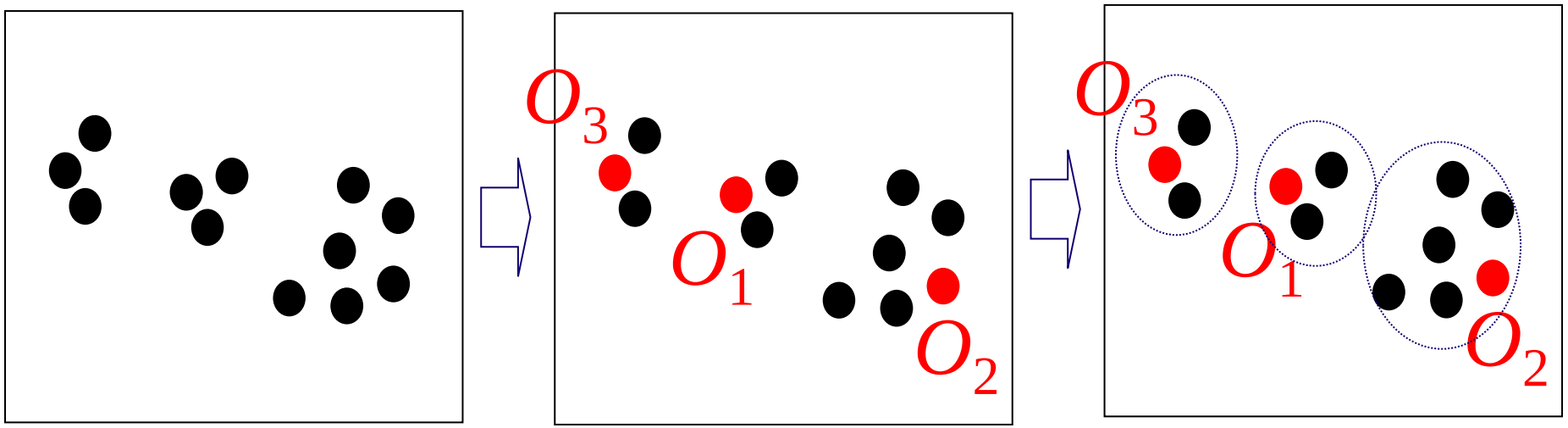
1. k 個のクラスタを持つように、データを分割し, 各クラスタの中心を求める
2. 各点 p_i について、k個のクラスタのうち、クラスタの中心が最も p_i に近いようなクラスタを求め、クラスタを作り直す
3. クラスタの中心を求めなおす
4. 上記の 2, 3 を収束するか、ある規準を満たすまで繰り返す

k-medoid アルゴリズム

- 代表データ (medoid) をデータベースから探す
  - k 個の medoid O_1, O_2, ..., O_k を, クラスタの代表データとして, S から選択
- medoid に対応するクラスタの作成
  - S の残りのデータを，最も相違度が低いクラスタ C_i に分類,
  - k 個のクラスタ C_1, C_2, ..., C_k を作成

$$\min_{1 \le i \le k} D(p_j, Oi), p_j \in S$$

k-medoid アルゴリズム



代表データ (medoid) を
データベースから探す

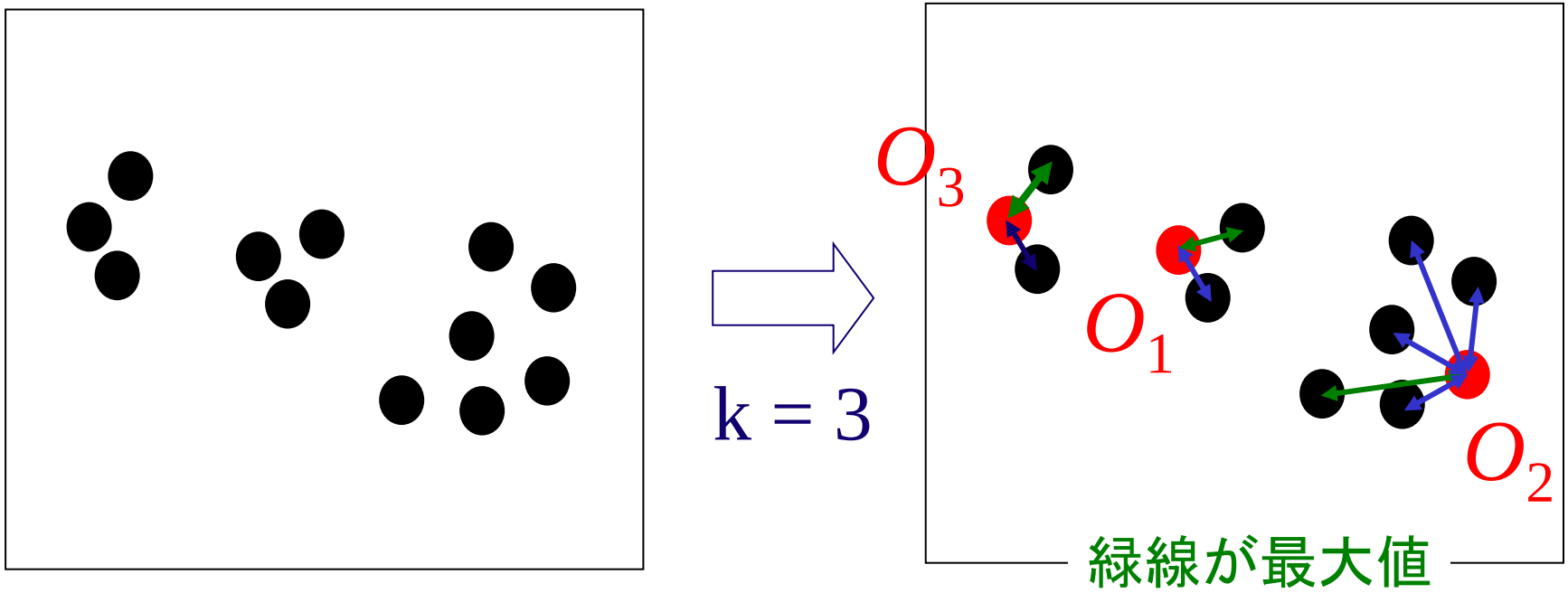
medoid に対応する
クラスタの作成

ここが問題で、種々の手法が
提案されてきた

Medoid とは

- クラスタの代表オブジェクト
- クラスタの「中心」にあるオブジェクトをいかに選ぶかが課題

中心：クラスタ内の他の要素との相違度の最大値が最も小さくなるような要素



Algorithm PAM

1. Select k representative objects arbitrarily.

2. Compute TC_{ih} for *all* pairs of objects O_i, O_h where O_i is currently selected, and O_h is not.

3. Select the pair O_i, O_h which corresponds to $min_{O_i,O_h}\ TC_{ih}$. If the minimum TC_{ih} is negative, replace O_i with O_h, and go back to Step (2).

4. Otherwise, for each non-selected object, find the most similar representative object. Halt.

CLARANS アルゴリズム

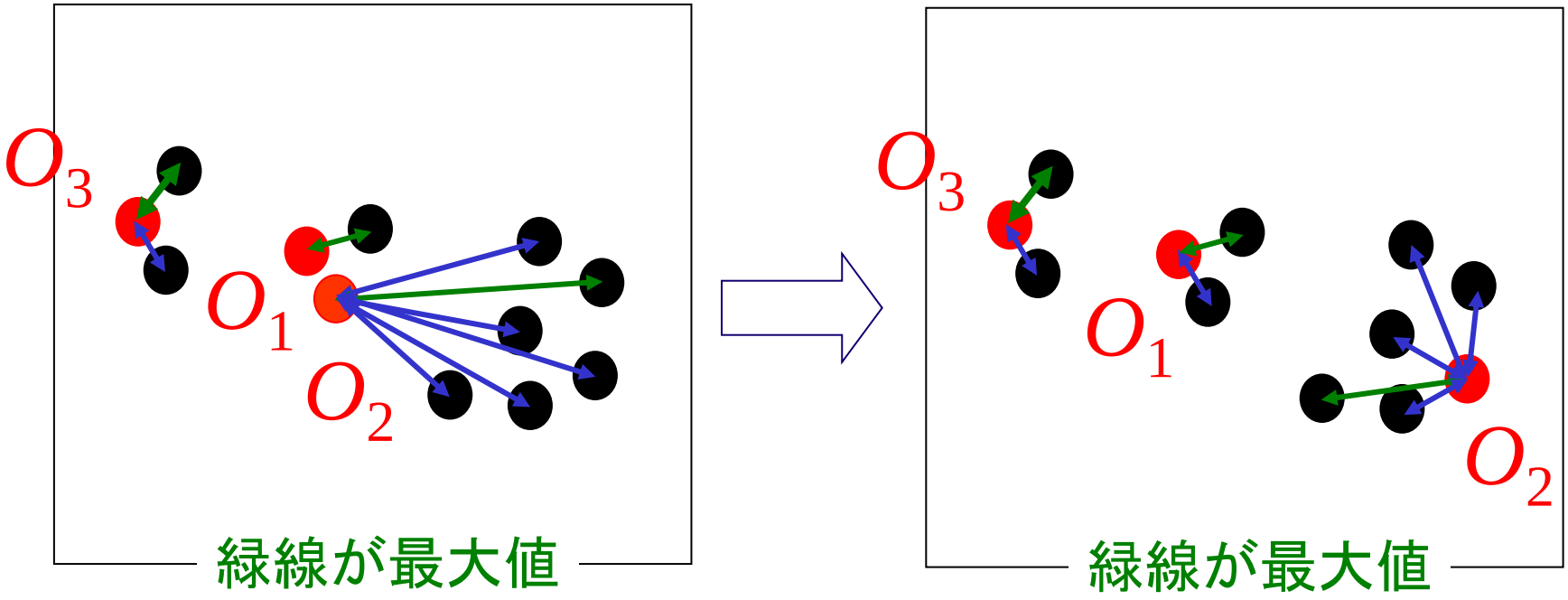
- k 個のデータ $O_1, O_2, ..., O_k$ を，S から適当に選択
- 選択した k 個のデータを，ポテンシャルが小さくなるように改良

$$\sum_{p_j \in S} \min_{1 \le i \le k} D(p_j, O_i)$$

  - S から，ランダムに取替え候補Xを選択

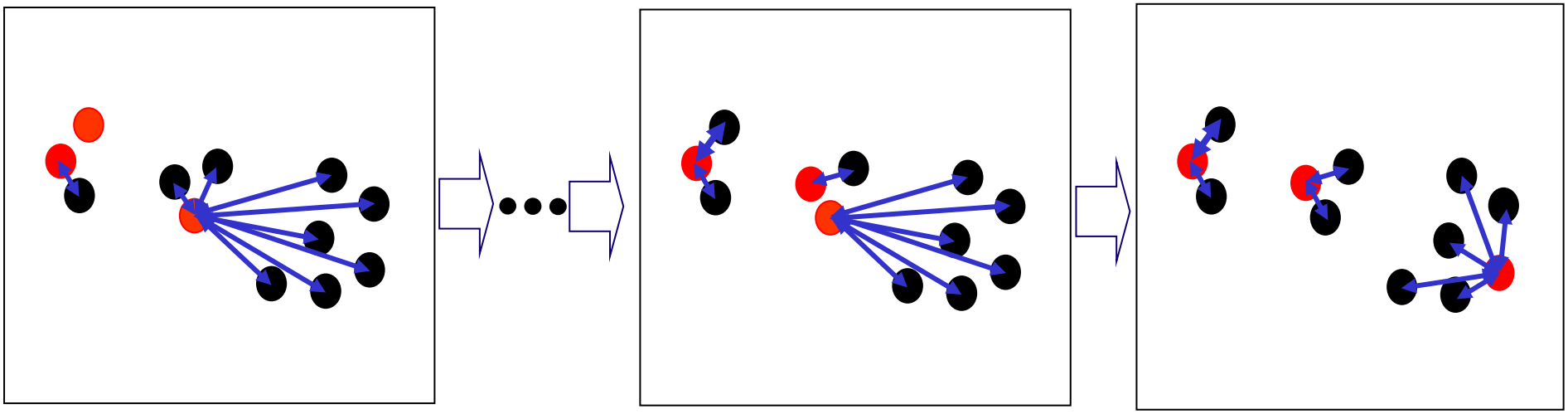
    ⇒ これが、CLARANS のアイデア
  - 選択されるXは，max(n/80, 250)個のランダムデータ (n は全データ数)．最適の相手と交換.
- 以上が済んだら、k 個の medoid が得られる

PAM, CLARA, CLARANS のアイデア



こちらの方が良い

PAM, CLARA, CLARANS のアイデア



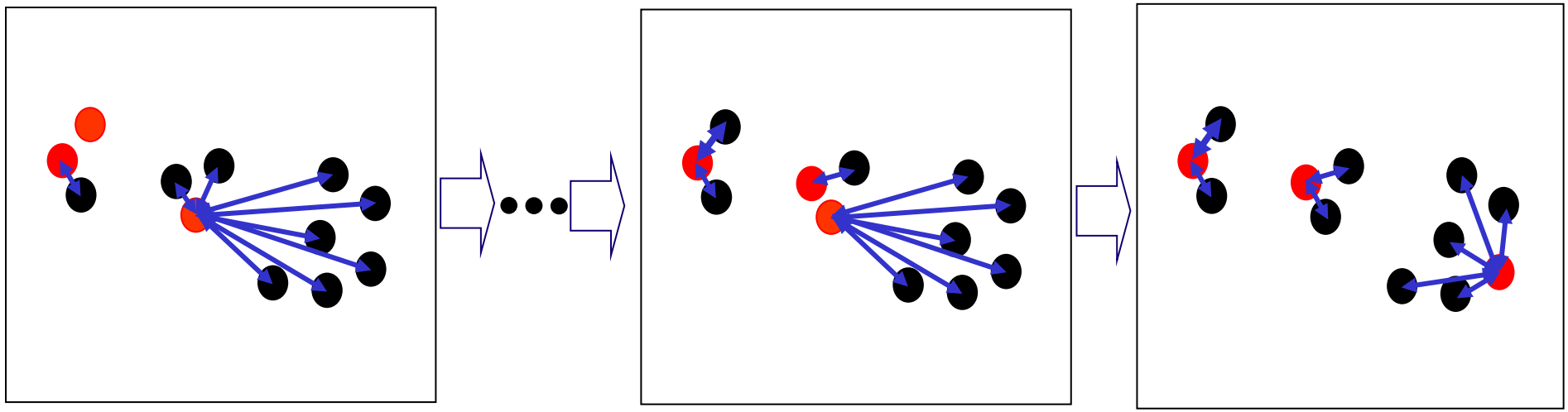
適当な k 個の点を選んで，開始

より良いクラスタリングが得られるように、代表点の交換を繰り返す

交換ができなくなったら終わり

- 定数の近似比(approximation rate) を保証することはできない（理論的なクラスタリングの最適性は保証しない）
- 良いクラスタリングが期待できる

PAM, CLARA, CLARANS のアイデア

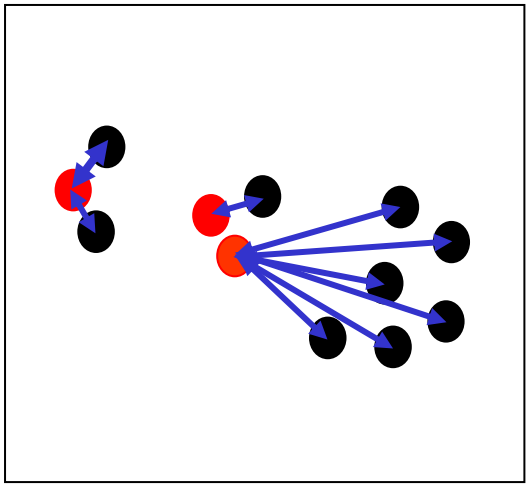


ポテンシャル：高 ⇒ ポテンシャル：低

ポテンシャルの定義：
代表点から、（仮の）クラスタ内の点までの相違度の総和
（青線の長さの和）

代表点 (k個) × 交換相手の候補 (n-k個) の組み合わせを調べ，ポテンシャルが最も低いものと交換する

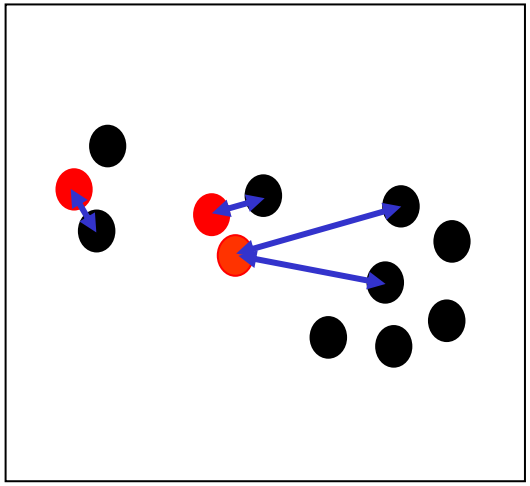
CLARANS のアイデア



PAM

ポテンシャル値を求める

⇒　遅さの原因



CLARANS

・ランダムサンプリング

・ポテンシャルの近似値を求める

・サンプル数は，max(n/80, 250)

Outline

Clustering に関する先行研究の紹介

- farthest-point clustering (1985)
  - 精度 : 近似比2を保証.　O($n\log k$) のアルゴリズム
- クラスタリングの精度は上限がある(1988)
  - 精度 :　最悪の場合の近似比→2は保証できる.
    2を大きく超えて改善することは困難
- CLARANS　(1994)
  - 精度 :　近似比2は保証できない
  - 高速
- その他

クラスタリング問題

入力

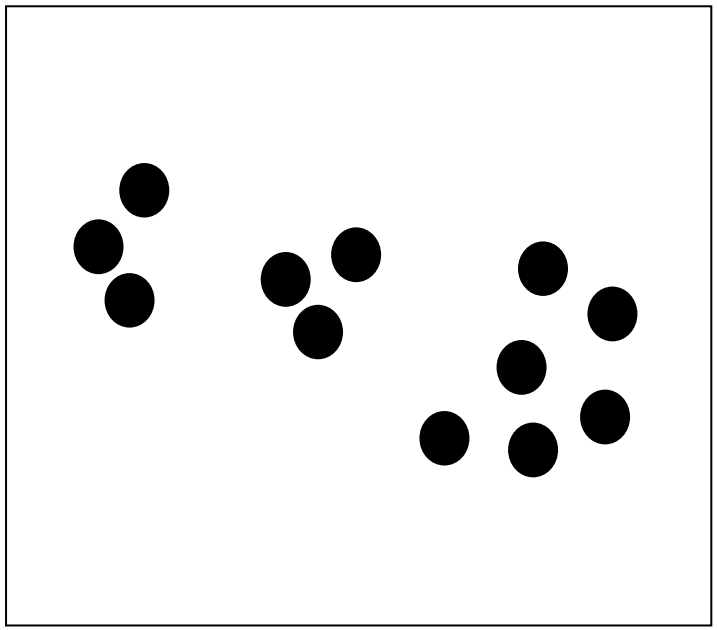
点集合 S

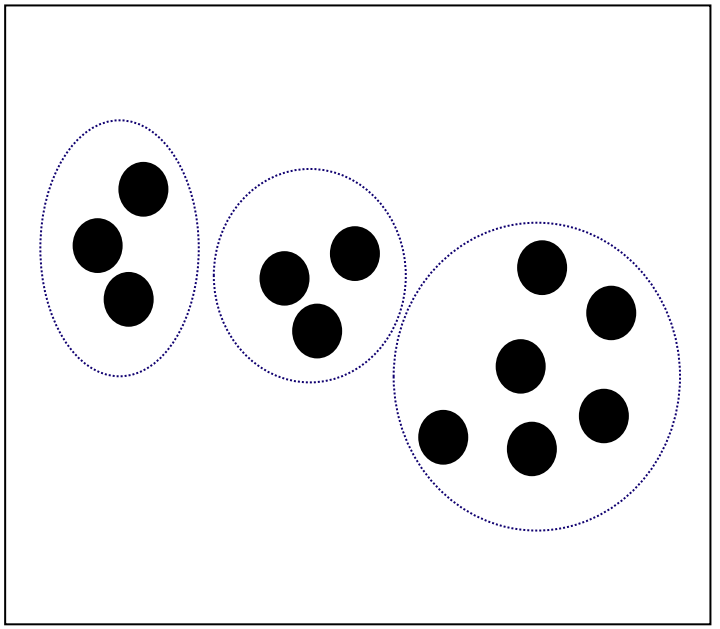
整数 k

出力

S の「良い」分割

$S_1,\ldots,S_k$





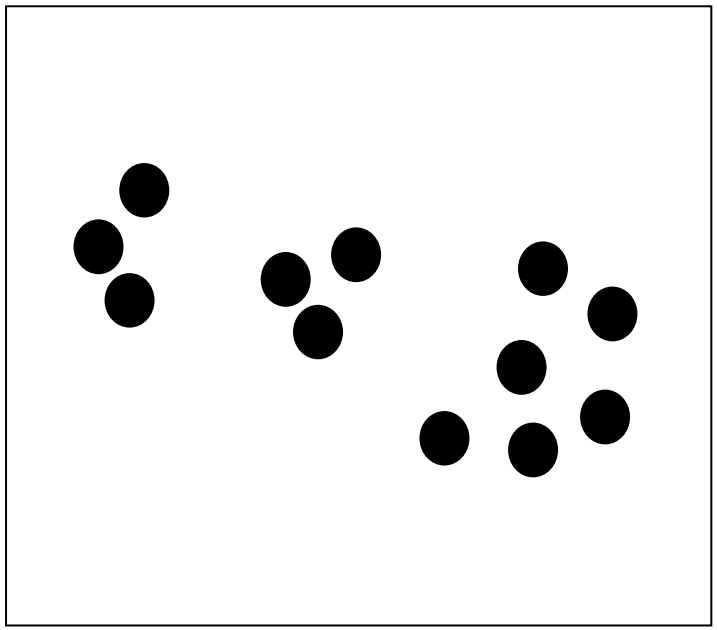
新しいクラスタリング問題

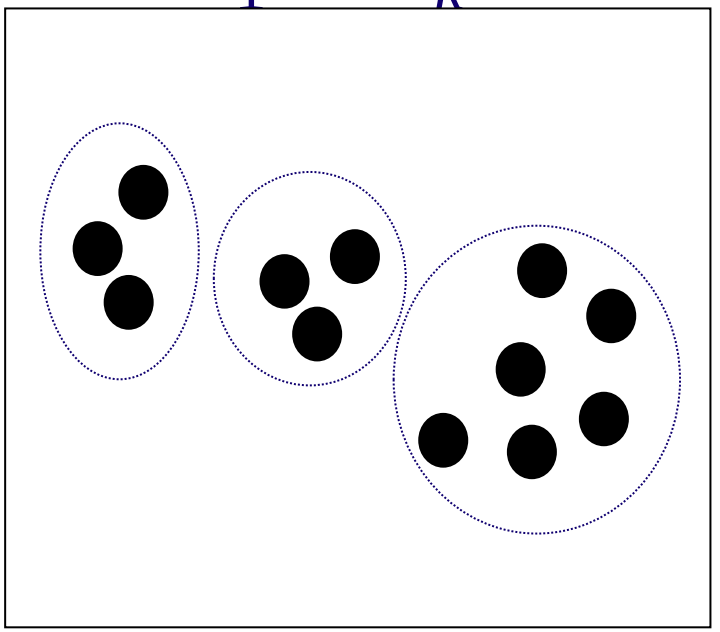
入力

- 点集合 S
- 各クラスタが備えるべき基準

出力

- クラスタ数 k
- S の「良い」分割 $S_1,\ldots,S_k$





クラスタリング問題

入力

点集合 S

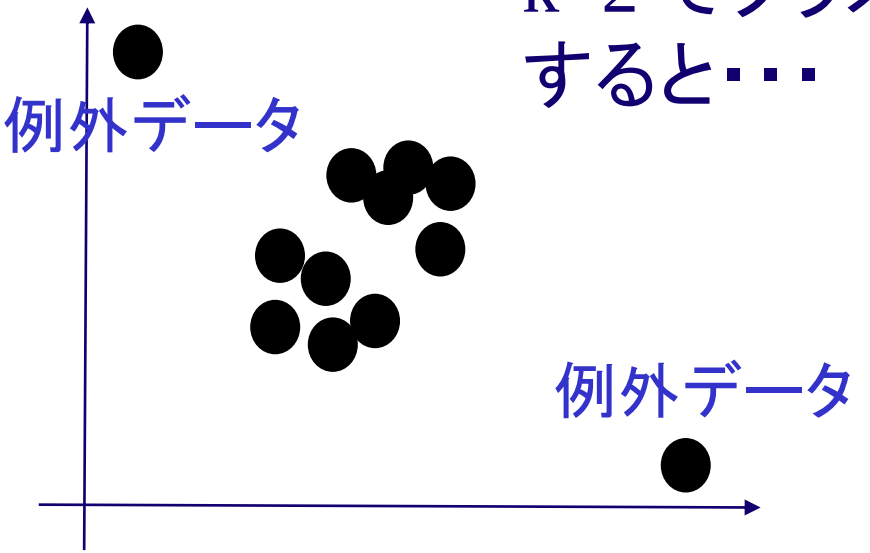
整数 k

出力

S の「良い」分割

$S_1,\ldots,S_k$

k=2 でクラスタリングすると・・・



クラスタリング問題

入力

点集合 S

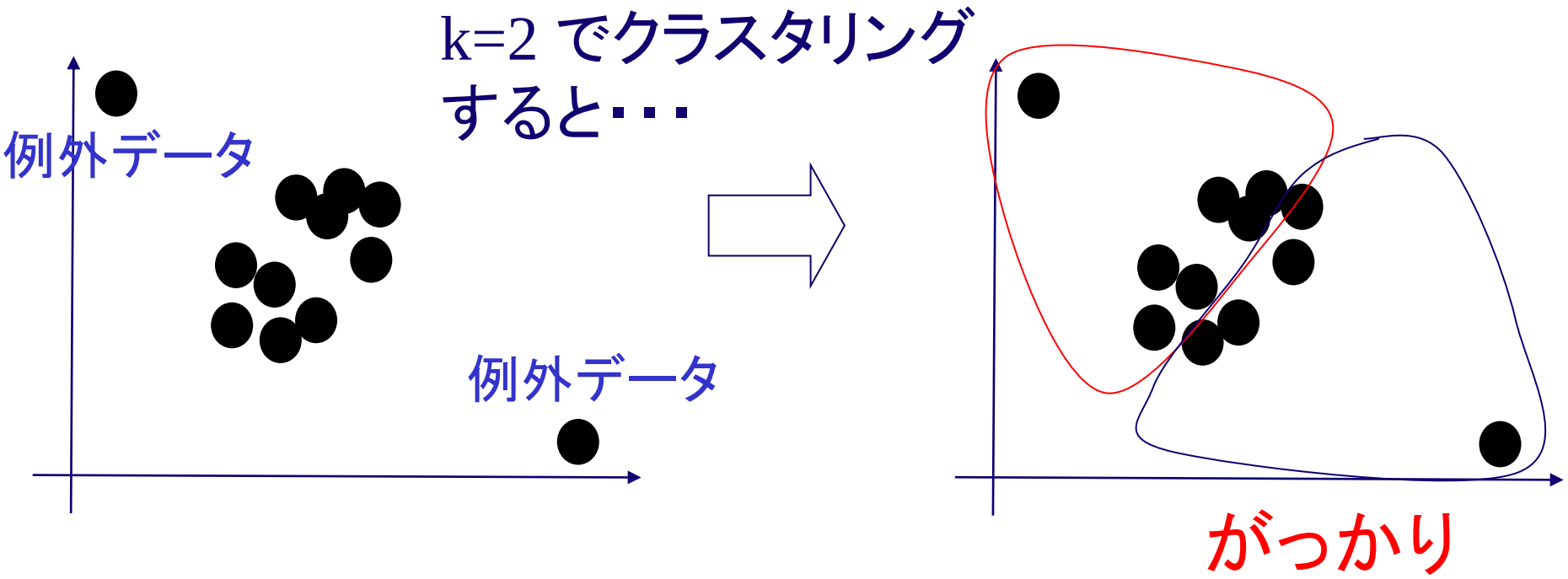
整数 k

出力

S の「良い」分割

$S_1,\ldots,S_k$

k=2 でクラスタリング
すると・・・



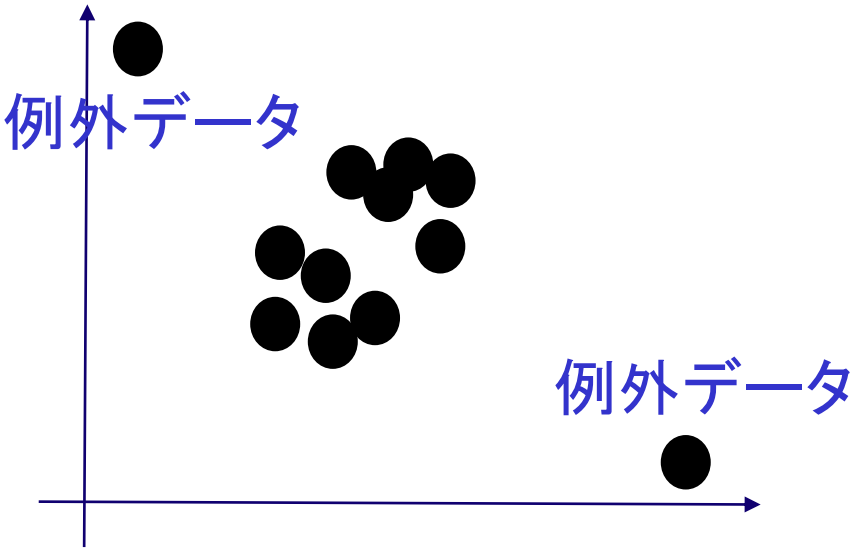
新しいクラスタリング問題

入力

- 点集合 S
- 各クラスタが備えるべき基準

出力

- クラスタ数 k
- S の「良い」分割 $S_1,\ldots,S_k$



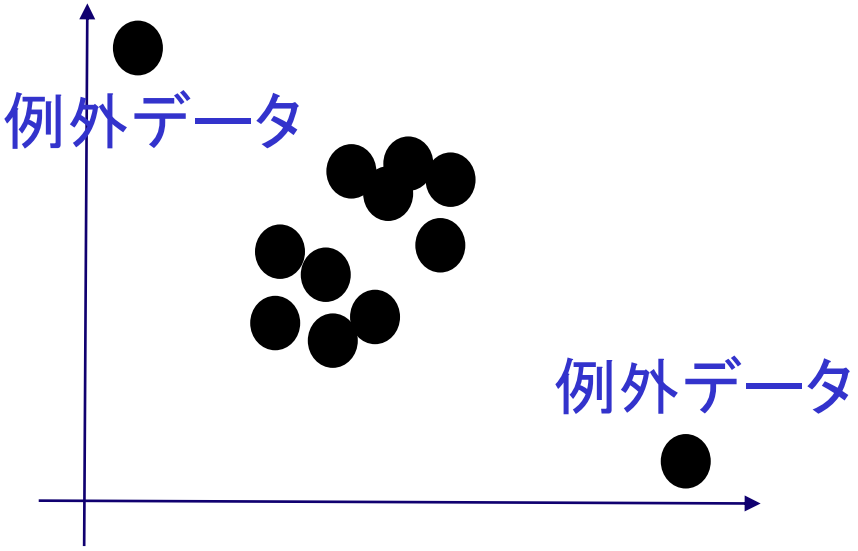
新しいクラスタリング問題

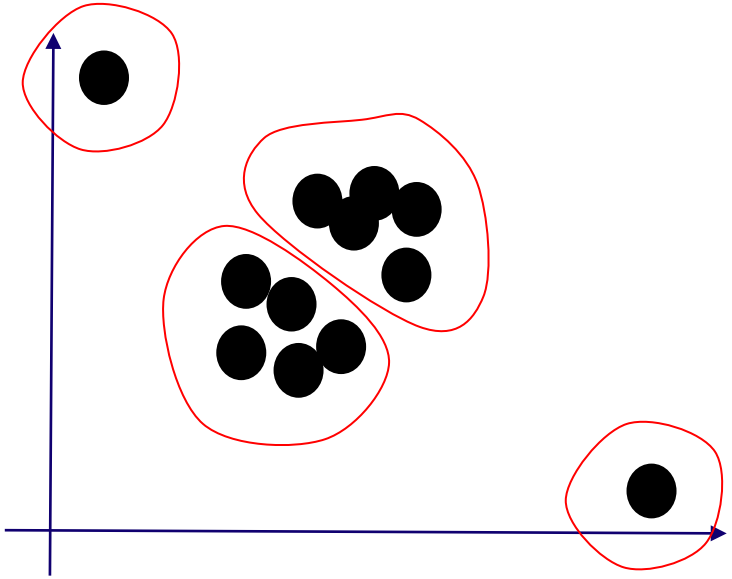
入力

- 点集合 S
- 各クラスタが備えるべき基準

出力

- クラスタ数 k
- S の「良い」分割 $S_1,\ldots,S_k$





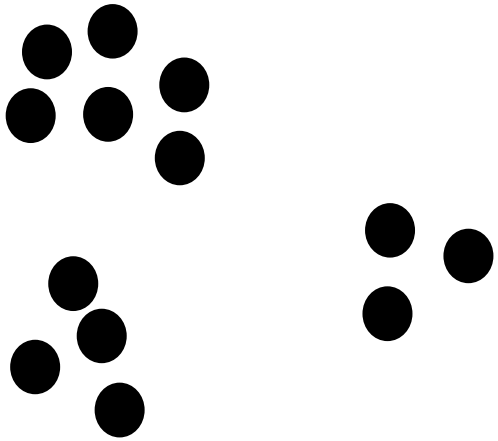
hierarchical algorithms

- BIRCH [7]
  - condensation-based approach, based on the Cluster-Feature Tree
    - 階層構造をもった Cluster-Feature Tree を作る
  - data partitioning according to the expected cluster structure of data
    - クラスタ内分散が、ある閾値以下
    - 閾値を超えそうなときは、leaf node を分割

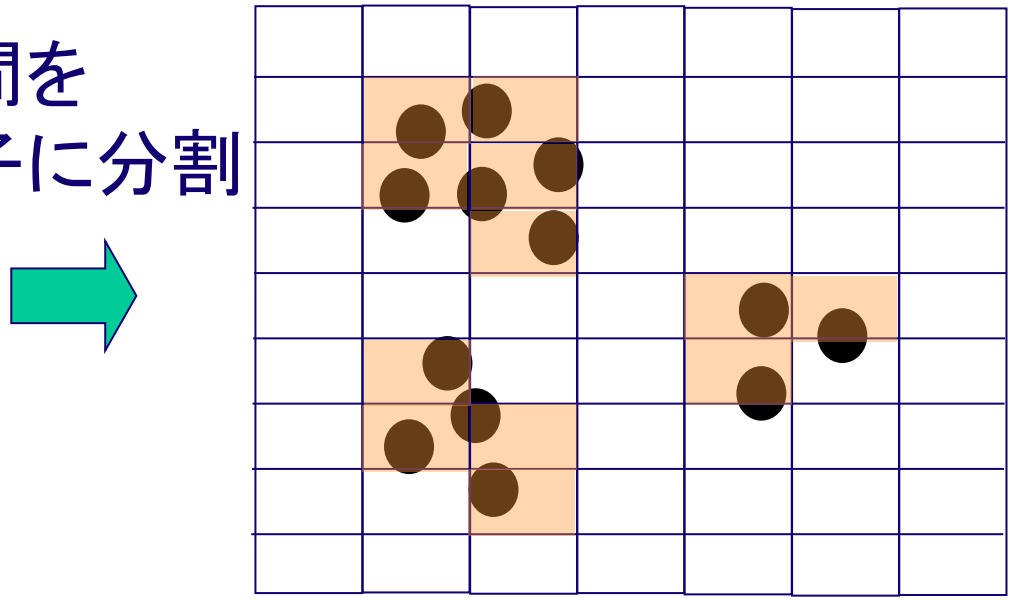
[7] BIRCH: an efficient data clustering method for very large databases,
International Conference on Management of Data,
Proceedings of the 1996 ACM SIGMOD international conference on Management of data,
pp. 103-114, 1996.

density-based algorithms

- データベースに次の性質が成り立てば、大幅な高速化が可能
  - クラスタ外: データの密度(deisity)が低
  - クラスタ内: データの密度(density)が高で，密度がだいたい同じ



空間を
格子に分割



density-based algorithms

- 空間を，格子構造によって区切り，各格子内の点の数（密度）を数える.
  - 密度の低い部分と密度の高い部分ができる
- 密度の高い格子の連結成分を取って，クラスタを得る

density-based algorithms

- DBSCAN(Density-based Spatial Clustering of Application with Noise)
  - locality based clustering algorithm, density-based notion of cluster (1994)
- DBCLASD
  - locality based clustering algorithm,
  - DBSCAN の改良，入力パラメータなしで動く (1998)

space partitioning methods

- grid-based algorithms:
  本来，低次元のために設計されたもの ⇒ grid 数が 次元のべき
  - STING(Statistical Information Grid)
    - used regular grid for an efficient clustering, condensation-based approach, using quadtree-like structure containing additional statistical information (1997)
  - WaveCluster
    - used regular grid for an efficient clustering, wavelet-based approach
- 高次元で動くもの
  - DENCLUE(Density-based Clustgering)
    - 密度情報による方法、condensation-based approach, uses a regular grid to improve efficiency (1998)
  - OptiGrid
    - grid-partitioning, 分割面を計算 (1999)
  - CLIQUE(Clustering in Quest)
    - 基本区間(basic interval) の格子構造を作り，不要属性を排除