クラスタ例からの学習 —分類対象集合全体に関わる特徴の利用—

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
- 分類対象集合を分割する規則を，分類対象集合とその集合に対する適切な分割の組からなる学習事例から獲得することを目的とする

問題点

- 適切な分割とはかけはなれた分割しか獲得できなかった

解決法

- 分類対象全体に関わる特徴を利用できる新しいアルゴリズムを開発

目次

クラスタ例からの学習

- クラスタリングとは
- クラスタリングの問題点
- クラスタ例からの学習とは

クラスタ例からの学習の方法

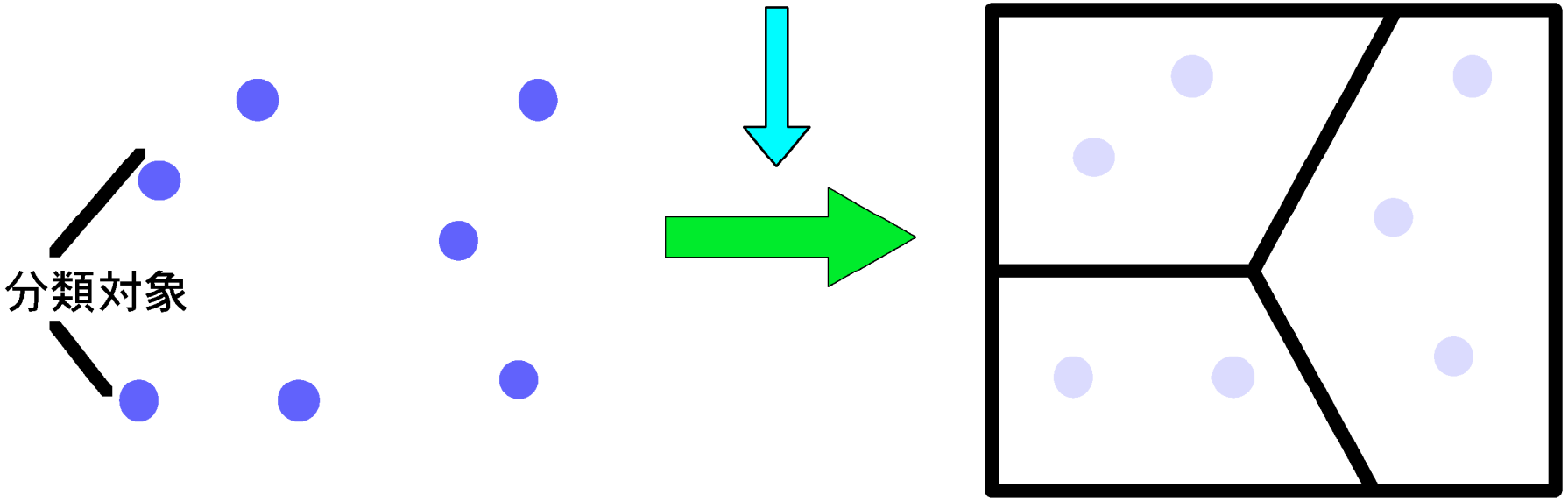
- 従来の方法
- 今回の方法（分類対象全体に関わる特徴を利用）

実験

- ドットパターンの分割問題への適用

クラスタリングとは

分類対象の類似性を判断するための規則



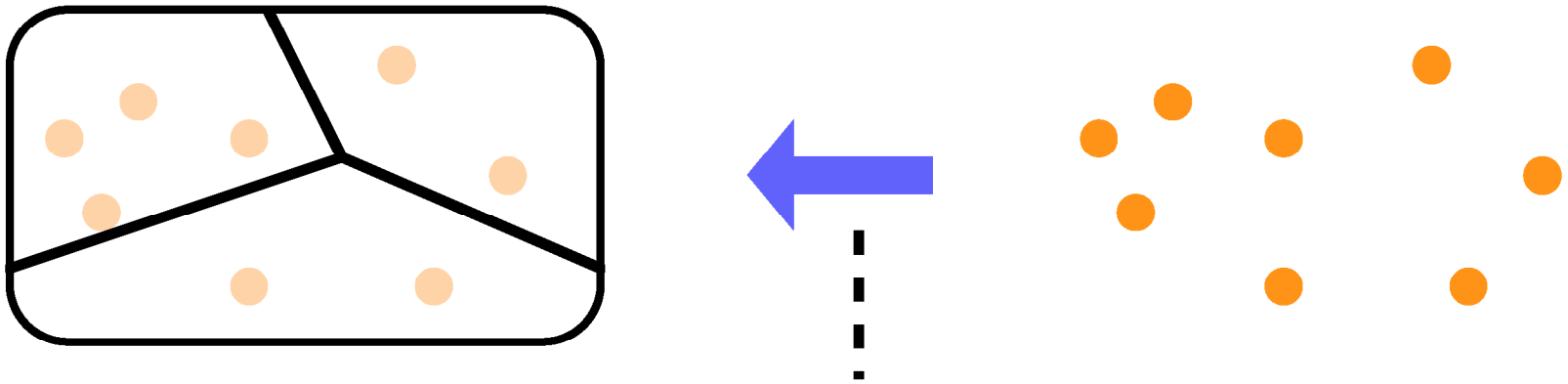
分類対象集合　　分割

事前に定めた規則をもとに，「類似したもの」を集めた部分集合(クラスタ）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

利用者が意図する分割

分類対象の集合



利用者は分割の規準は知らない

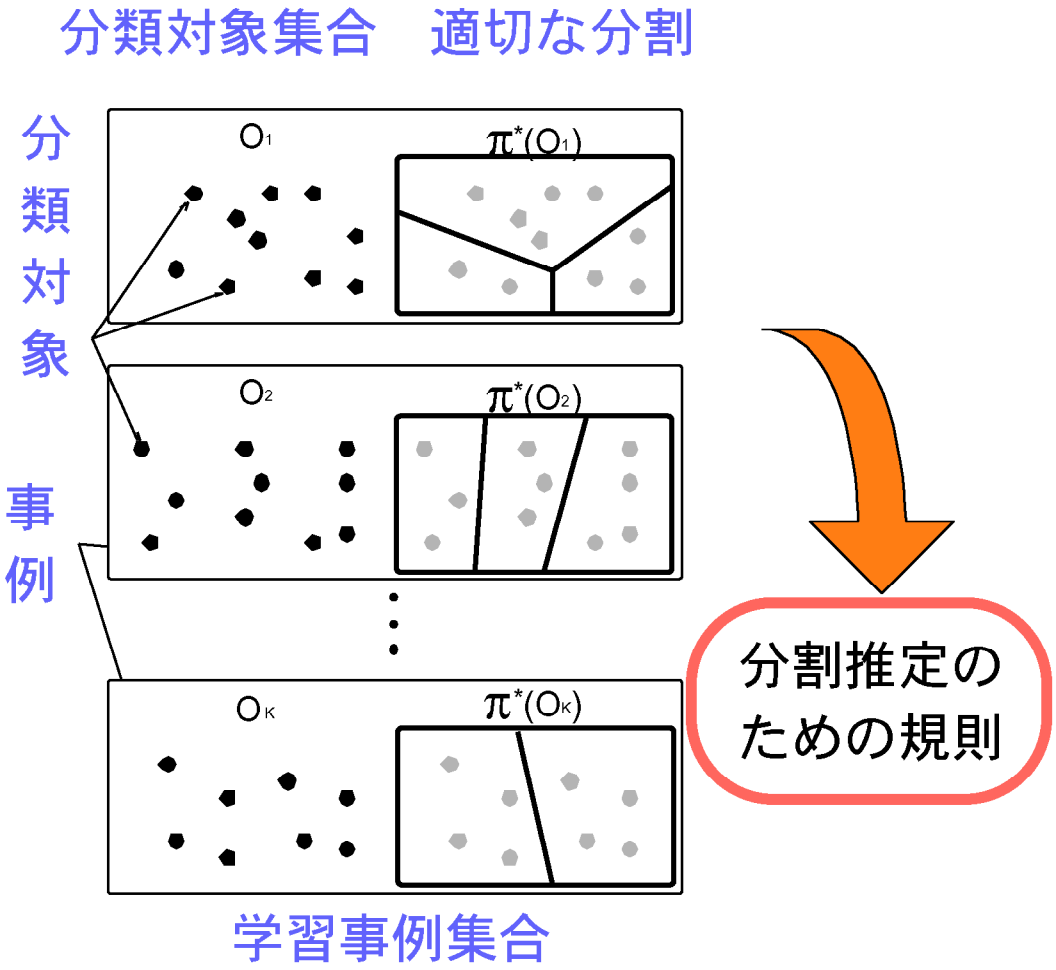
- 分割の規準・規則が事前にあるのではなく，利用者が意図する分割があり，それを自動的に導く手段としてクラスタリングを利用

クラスタリングの利用の問題点

- クラスタリング手法を利用するには，事前に分類対象の類似性を定める規則が必要
- この規則は，利用者の意図に則した適切な分割を導くように定める必要がある

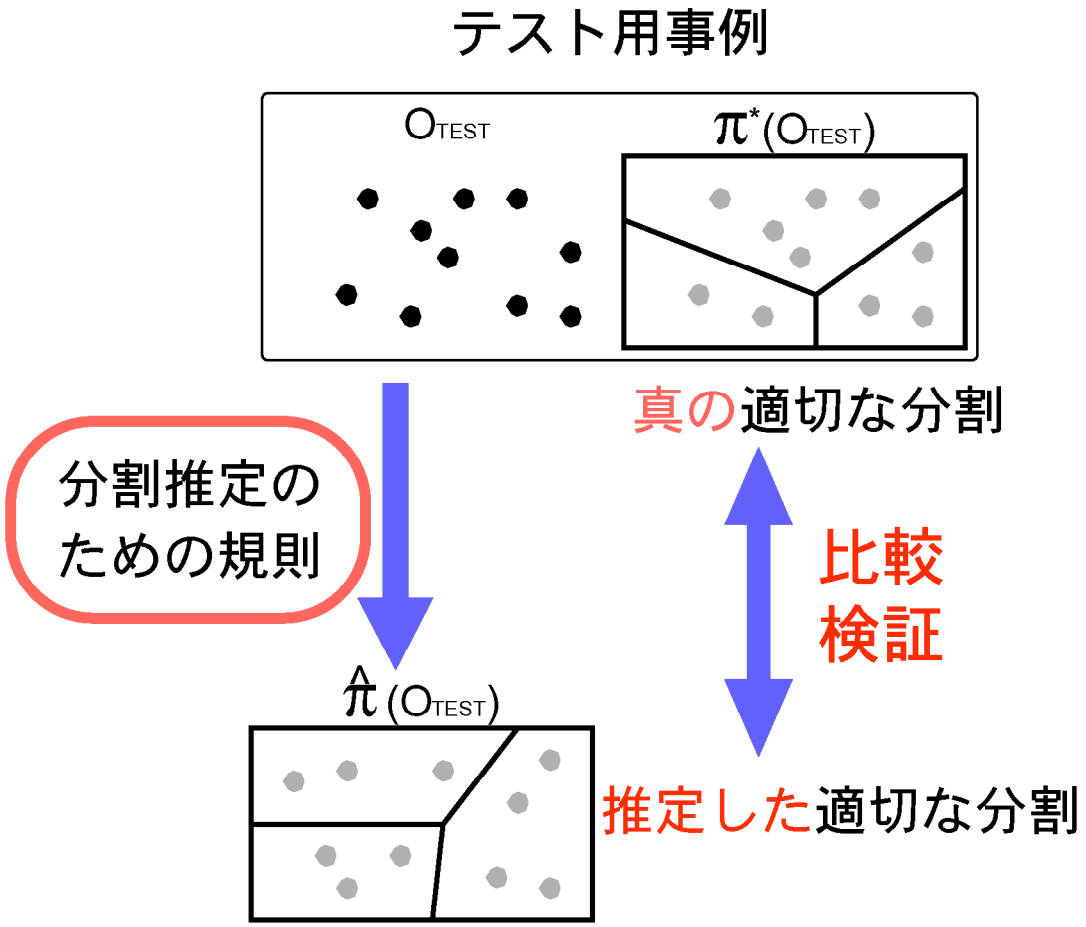
- クラスタリングの利用者が，その分割問題に対する知識をもとに，試行錯誤によって，この規則を定める
- 過去に与えられなかった，未知の分類対象集合を適切に分割できるか？

クラスタ例からの学習　(学習)



- 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　(分割の獲得と検証)



- 学習事例の要素ではない集合も分割できるかどうかを検証
- 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

前回の方法

分類対象の対を基本に分割を推定する手法

学習

- 分類対象集合の中の，任意の分類対象の対が，適切な分割で同じクラスタの要素になる確率 P1を推定するための規則を獲得

分割の推定

- 分割する集合の，すべての要素の対に対して確率P1を求める
- P1の大きな対が同じクラスタの要素となる分割を，代表的なクラスタリング手法 最大距離法，最小距離法，及び，群平均法により求める

前回の方法の問題点

分類対象の対を基本に分割を推定することに起因する問題

- 局所的な特徴をもとに分割を推定するので大域的な特徴を参照できない

分割の推定に既存のクラスタリング手法を利用することに起因する問題

- 最小距離法，最大距離法は，確率P1のわずかな推定誤りが，真の分割とは大きくかけ離れた分割を導く場合がある (ロバスト性の問題)
- 群平均法を用いた場合，分割中のクラスタ数を決定できない

今回の方法

次の確率を最大にする分割を適切な推定分割として採用

$$\Pr[\langle \text{集合全体の特徴} \rangle \mid \pi] \times \Pr[\pi \mid \langle \text{分類対象の対の特徴} \rangle]$$

- 〈集合全体の特徴〉… 分割中のクラスタ数の最大値など
- 〈分類対象の対の特徴〉… 分類対象間の距離など
- Pr[〈集合全体の特徴〉| π] …条件付き確率密度
- Pr[π|〈分類対象の対の特徴〉] …条件付き確率

ドットパターンの分割実験

クラスタの重なりの度合いが異なる三種類のデータを準備

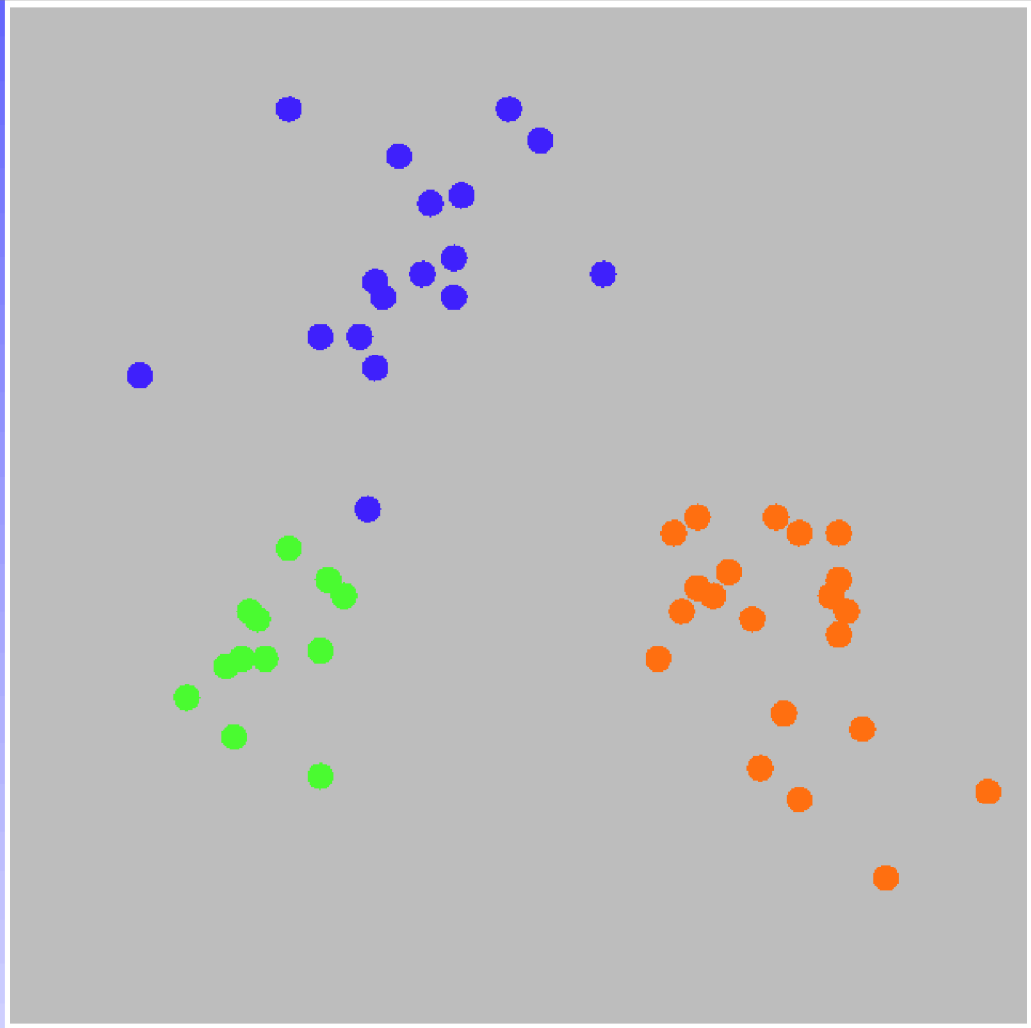
各データは100個の事例で構成

各事例はドットパターンとその適切な分割の組

各データに対し Leave-One-Out 試験を行った

- 最初の事例を取り除き，残りを学習事例集合とする
- 学習事例から，分割推定のための規則を獲得
- 取り除いたドットパターンの推定分割と，事例中の適切な分割の間の情報損失量を計算
- これらの手順を残り全ての事例すべてについて行う
- 情報損失量の平均を求める

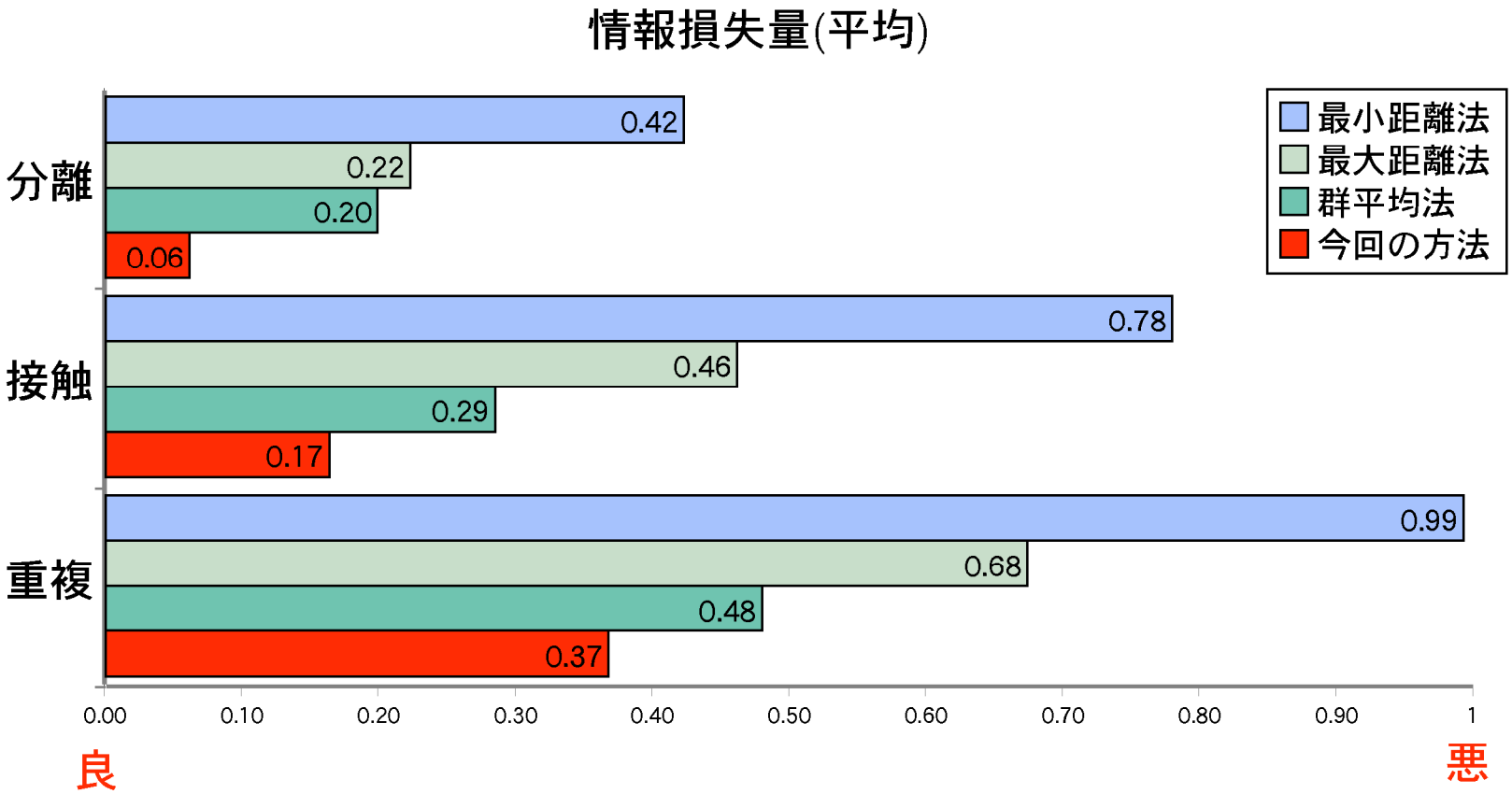
ドットパターンの分割の例



- ドットは２次のベクトル (X,Y)
- 適切な分割の同じクラスタの要素であるドットを同じ色で表示
- 同じクラスタ内のドットは，円形のガウス分布に従う
- ドットの位置の情報から，元のクラスタを推定する実験

ドットパターンの例

実験結果　(情報損失量の平均の比較)



実験結果(真に適切な分割が獲得できた割合)

各データ100回の試行のうち，どれくらいの割り合いで真に適切な分割を獲得できたか

	重複	接触	分離
最小距離法	0	11	42
最大距離法	0	9	41
群平均法	1	10	26
今回の方法	11	45	74

単位%

まとめ

結論

- 分割の具体例から，分類対象集合を分割する「クラスタ例からの学習」をドットパターンに対して適用する実験を行った
- 分類対象全体に関わる特徴を考慮することのできる，分割獲得アルゴリズムの開発により，従来よりも結果が改善された

今後の予定

- 「クラスタ例からの学習」をより多くの問題に適用する
- 分類対象全体だけではなく，個々のクラスタに関わる特徴を考慮した分割の獲得