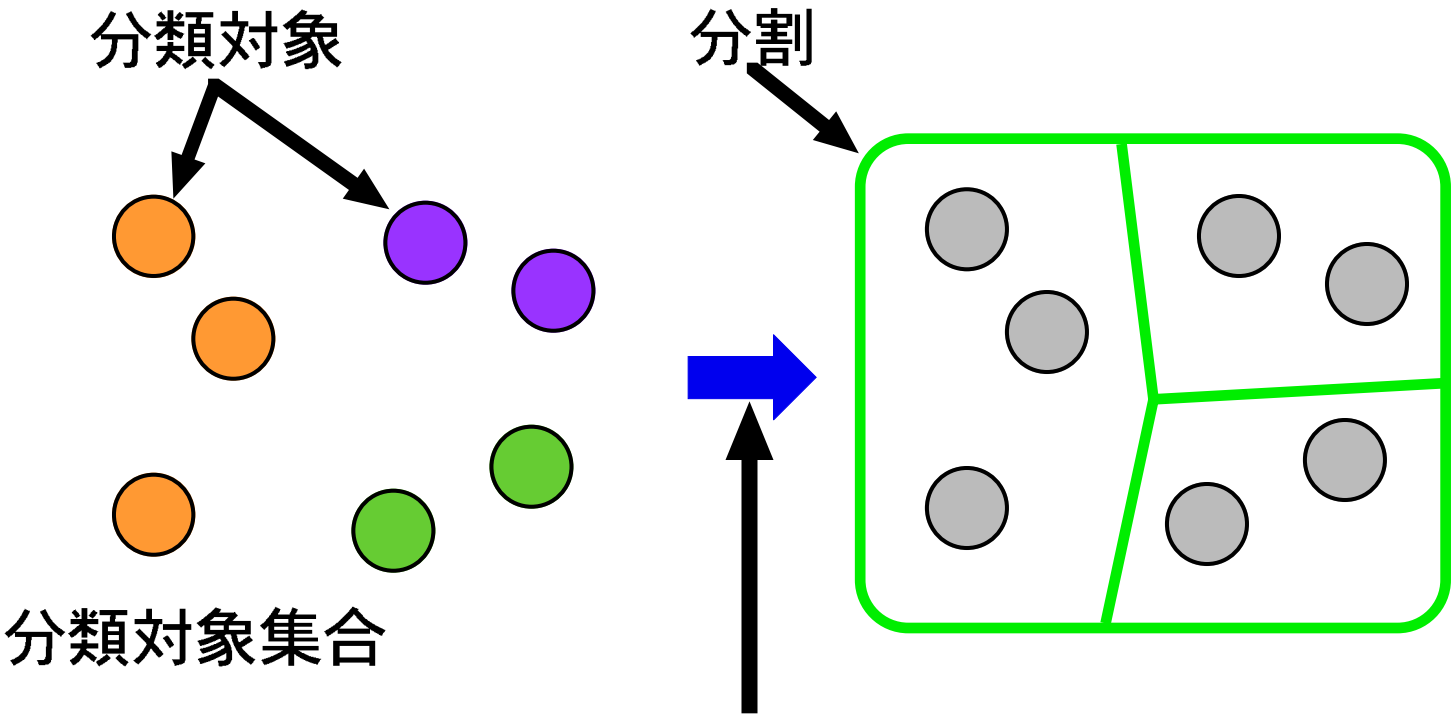
クラスタ例からの学習ークラスタ属性の利用法の改良(2)

神嶌敏弘　赤穂昭太郎 (産業技術総合研究所)

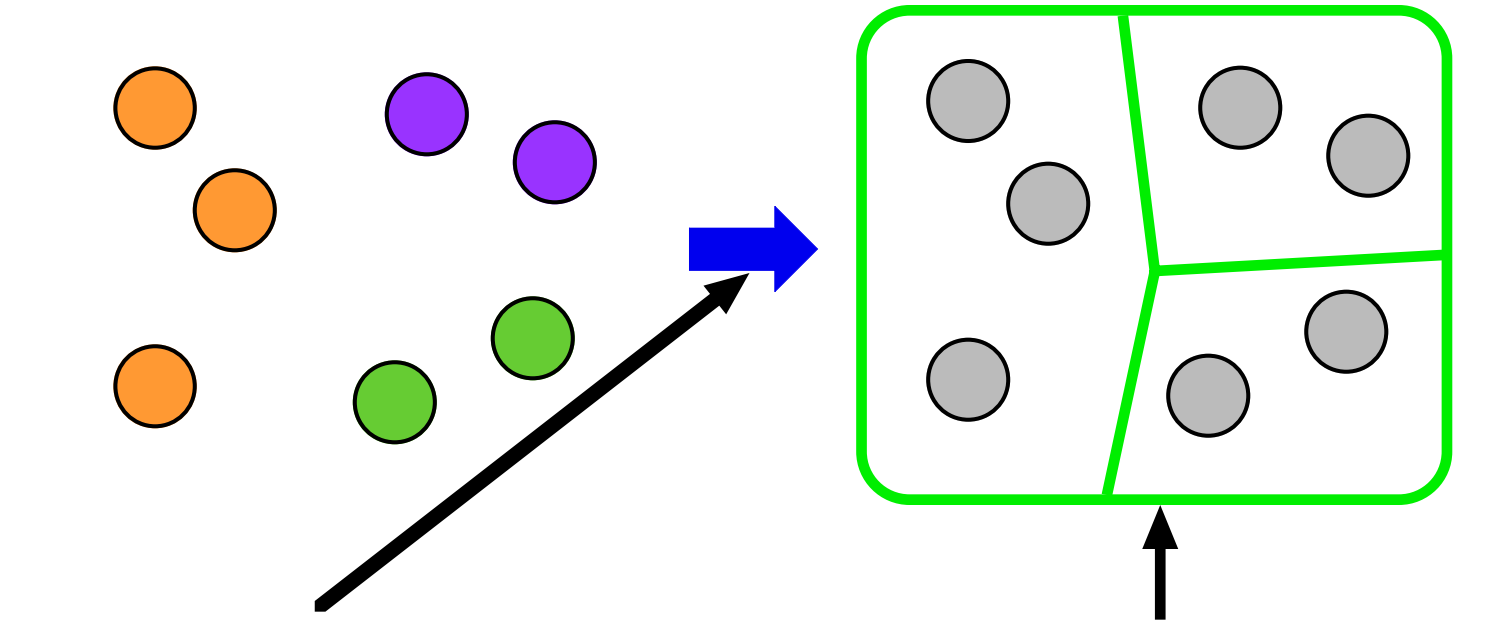
- 研究の経過
  - クラスタ例からの学習：クラスタリングと例からの学習を合成した学習問題
    （分割の事例から分割のための規則を推定）
- 問題点
  - 推定したクラスタの精度が不十分
    ➡クラスタ属性を導入したが効果はみられず
- 今回の手法
  - クラスタ属性の分布の族の改良

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

利用者の意図する分割の導出



クラスタリング手法に
内在する基準

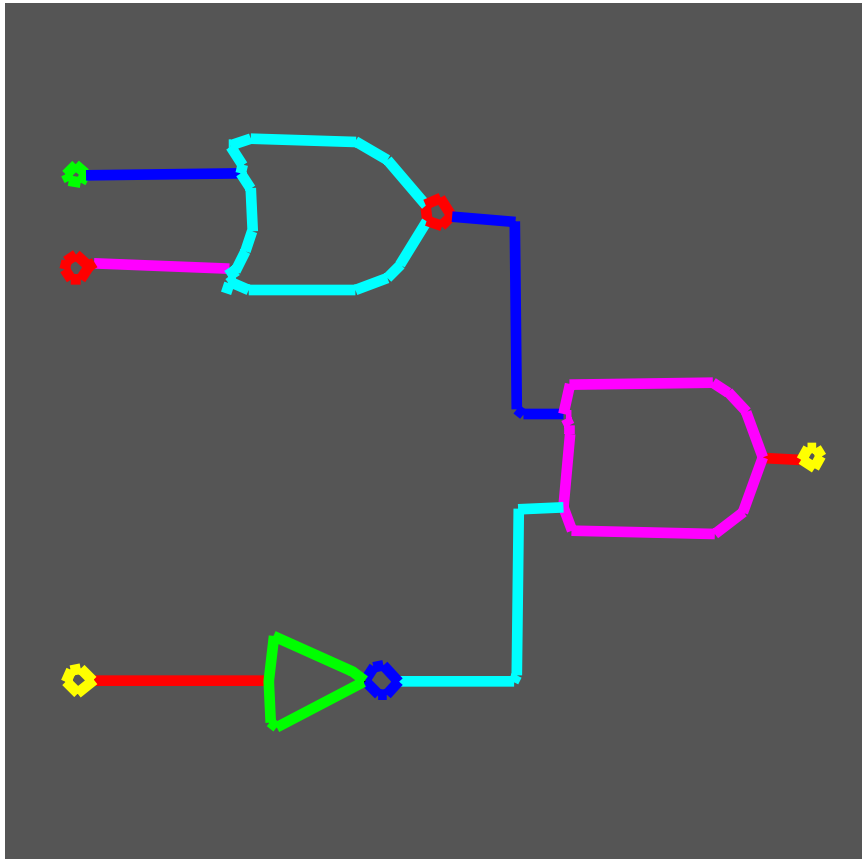
利用者の意図する分割
分割の基準は未知

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



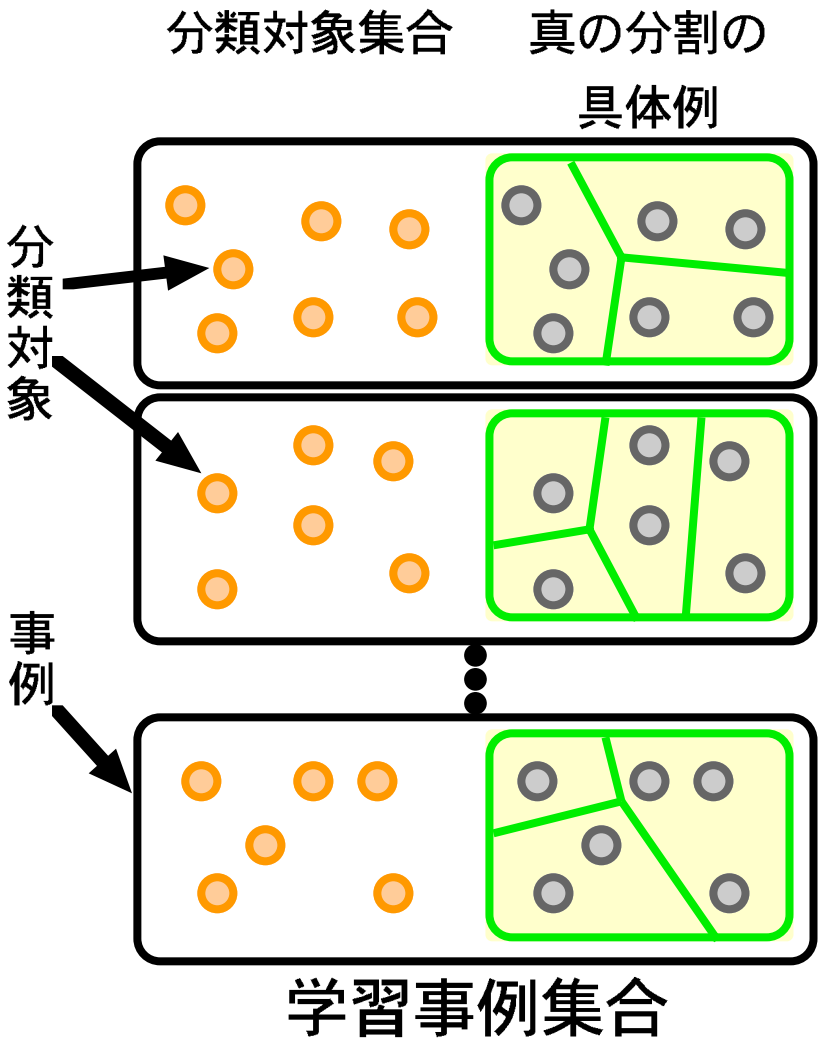
クラスタリングには分割の規準が必要

⇕

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)

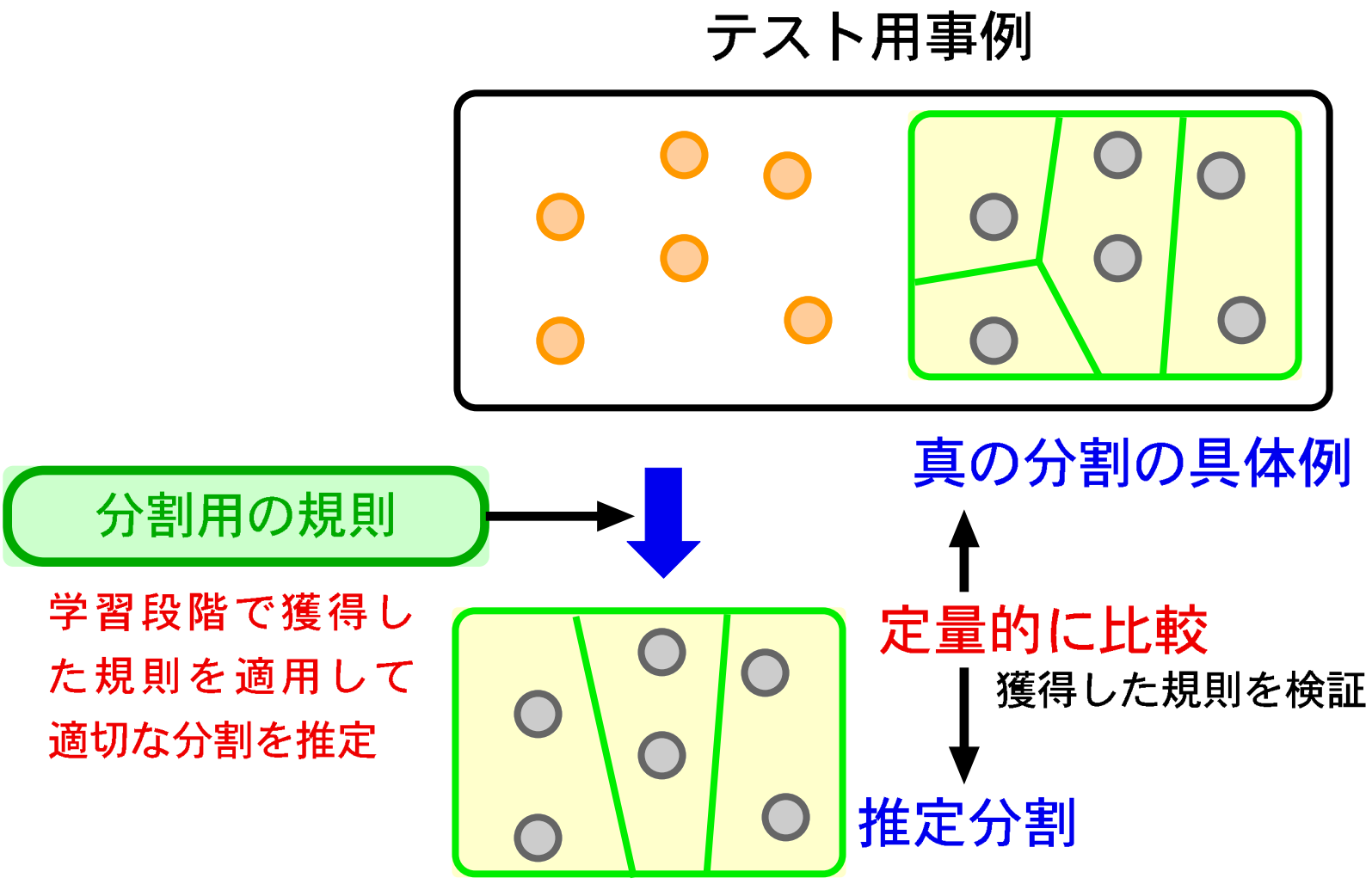


分割用の規則

クラスタ例からの学習

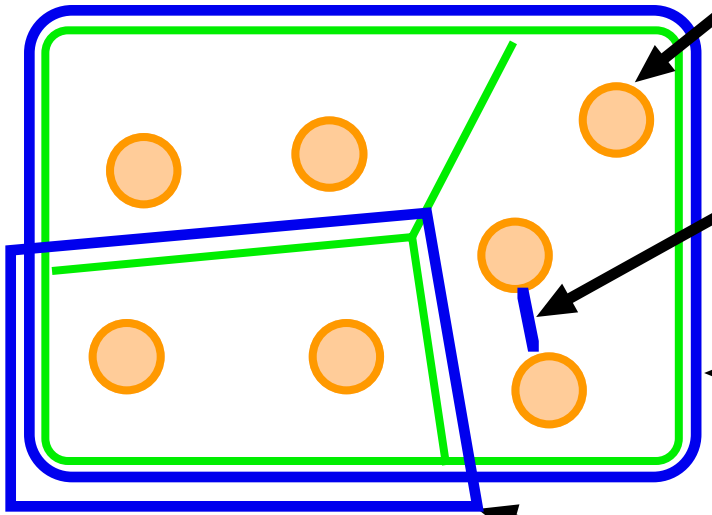
学習事例集合(分類対象集合とその適切な分割の組)から分割のための規則を学習

クラスタ例からの学習(分割段階と検証)



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.クラスタ中の分類対象数

※ 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
$Pr[\pi=\pi^*, A(\pi), \{A(C)\}, \{A(p)\},\{A(o)\}]$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

学習段階：学習事例から評価関数を推定

分割段階：分割の中で評価関数を大きくする分割を探索

結合確率の分解・簡略化

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

➡以下の3個の確率/確率密度の積に変換

$Pr[\pi=\pi^*| \{A(p)\},\{A(o)\}]$

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

$Pr[A(\pi) | \pi=\pi^*]$

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

$Pr[\{A(C)\} | \pi=\pi^*]$ ⬅現在の研究対象

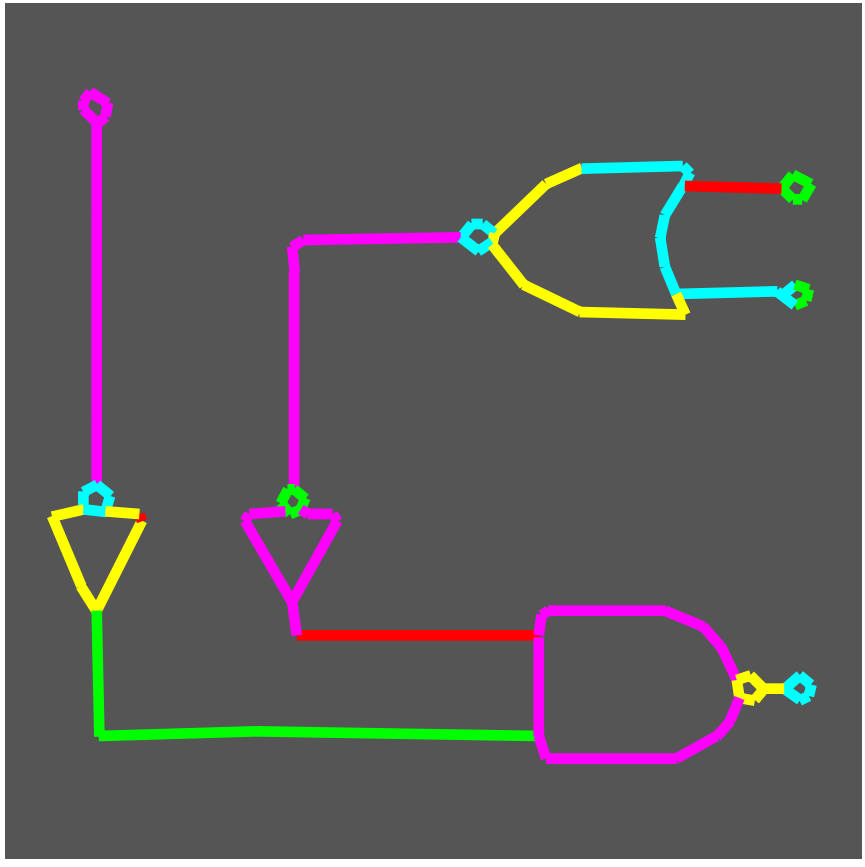
事例集の各要素について$\{A(C)\}$を求め，その集合から確率密度関数を推定

（$Pr[\{A(p)\},\{A(o)\}]$ 定数なので無視）

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，ひとまとまりのクラスタである分割
- メジアンの結果
- {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

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

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$Pr[\{a(C)\}|\pi=\pi^*]$

Pr[{a(C)}|π=π*]の計算

Pr[{a(C)}|π=π*]の計算

- {a(C)}の要素は，Θをパラメータとする分布に従い独立に発生
- パラメータΘは，Hを超パラメータとする分布に従って発生

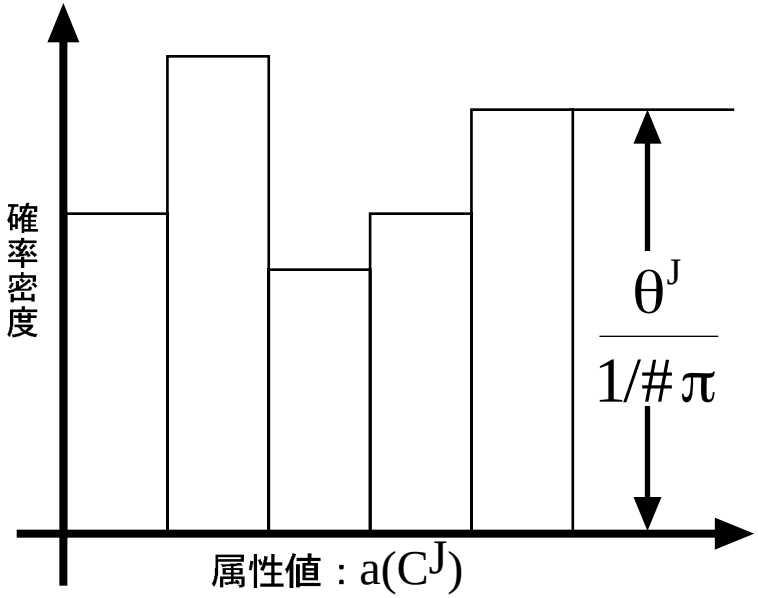
$$\Pr\left[\{a(C)\}\middle|\pi=\pi^*\right]=\left(\prod_{J=1}^{\#\pi}\Pr\left[a\left(C^J\right)\middle|\pi=\pi^*;\Theta\right]\right)\Pr[\Theta;H]$$

学習事例EX中の分類対象のクラスタ属性の属性値から，超パラメータをEMアルゴリズムによって計算

確率密度の族

$$\Pr\left[a\left(C^J\right)\middle|\pi=\pi^*;\Theta\right]$$

ヒストグラム型の分布



[0,1]の区間を，#π個に等分したヒストグラム

- 比較的少数のパラメータで，多様な形状の分布を扱える
- 密度が無限大などにならない

※ θ_Jの値の和は1

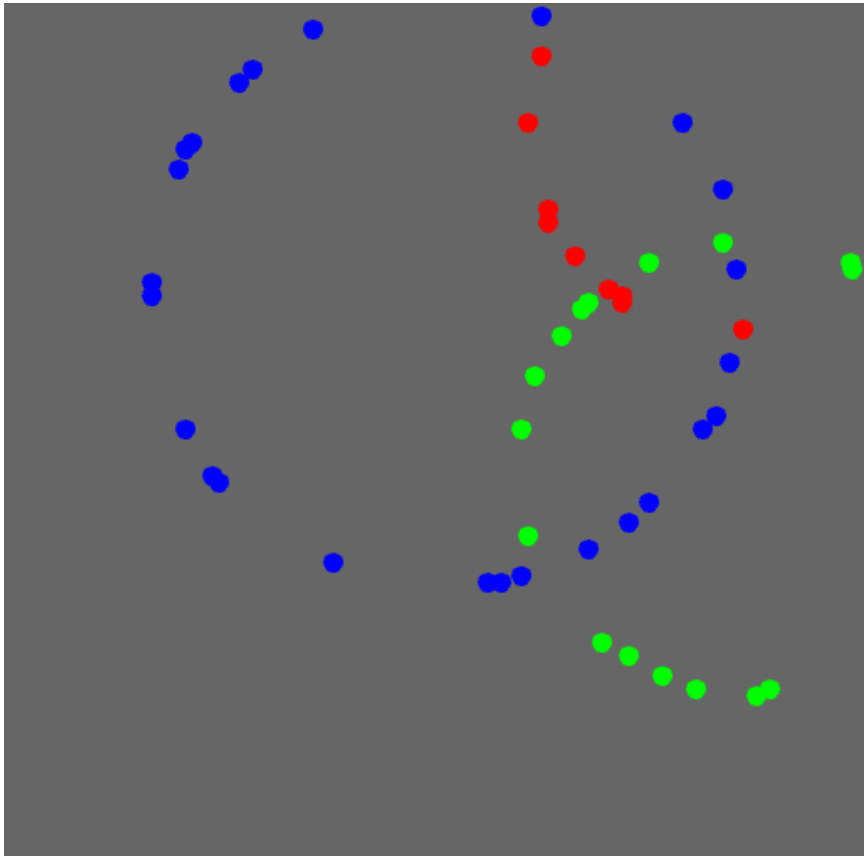
$$\Pr[\Theta;H]$$

多変量β分布（Dirichlet分布）

- 和が1となる値を表す分布
- 指数族分布はEMを適用する際に便利

実験対象

クラスタ属性の有効性を検証するための実験対象



ドットパターン

- 同じクラスタの点が円周上に並ぶ
- 中心方向に若干のガウスノイズ
- 分類対象数は50

クラスタ属性の有効性の検証実験の結果

1000事例でのleave-one-outの交叉確認による実験

	情報損失量 分割の間の類似性の尺度 0から1の範囲で0が最も良い
クラスタ属性なし	0.860
クラスタ属性あり	0.839

差のt検定：t-値 = 7.51

危険率1%での有意水準 = 2.33

クラスタ属性の導入：有意に情報損失量は減少

➡ クラスタ属性は分割用規則の獲得に有用

まとめと今後の予定

- クラスタ例からの学習へのクラスタ属性の導入
- クラスタ例の学習で，クラスタ属性を取り扱う方法を改良
- 人工データを対象にした実験により，クラスタ属性の導入によって，真の分割に近い推定が可能となることを示した
- 人工データだけでなく，より実問題に近いデータでの有効性を検証する予定