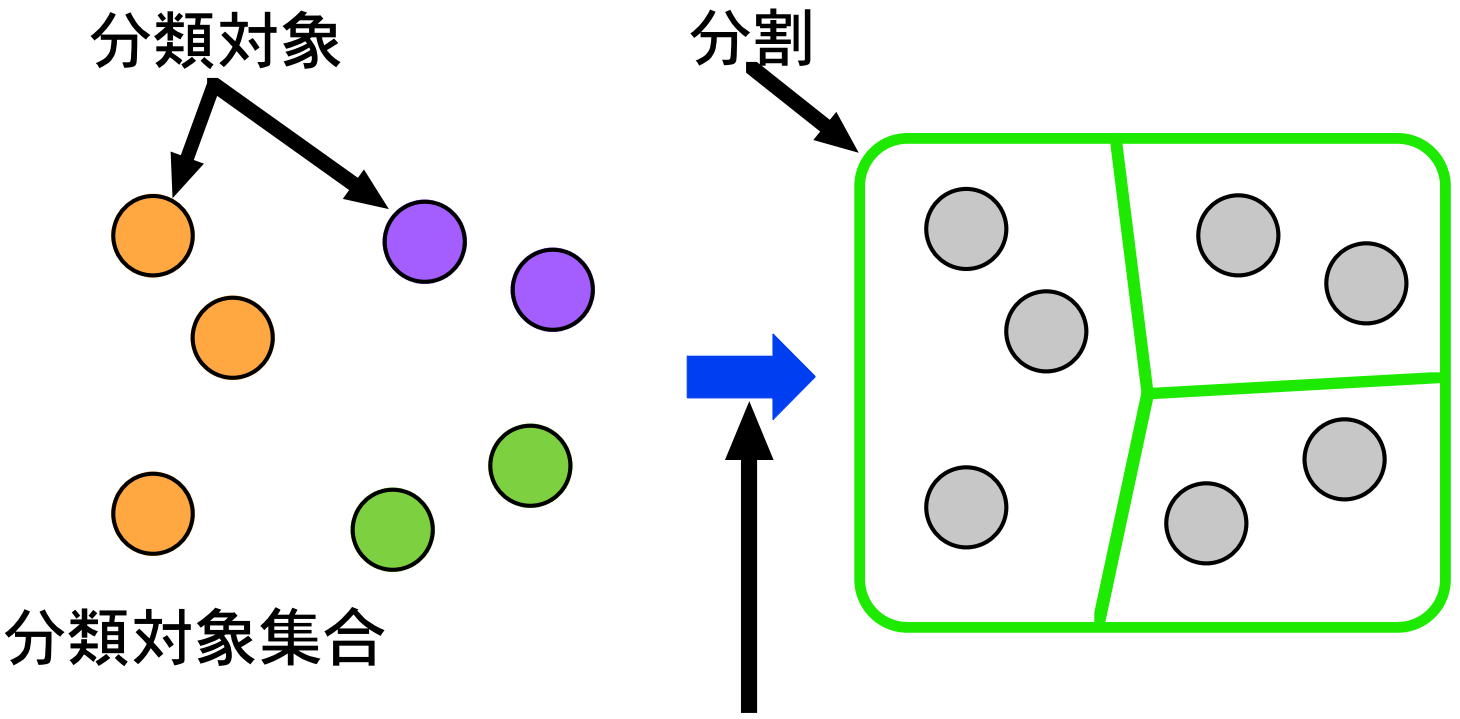
クラスタ例からの学習—クラスタ属性の利用法の改良

神嶌敏弘　元吉文男 (電子技術総合研究所)

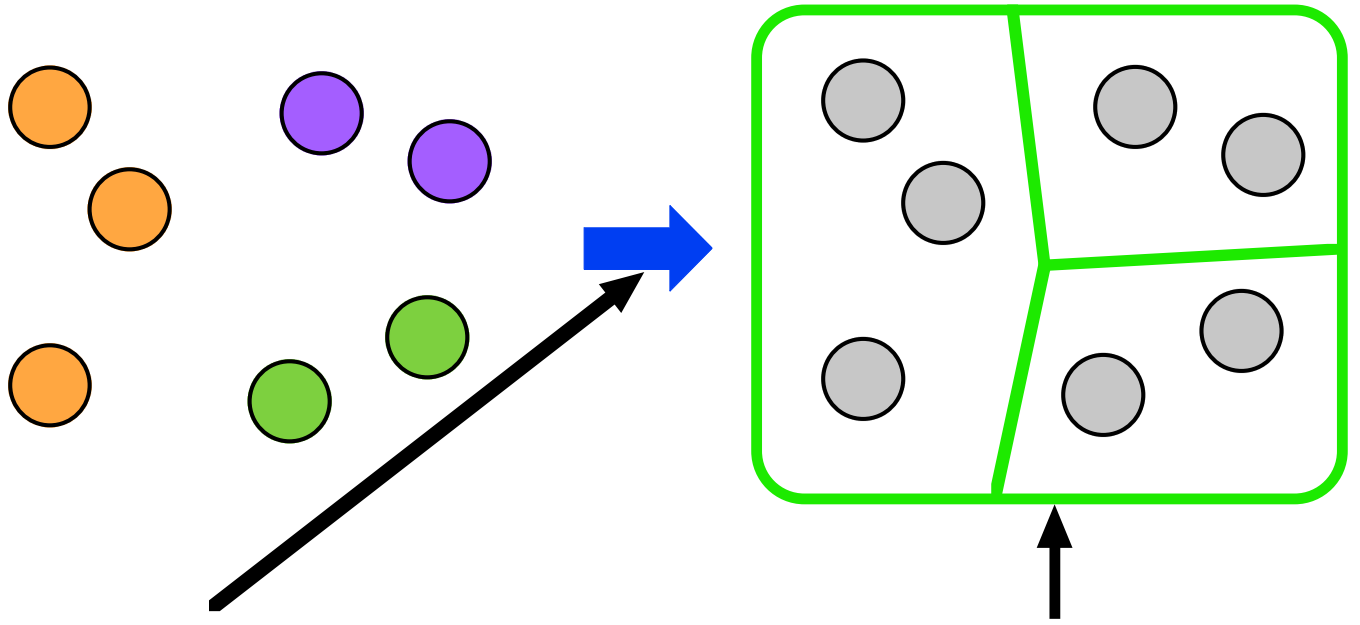
- 研究の経過
  - クラスタリングと例からの学習を合成した学習問題
    （分割の事例から分割のための規則を推定）
- 問題点
  - 推定の精度が不十分
  - クラスタ属性を導入したが効果はみられず
- 今回の解決法
  - クラスタ属性とクラスタ数の依存性を考慮

クラスタリング



事前に定めた基準に基づき「似ているもの」を集めた部分集合（クラスタ）に分類対象集合を分割

クラスタリングによる適切な分割の導出



クラスタリング手法に
内在する基準

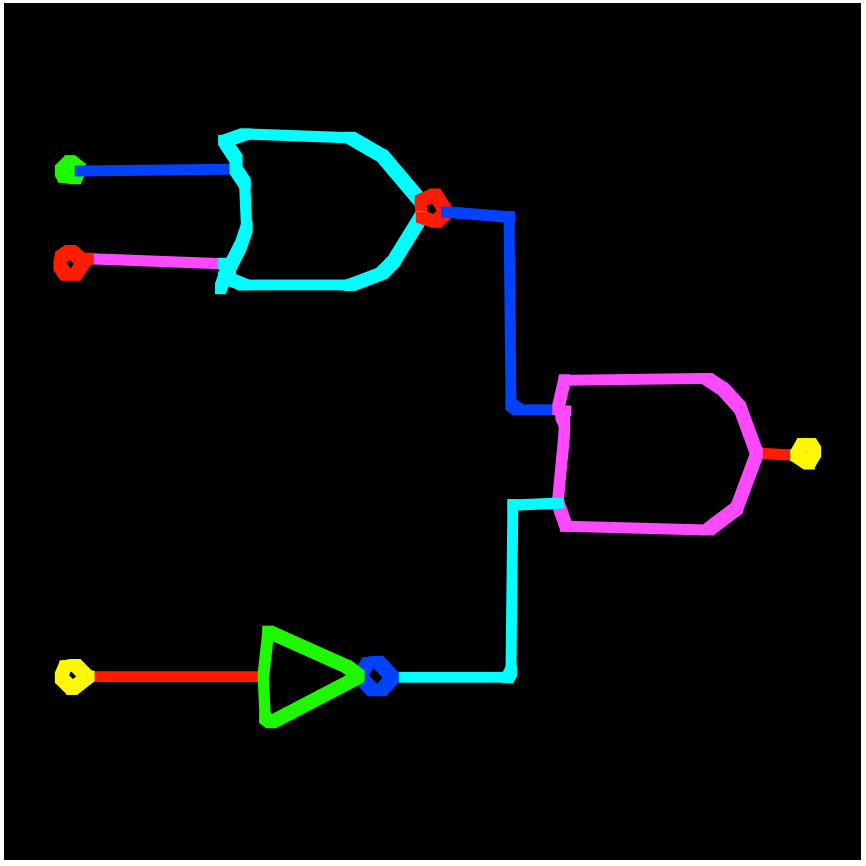
利用者の意図する分割
分割の基準は未知

試行錯誤で一致させるのは困難

適切な分割の導出が必要な状況の例

画像のセグメンテーション……画像の構成要素を何らかの意味をもつ集団ごとにまとめる操作

※画像認識の課程で利用される手法



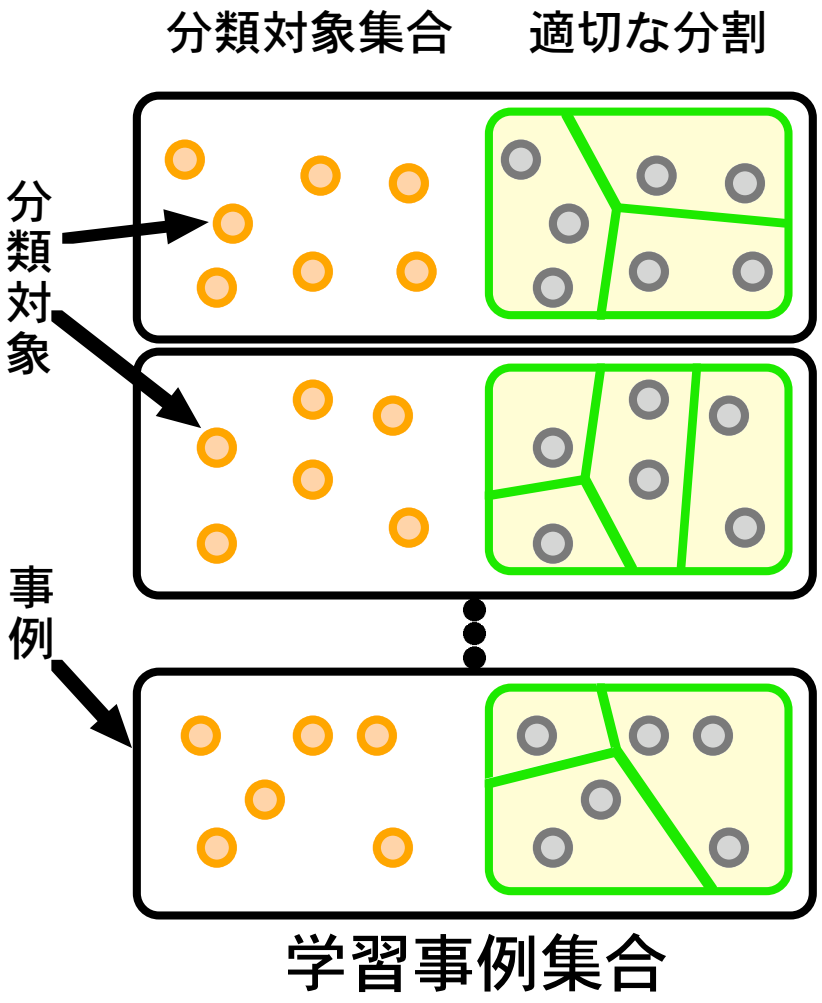
意味をもつ要素を集める基準は未知

⇕

意味をもつ要素を集める基準は未知

・線分の集合で対象を表現したベクトル画像
・図面部品ごとに分割する例

クラスタ例からの学習(学習段階)



分割のための規則

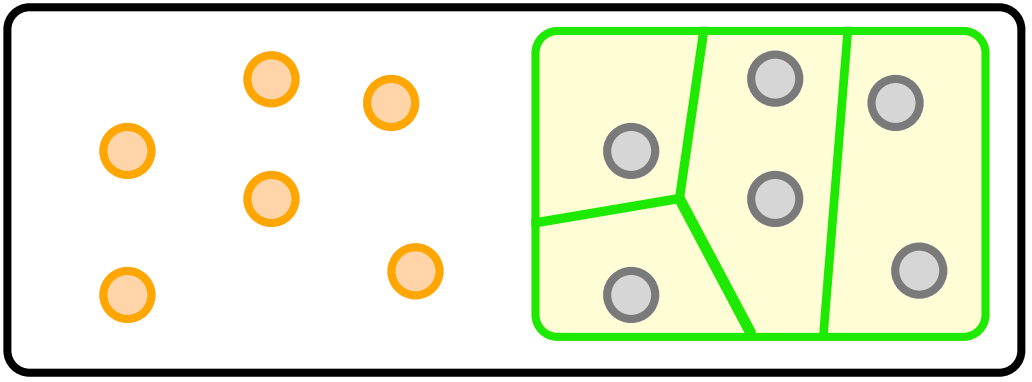
※具体的には確率密度関数

クラスタ例からの学習

学習事例集合(分類対象集合とその適切な分割の組)から分割のための規則を学習

クラスタ例からの学習(推定段階と検証)

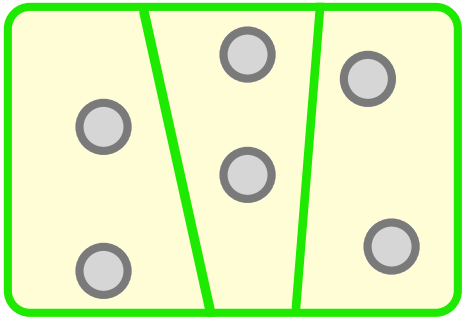
テスト用事例



真に適切な分割

分割のための規則

学習段階で獲得した規則を適用して適切な分割を推定



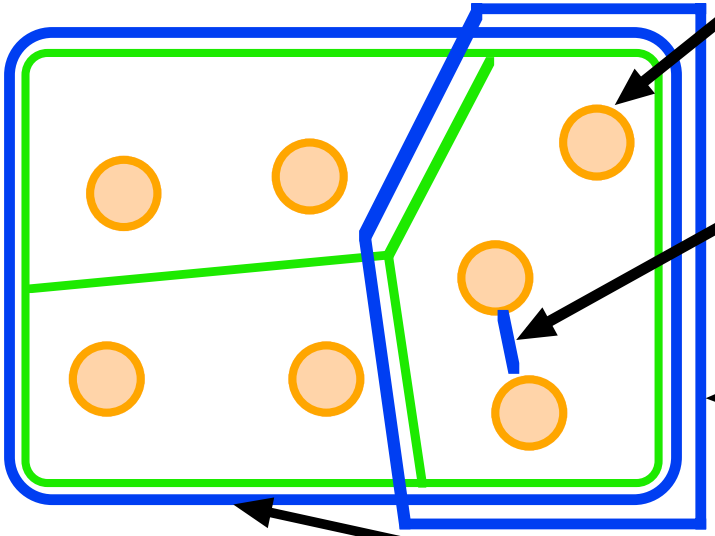
定量的に比較

獲得した規則を検証

推定分割

分類対象集合の表現方法

分類対象集合 O を 4 種類の多数のベクトルで表現



A(o)…分類対象属性

各分類対象の特徴

ex.位置

A(p)…分類対象対属性

各分類対象対の特徴

ex.距離

A(π)…分割全体の属性

分類対象全体の特徴

ex.クラスタ数

A(C)…クラスタ属性

各クラスタ対の特徴

ex.クラスタ内要素数

※ A(π)とA(C)は集合を(仮に)πに分割したとき初めて定まる

分割推定のための規則

与えられた分類対象集合Oの全ての可能な分割の中で
$P(\pi=\pi^*, A(\pi), \{A(C)\}; \{A(p)\},\{A(o)\})$を最大にする分割
を推定分割とする

$\pi=\pi^*$ … πが真に適切な分割であるという事象
$A(\pi)$ … 分割全体の属性
$\{A(C)\}$… 全てのクラスタ属性の集合
$\{A(p)\}$… 全ての分類対象対属性の集合
$\{A(o)\}$… 全ての分類対象属性の集合

結合確率$P(\pi=\pi^*, A(\pi), \{A(C)\}; \{A(p)\},\{A(o)\})$
は複雑なので分解・簡略化➡

結合確率の分解・簡略化

$P(\pi=\pi^*, A(\pi), \{A(C)\}; \{A(p)\},\{A(o)\})$を分解・簡略化

➡以下の3個の確率/確率密度の積に変換

$P(\pi=\pi^*; \{A(p)\},\{A(o)\})$

分類対象対が同じクラスタの要素となる確率の積に分解

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

事例集の各要素について$A(\pi)$を求め，その集合から確率密度関数を推定

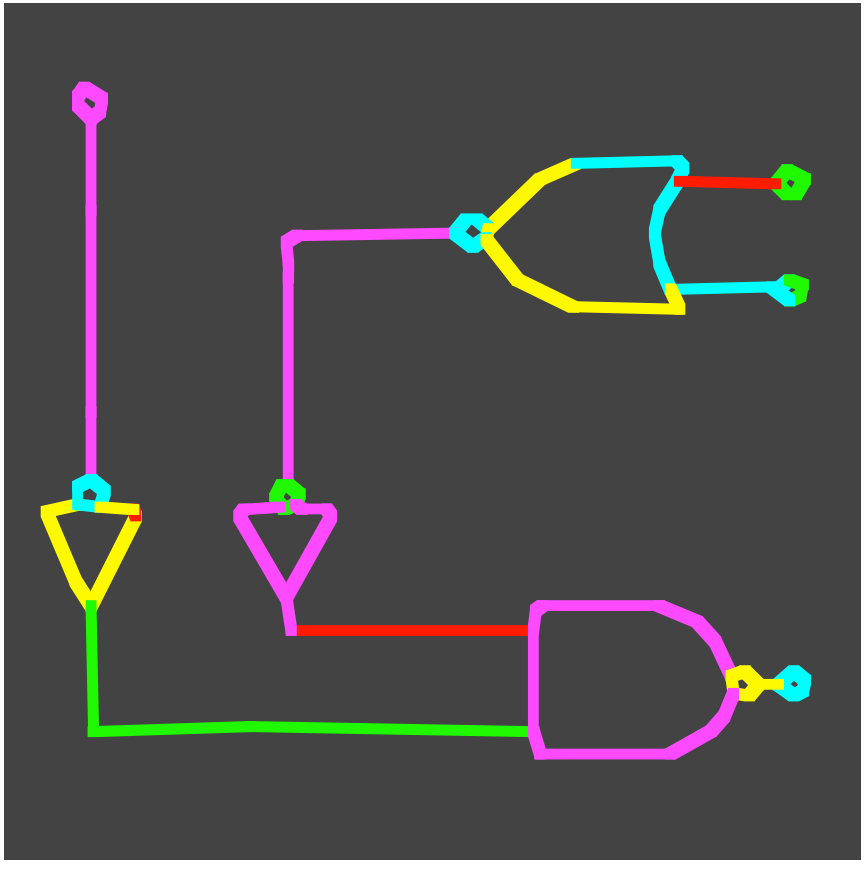
$P(\{A(C)\} \mid \pi=\pi^*)$ ⬅現在の研究対象

事例集の各要素について$\{A(C)\}$を求め，その集合から確率密度関数を推定

従来の方法で推定した分割の例

ベクトル画像の分割例

※線分の集合で対象を表現した画像



◇同じ“部品”を構成する線分がクラスタを構成する分割を求める

◇中間的な結果

◇{A(C)}の項は利用していない

※同色でまとまっている線分が一つのクラスタを表す

P({a(C)}|π=π*)の計算（問題設定）

入力　・学習事例集合の各要素について{A(C)}を計算
・属性ベクトルの各要素を独立とみなしi番目の要素だけに注目

$$
\begin{aligned}
a(C_1) &= \{a(C_1^1), a(C_1^2), \ldots, a(C_1^{\#\pi^{*1}})\} \\
a(C_2) &= \{a(C_2^1), a(C_2^2), \ldots, a(C_2^{\#\pi^{*2}})\} \\
&\vdots \\
a(C_{\#EX}) &= \{a(C_{\#EX}^1), a(C_{\#EX}^2), \ldots, a(C_{\#EX}^{\#\pi^{*3}})\}
\end{aligned}
$$

出力　与えられた分類対象集合をπに分割したときの属性値集合$\{a(C)\}=\{a(C^1), a(C^2), \cdots, a(C^{\#\pi})\}$を引数とする確率密度関数$P(\{a(C)\}|\pi=\pi^*)$

P({a(C)}|π=π*)の計算(1)

集合{a(C)}の要素数#πは分割πに依存して変化

➡P(#π)の分布を別に考え，#πが既知の場合の分布を学習事例から獲得

$$P\left(\{a(C)\}\middle|\pi=\pi^*\right)=P(\#\pi)P\left(\{a(C)\}\middle|\#\pi,\pi=\pi^*\right)$$

※クラスタ属性分布の扱いの改良点 1

- #πは均一分布としていたが，厳密に学習事例から計算
- 学習事例を#π*に応じて分け，それぞれに{a(C)}の密度を計算することで条件付きの分布を計算

P({a(C)}|π=π*)の計算(2)

P({a(C)}|#π, π=π*)の計算

・パラメータΘとそのパラメータの分布を導入

・パラメータ分布を事前分布に{a(C)}の要素が独立に発生するモデルを採用

$$P\left(\{a(C)\}\middle|\#\pi,\pi=\pi^{*}\right)=P(\Theta)\prod_{J=1}^{\#\pi}P\left(a\left(C^{J}\right)|\Theta\right)$$

(P(a(C^J)|Θ)はベータ分布，P(Θ)は対数正規分布)

※クラスタ属性分布の扱いの改良点2

・パラメータの分布を暫定的な方法ではなくEMアルゴリズムで計算

実験結果

ドットパターンの分割問題で予備実験

・二つの改良点(P(#π)の分布の導入，EMアルゴリズムの採用)のいずれも効果はなかった

分割の適切さを評価する結合確率を計算するための三つの確率・確率密度について調査

・真に適切な分割でのP({a(C)}|π=π*)の値が，推定された分割よりも大きかったが，その差は他の項の誤差分よりも小さく有効に作用していない

考察と今後の予定

- {A(C)}があまり有効でない詳細な原因は不明
- 推測➡ベータ分布の問題
  - 連続関数なので急激な変化の多い属性値の分布を十分に表現できない
  - 無限大になることが多く積分で誤差が集積する
- ヒストグラムのような形の分布を試す予定