クラスタ例からの学習 —クラスタに関する属性の利用—

神嶌敏弘 元吉文男

◆クラスタ例からの学習

- 「クラスタリング」と「例からの学習」を合成した学習問題

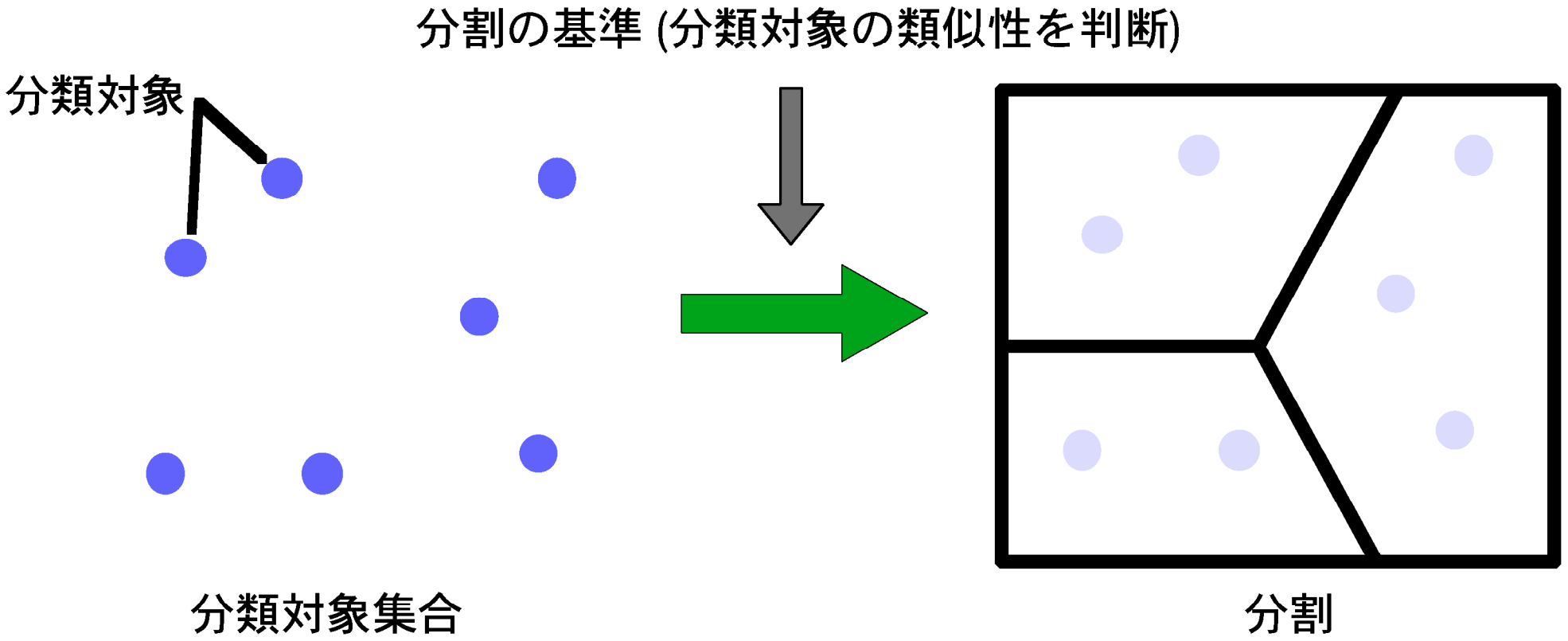
◆問題点

- もっと適切な分割が必要

◆解決法

- クラスタに関する属性を利用できる新しいアルゴリズムを開発

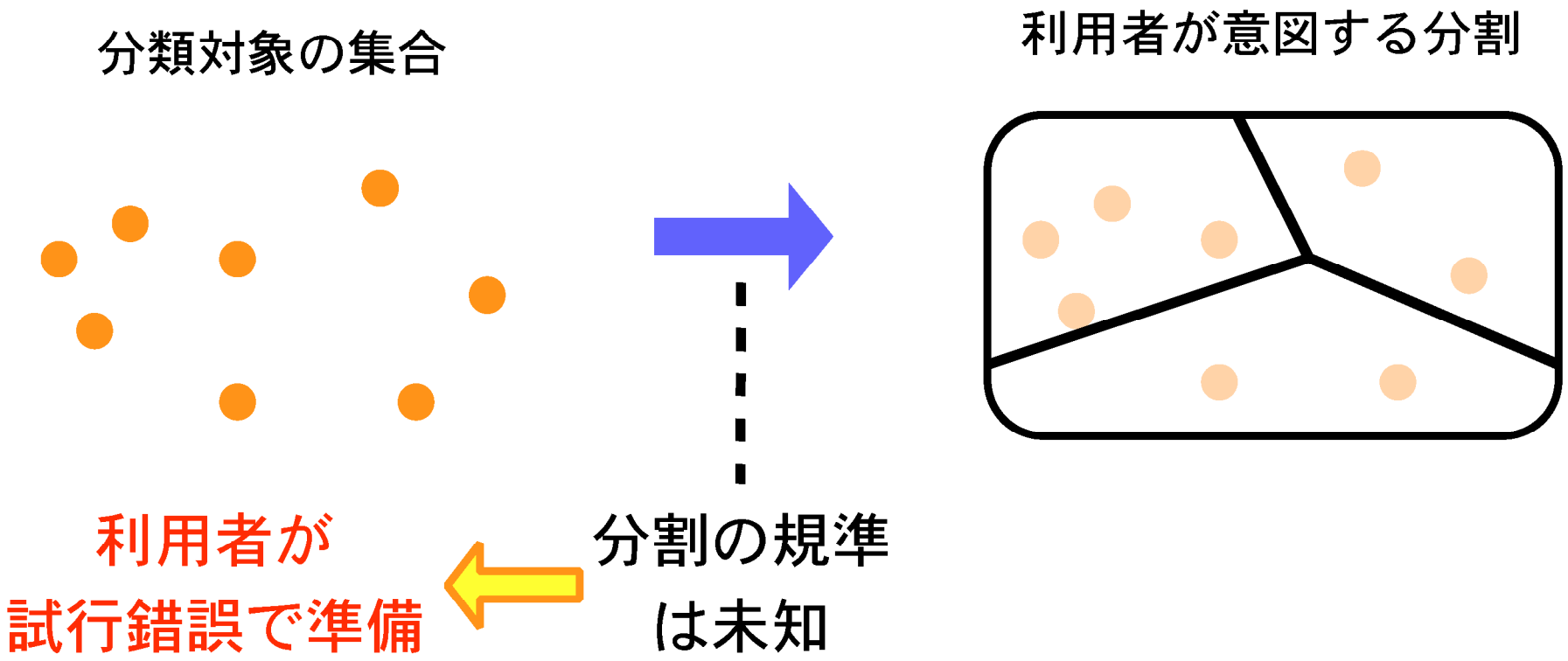
クラスタリング



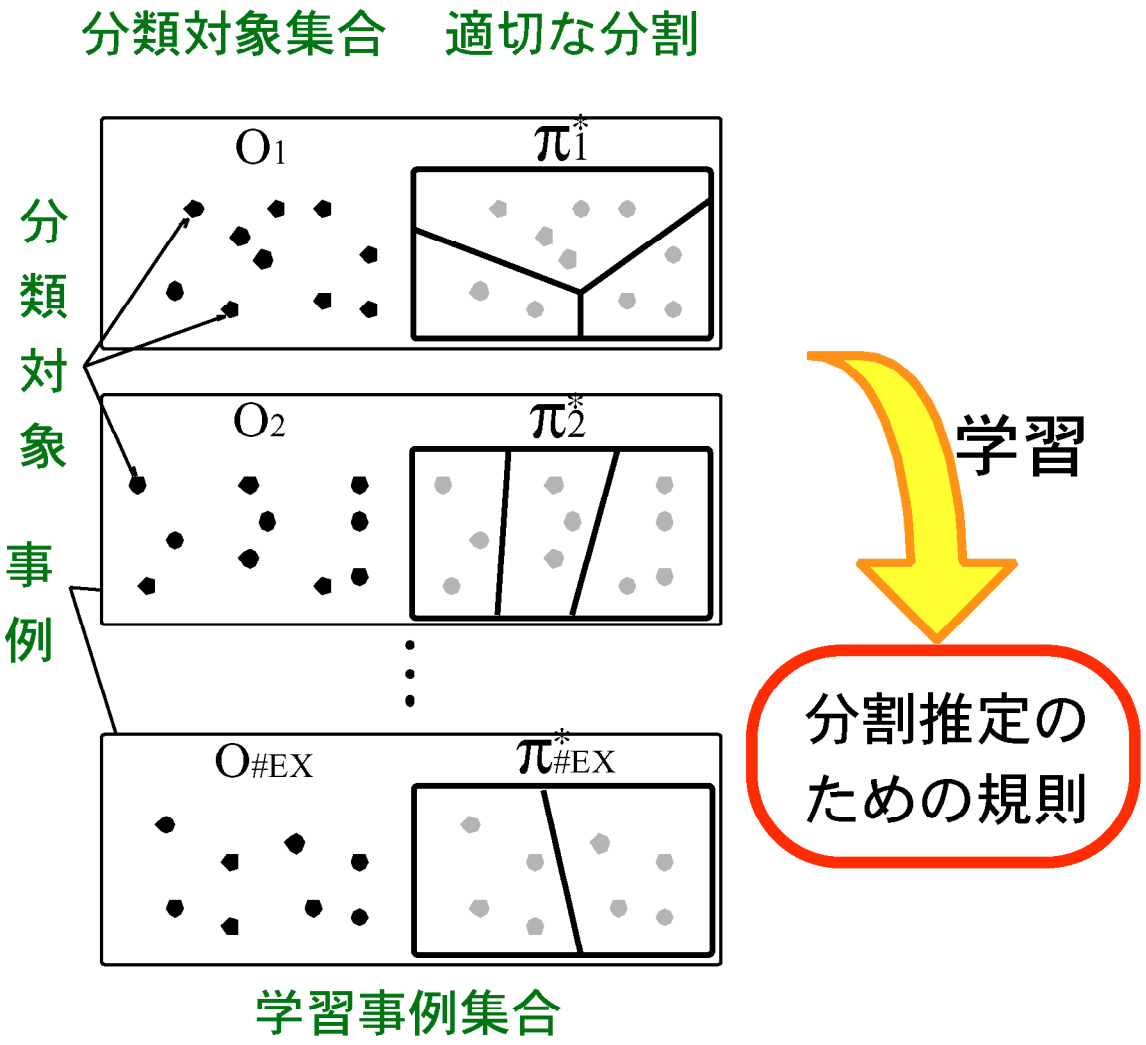
事前に定めた基準をもとに「類似したもの」を集めた部分集合(クラスタ)に分類対象集合をに分割する

適切な分割を導くクラスタリングの利用例

- 分割の規準が事前にあるのではなく，利用者が意図する分割を導く手段としてクラスタリングを利用



クラスタ例からの学習　(学習段階)



学習

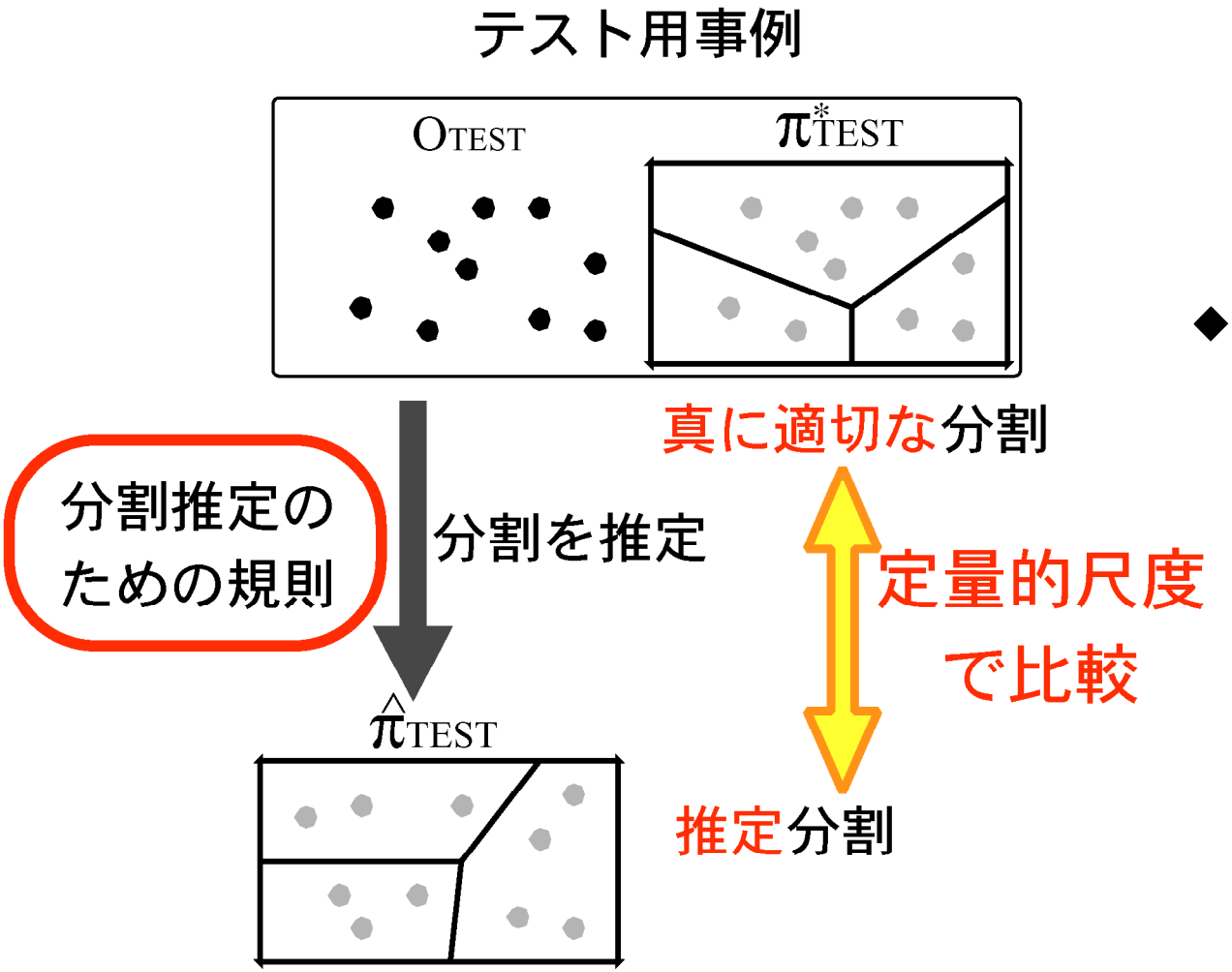
分割推定のための規則

・・・・具体的には
確率密度・分布関数

学習事例集合から，適切な分割を推定する規則を獲得する

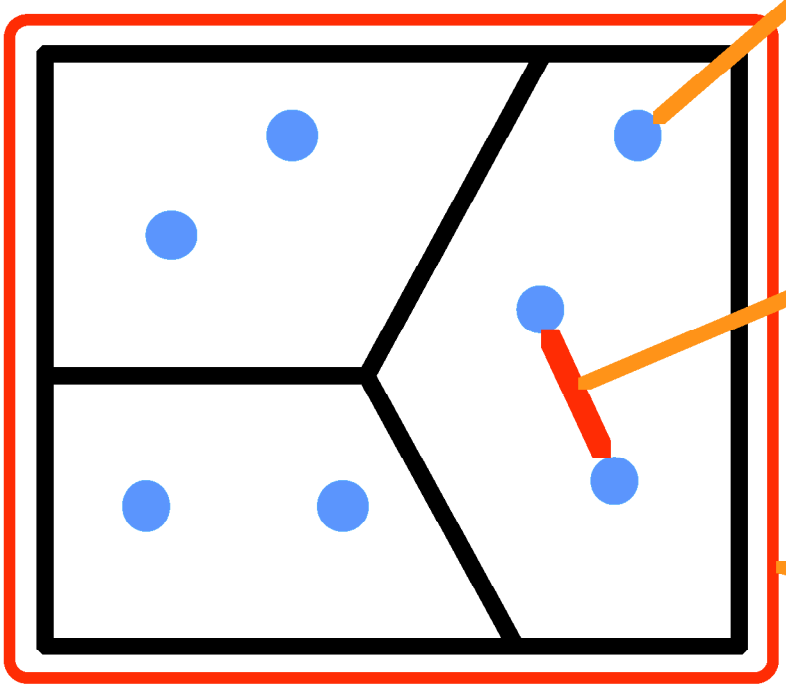
- 各事例は，分類対象集合とそれに対する適切な分割の具体例

クラスタ例からの学習　(推定段階と分割の検証)



- 未知の分類対象集合も分割できるか検証
  - テスト用事例の集合を獲得した規則を利用して分割
  - テスト用事例の真の分割と推定分割を比較して検証

分類対象集合の表現方法



A(o) — 分類対象属性

分類対象の特徴

（例．大きさ）

A(p) — 分類対象対属性

分類対象対の特徴

（例．距離）

A(π) — 分類対象集合全体の属性

分割と分類対象集合全体の特徴

（例．クラスタ数）

分類対象集合の推定方法

分割推定のための規則

$P(\pi=\pi^*, A(\pi), A(P), A(O))$ を最大にする分割

$\pi=\pi^*$ ― πが真に適切な分割であるという事象

$A(\pi)$ ― 分類対象集合全体の属性の属性値

$A(P)$ ― 分類対象対属性値の集合

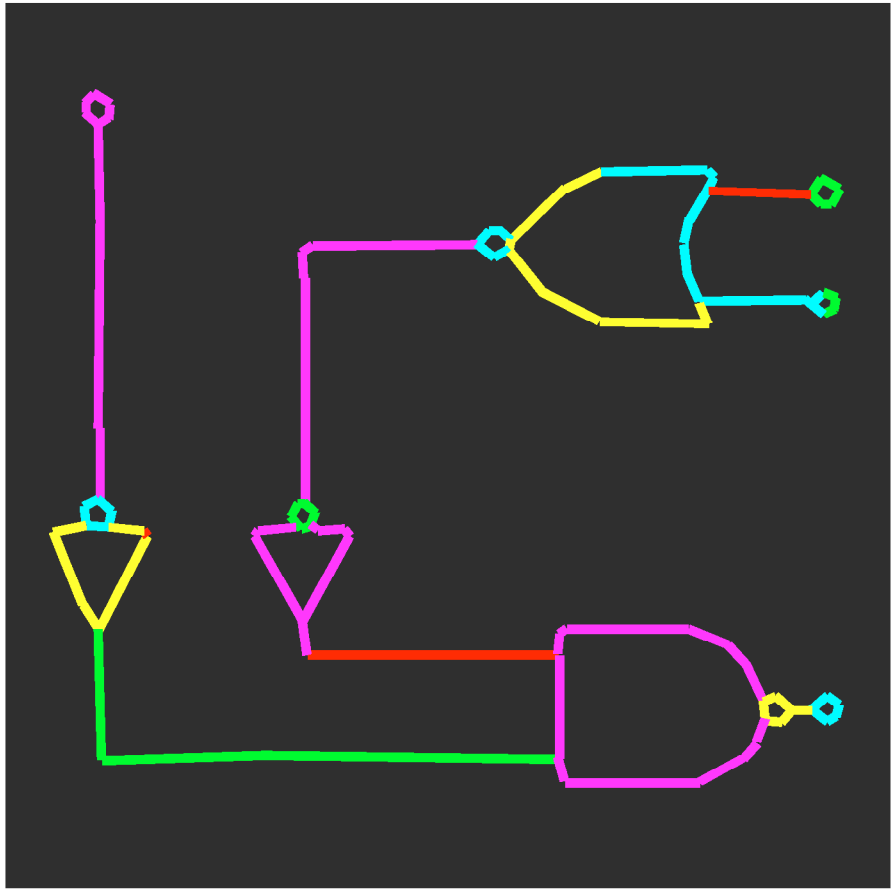
$A(O)$ ― 分類対象属性値の集合

⇩ 簡略化

$$P(A(\pi) \mid \pi=\pi^*) \times P(\pi=\pi^* \mid A(P), A(O))$$

これらの確率密度・分布関数を事例集合から学習

従来の方法で推定した分割の例



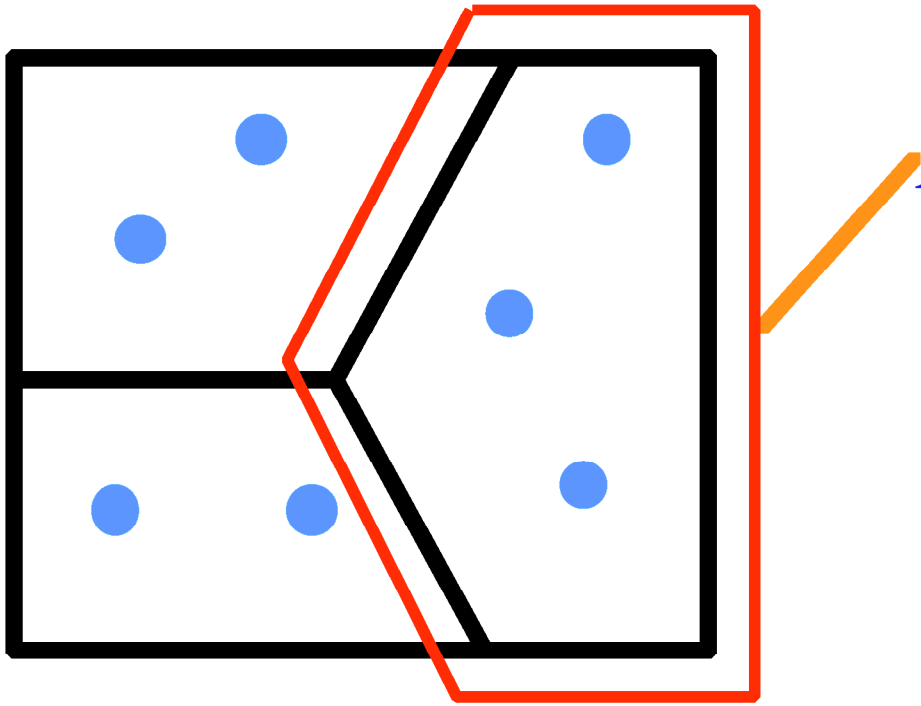
線分で構成されたベクトル画像

図面の各部品が一個のクラスタを構成している分割を求める

より適切な分割を獲得するには？

クラスタごとの特徴（部品の形など）も考慮すればいい

クラスタ属性の導入



$A(C^J)$ ― クラスタ属性

一個のクラスタを構成する
分類対象の集合の特徴

例. クラスタ内のドットの分布
部品の形状

クラスタ属性を利用した分割の推定

分割推定のための規則

$P(\pi=\pi^*, A(\pi), A(C), A(P), A(O))$を最大にする分割

$A(C)$ ― クラスタ属性値の集合

簡略化

$$P(A(C) \mid \pi=\pi^*) \times$$

$$P(A(\pi) \mid \pi=\pi^*) \times P(\pi=\pi^* \mid A(P), A(O))$$

$P(A(C) \mid \pi=\pi^*)$の分布関数を事例集合から学習する必要性

クラスタ属性の確率密度の計算

$P(A(C) \mid \pi=\pi^*)$ はクラスタ数 $\#\pi$ が不定のため計算が困難

↓ 簡略化

クラスタ数 (分類対象集合全体の属性値 $A(\pi)$ の一部)

$$P(A(C) \mid \pi=\pi^*) = \prod_{J=1}^{\#\pi} P(A(C^J) \mid \pi=\pi^*)$$

クラスタ属性値の
集合の確率密度

各クラスタの
属性値の確率密度

実験結果

実験的な事例集合を対象に実験
(ドットパターン)

推定分割の定量的評価　(0から１の範囲，小さい方が優れた結果

以前の結果	0.205
今回の結果	0.223

クラスタ属性を
導入した方が悪い

分割中のクラスタ数の平均

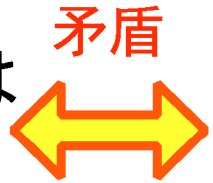
正解	3
今回の結果	3.7

クラスタ数が多めに
推定される傾向

今回の手法の問題点と今後の改良方針

問題点

$P(A(C)|\pi=\pi^*)$ の計算は $\#\pi\ (\in A(\pi))$に依存



分割推定のための規則の簡略化で$A(\pi)$と$A(C)$の独立を仮定

改良方針

$P(A(C)|\pi=\pi^*)$ではなく$P(A(C)|A(\pi), \pi=\pi^*)$を計算

具体的な方法

学習事例の$A(\pi)$の値に応じて

複数のクラスタ属性値の確率密度関数を学習段階で獲得

まとめ

- **結論**
  - クラスタ属性を扱う手法を考案
  - クラスタ属性を採用した利点を生かせなかった
  - クラスタ属性の分布に関し，分類対象集合全体の属性への依存性を無視した点が問題
- **今後の予定**
  - 分類対象集合全体の属性値を条件とするクラスタ属性の条件付確率分布を学習