パターン認識

早稲田大学講義 – 平成19年度

(独) 産業技術総合研究所　脳神経情報研究部門
栗田多喜夫、赤穂昭太郎

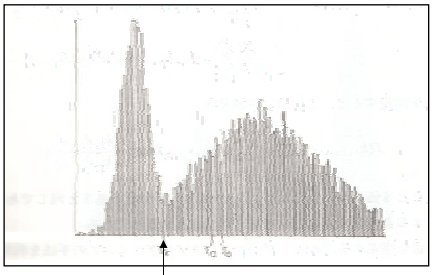
クラスター分析

クラスター分析

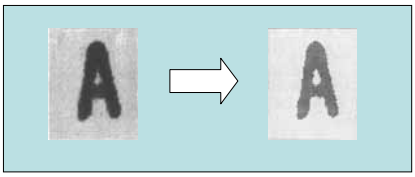
- クラスター分析（クラスタリング）
  - 複数の特性によって決定された個体間の類似性の指標をもとに、個体の集合をいくつかのグループに分類するための手法
  - 生物学、植物学、医学、社会科学、地球科学、政治経済学等で利用されている
  - 通信の分野では、ベクトル量子化と呼ばれ、情報の圧縮のための基本的な手法
- クラスタリング手法
  - 反復的クラスタリング
    - K-meansクラスタリング
    - Fuzzyクラスタリング
    - Vector Quantization
  - 階層的クラスタリング
    - 凝集的クラスタリング <= ヒープを利用した高速アルゴリズム[Kurita1991]
    - 分割的クラスタリング
  - オンラインクラスタリング
    - 競合学習
    - Self-Organizing Maps (SOM)
    - Learning Vector Quantization (LVQ)

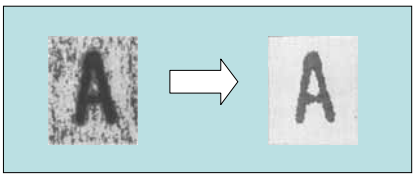
ヒストグラムの分割問題

- 濃淡画像を対象領域と背景に分離するためのしきい値選定に利用可能



しきい値





判別基準に基づく二値化(大津の二値化)

- 二値化のためのしきい値の選定
  - 濃淡画像を各画素の輝度値により対象領域と背景とに分離
- 判別基準
  - 輝度値をあるしきい値により分類したときの判別基準

$$\lambda = \frac{\sigma_B^2}{\sigma_W^2},\ \kappa = \frac{\sigma_T^2}{\sigma_W^2},\ \eta = \frac{\sigma_B^2}{\sigma_T^2} \quad \Rightarrow \quad \boxed{\text{最大化}}$$

  - 最小二乗基準との関係

$$\sigma_W^2 + \sigma_B^2 = \sigma_T^2 \quad \Rightarrow \quad \boxed{\sigma_W^2 \ \text{最小}}$$

  - ヒストグラムから簡単に計算可能

平均誤識別率に基づく二値化 (Kittlerの二値化)

- 平均誤識別率を最小とする基準
  - 対象領域の輝度値を背景の輝度値がともに正規分布に従うと仮定し、各画素をしきい値によって2つのクラスに分類したときの平均誤識別率から最適なしきい値を決定
- 平均誤識別率

$$P_e = 1 - \sum_{g=1}^{k} P(C_1 \mid g)p(g) - \sum_{g=k+1}^{L} P(C_1 \mid g)p(g)$$

$$\le 1 - \sum_{g=1}^{k} [1 + \log P(C_1 \mid g)]p(g) - \sum_{g=k+1}^{L} [1 + \log P(C_1 \mid g)]p(g)$$

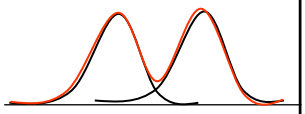
$$= \omega_1(k)\log\frac{\sigma_1(k)}{\omega_1(k)} + \omega_2(k)\log\frac{\sigma_2(k)}{\omega_2(k)} + \frac{1}{2}(1 + \log(2\pi)) + \sum_{g=1}^{L} p(g)\log(p(g))$$

- Kittlerの二値化基準

$$\boxed{J(k) = \omega_1(k)\log\frac{\sigma_1(k)}{\omega_1(k)} + \omega_2(k)\log\frac{\sigma_2(k)}{\omega_2(k)}}$$

最大尤度しきい値選定法

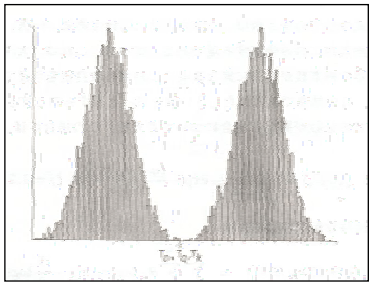
- 混合分布モデル (Population Mixture Model)
  - 画素iがどのクラスに属しているかの情報がθで与えられたとき、その画素が輝度値g_iを取る条件付確率分布

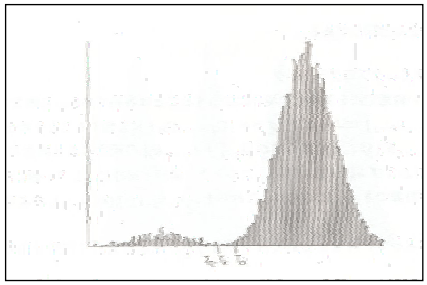


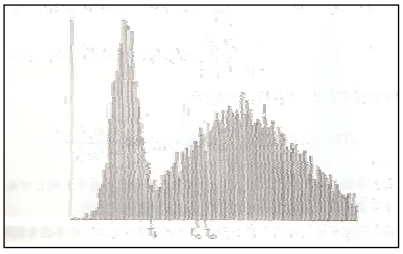
$$p(g_i \mid \boldsymbol{\theta}_i) = \theta_{1i}\, p(g_i \mid C_1) + \theta_{2i}\, p(g_i \mid C_2)$$

- 正規分布を仮定して、尤度を最大とするようなしきい値を選定
  - 各クラスの分布が平均は異なるが同じ分散を持つと仮定して、条件付分布の尤度を最大化 ＝＞ 大津のしきい値選定法
  - 異なる分散を持つと仮定して、同時分布の尤度を最大化 ＝＞ Kittlerのしきい値選定法
  - 各クラスの分布が平均は異なるが同じ分散を持つと仮定して、同時分布の尤度を最大化 ＝＞ 栗田のしきい値選定法

しきい値選定結果

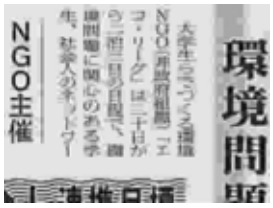






BTC (Block Truncation Coding)による画像圧縮への応用
～JPEGとの比較～

JPEG



(3.65KB)

(4.29KB)

(2.38KB)

提案方式

(3.32KB)

(4.14KB)

(2.24KB)

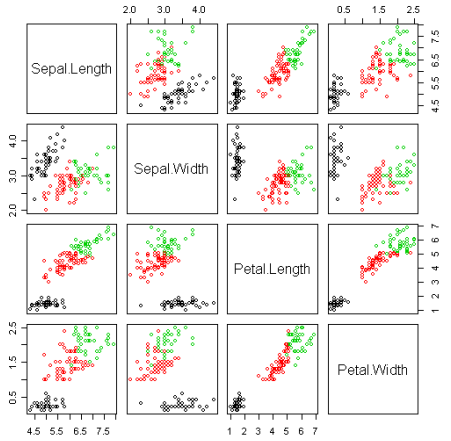
k-means法

- 反復的クラスタリング
  - データを繰り返し参照して、K個のクラスターに自動分類

- アルゴリズム
  - 初期化:
    - クラスター数Kを決定
    - K個のクラスター代表ベクトルを選択
  - 繰り返し
    - N個のサンプルを最も近いクラスター代表ベクトルのクラスターに分類
    - 分類されたサンプルを用いてクラスター平均ベクトルを計算し、それを新しいクラスター代表ベクトルとする
  - 終了
    - 終了条件の適合したら、クラスター代表ベクトルを返す

アヤメのデータのk-means法でのクラスタリング

- k-means法 – クラス代表ベクトルへの近さでデータを分類することを繰り返すクラスタリング手法



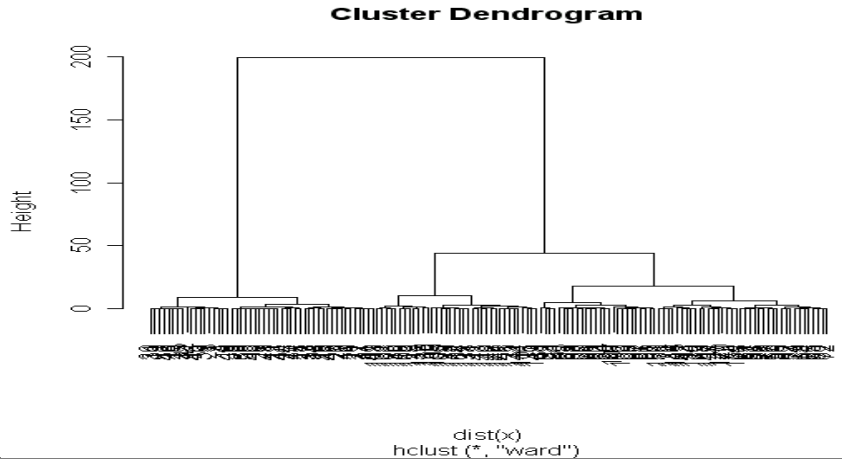
階層的クラスタリング

- 階層的クラスタリング
  - N個のサンプルに対して、サンプル1個のみからなるNクラスターからはじめて、近いクラスター対を順次統合するクラスタリング手法

- アルゴリズム
  - 初期化
    - サンプル1個のみからなるN個のクラスターを作成し、
    - 各クラスターに含まれるサンプル数、および平均ベクトルを、それぞれ、1個、サンプルの特徴ベクトルそのものとする
  - 繰り返し
    - 最も近いクラスター対を見つける
    - そのクラスター対を統合して、新しいクラスターを作成し、そのクラスターのサンプル数、平均ベクトルを計算
  - 終了
    - クラスター数が1個になれば終了

アヤメのデータの階層的クラスタリング

- Ward法 ＝ クラス平均とデータとの平均2乗誤差を最小とするようなクラス対を統合する手法



ベクトル量子化

- ベクトル量子化
  - データ集合をK個のクラスターに分割し、各クラスターをK個の代表ベクトル（コードブック）で近似
- 最適なコードブック
  - ベクトルデータ集合 $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
  - 平均2乗誤差最小の意味で最適な代表ベクトル $\mathbf{x}_l = \frac{1}{N_l} \sum_{i \in C_l} \mathbf{x}_i$
- ベクトル量子化のアルゴリズム
  - Linde等が提案したK-means法に基づくアルゴリズムが有名（LBGアルゴリズム）

One-Pass ベクトル量子化法

- データ集合を1度参照するだけである程度良いベクトル量子化を実現
- 考え方
  - 新たなデータに対して、現在のクラスタリングの結果を修正する
  - 修正法1
    - 既存のクラスターに新しいデータを追加

$$\Delta S = \frac{1}{N_l + 1}(\mathbf{x} - \bar{\mathbf{x}}_l)^T(\mathbf{x} - \bar{\mathbf{x}}_l)$$

  - 修正法2
    - 既存の2つのクラスターを統合し、新たなデータのみのクラスターを作成

$$\Delta S = \frac{N_p N_q}{N_p + N_q}(\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_q)^T(\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_q)$$

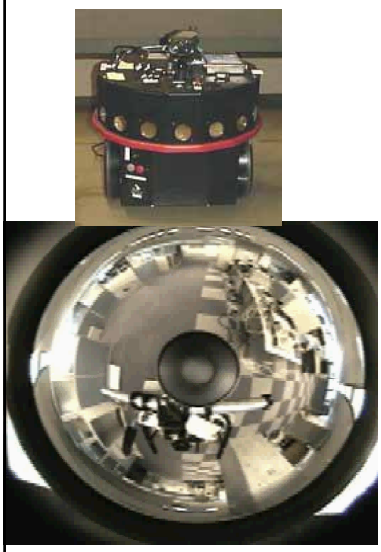
画像データの圧縮への応用

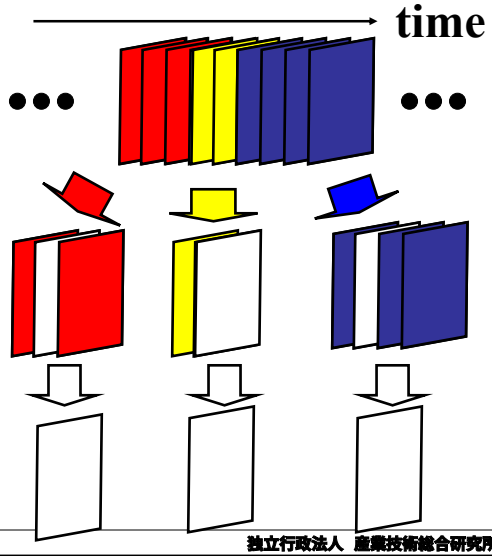
- M/RVQ (Mean / Residual Vector Quantizer)
  - 画像を小ブロックに分割し、各ブロックをその平均値と各画素の輝度値の平均からのズレのベクトルに分けて、ズレのベクトルをベクトル量子化する

元画像 (256x256画素、8ビット／画素)

圧縮画像 (256x256画素、1.5ビット／画素)

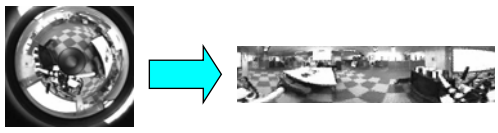
移動ロボットに搭載した全方位カメラ画像系列の分割



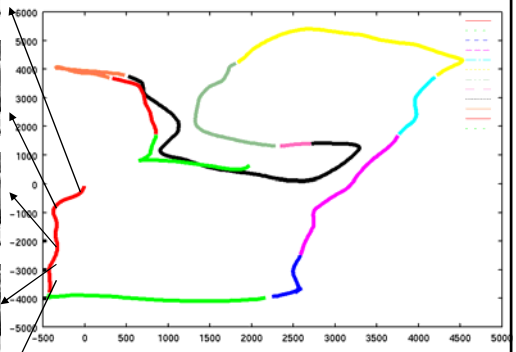


向きの変化に不変な特徴ベクトルの抽出

1. 画像の展開
   - 扱い易さ
2. ソーベルフィルタ
   - 柱などの形状を強調した特徴
3. 周囲方向への自己相関特徴
   - 向きの変化に不変な特徴
4. 主成分分析
   - 次元数の圧縮

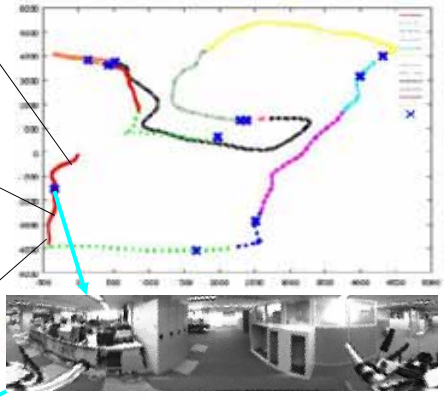


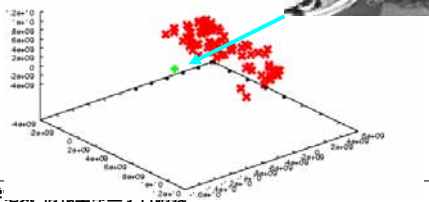
画像列のクラスタリング結果



画像　256×246
画像(展開)　359×79
クラスタ　759→12

情報量による代表画像の選択結果



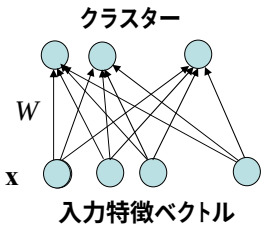


代表画像

競合学習 (Competitive Learning)

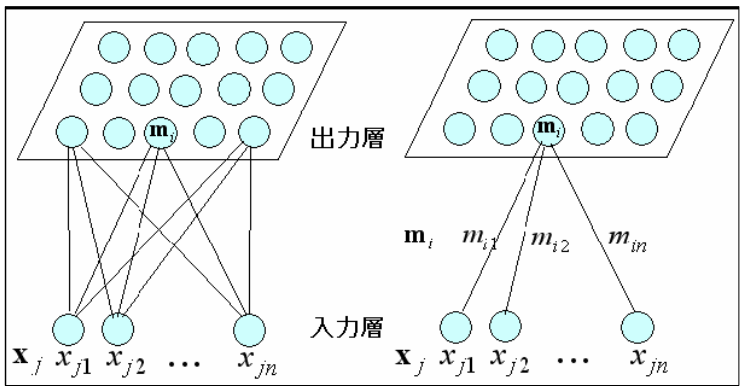
- オンラインクラスタリング
  - 時々刻々と与えられるサンプルに適応的にクラスター構造を変更

- アルゴリズム
  - 初期化
    - 重みの初期化
  - 繰り返し
    - 与えられた入力特徴ベクトルに対して最大の出力を与えるクラスター (勝者) を探す
    - そのクラスターに対応する重みを入力ベクトルの方向にちょっとだけ動かす
    - 重みを正規化

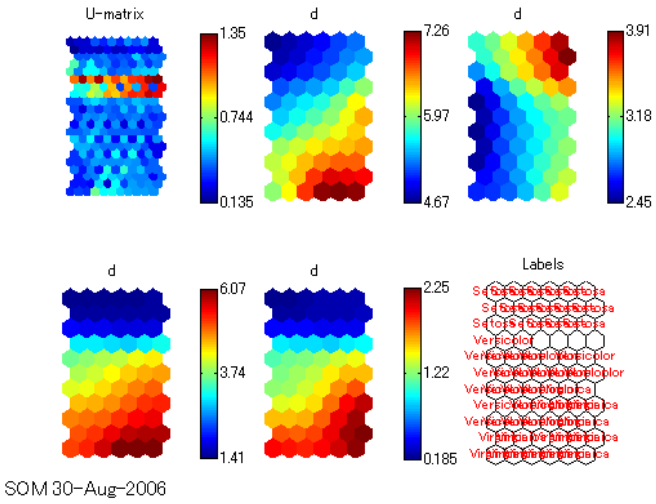


自己組織化マップ (SOM)

- 近傍との隣接関係を維持して、高次元のデータを低次元 (2次元) に非線形に写像するデータ解析法



アヤメのデータに対するSOMの適用例



ここまでのまとめ

- **ベイズ識別の理論**
  - 事後確率が重要。事後確率最大のクラスに識別すればよい。
  - そのためには、確率密度関数の推定が必要。
    - パラメトリックモデル、ノンパラメトリックな手法、セミパラメトリック
  - 各クラスの条件付確率が正規分布の場合
    - 事後確率の対数をとると、特徴量に関して2次の関数（識別関数）
    - クラスの分散共分散行列が等しい場合には、1次（線形）の識別関数
- **線形識別関数の学習**
  - 訓練データから直接識別関数のパラメータを求める
    - パーセプトロン、最小2乗判別関数の学習、ロジスティック回帰
    - 多層パーセプトロン
- **汎化性能**
  - 訓練データに対する識別性能ではなく、未学習データに対する性能が重要
    - 汎化性能の評価（Cross-Validation、ブートストラップ、情報量基準）
    - 汎化性能の向上（Shrinkage法、ノイズの付加）
    - 変数選択
- **統計的特長抽出**
  - 最小2乗判別関数、主成分分析、判別分析
- **クラスター分析**
  - クラスター分析、ベクトル量子化、自己組織化マップ

質問等

- **電子メール**
  takio-kurita@aist.go.jp

- **ホームページ**
  http://staff.aist.go.jp/takio-kurita/index-j.html

- **連絡先**
  〒305－8568
  茨城県つくば市梅園1-1-1 **つくば中央第**2
  産業技術総合研究所 脳神経情報研究部門
  栗田 多喜夫

- **電話**･FAX
  電話 029-861-5838 FAX 029-861-5842

6限目終了