楕円領域の解析による特徴量の選択

5M－10

阿部重夫
日立製作所　日立研究所

Ruck Thawonmas
理化学研究所　知能実現機能研究チーム

1．はじめに

パターン認識の分野では最適な特徴量の決定は、最小規模で最良の認識系を実現するうえで重要な課題である.特徴量を決定する方法としては元の入力を低次元の特徴量に変換する特徴抽出と元の入力から必要な特徴量を選択する特徴選択とがある.我々は文献[1]で領域を超直方体で近似してクラス間の重なりの度合いを定量化する指標により不要な特徴量を削除する方法を提案した.本論文では領域を楕円で近似して特徴量を選択する方式を述べる.

2．楕円領域の近似

m次元のデータ$\mathbf{x}$をn個のクラスに分離するとし，各々に属する教師データがあるとする.ここでクラスiの中心 $\mathbf{c}_i=(c_{i1},\ldots, c_{im})^t$ を

$$c_{ik}=\frac{1}{N_i}\sum_{\mathbf{x}\in \text{class } i} x_k \tag{1}$$

で求める．このときtは行列の転置を示しN_iはクラスiに属するデータ数とする．さらに$m \times m$共分散行列を

$$Q_i=\frac{1}{N_i}\sum_{\mathbf{x}\in \text{class } i}(\mathbf{x}-\mathbf{c}_i)(\mathbf{x}-\mathbf{c}_i)^t \tag{2}$$

により求める．クラスタの中心$\mathbf{c}_i$から入力$\mathbf{x}$への重みつき距離を

$$d_i^2(\mathbf{x})=(\mathbf{x}-\mathbf{c}_i)^t Q_i^{-1}(\mathbf{x}-\mathbf{c}_i) \tag{3}$$

で定義する．ここでQ_iを正則とするとQ_iは正定行列となるから平均の重みつき距離は

$$\frac{1}{N_i}\sum_{\mathbf{x}\in \text{class } i} d_i^2(\mathbf{x})=m \tag{4}$$

となることが分かる.
次にクラスiに属する入力$\mathbf{x}$の成立度を

$$m_i(\mathbf{x})=\exp\left(-\frac{1}{m}d_i^2(\mathbf{x})\right) \tag{5}$$

で定義する．ここで入力数が変わったときに成立度が比較できるように重みつき距離をmで割る.

ここで

$$m_i(\mathbf{x})>m_j(\mathbf{x}),\quad j=1,\ldots,n, j\neq i \tag{6}$$

が成立するときに入力$\mathbf{x}$がクラスiに属するとすると，クラスのデータを分割せず，また学習を行わないときの文献[2]のクラシファイアに一致する.

3．領域の複雑度

クラスiの領域のクラスjの領域に対する重なりの度合$o_{ij}(F)$を

$$o_{ij}(F)=\frac{\sum_{\mathbf{x}\in \text{class } i} m_j(\mathbf{x})}{\sum_{\mathbf{x}\in \text{class } i} m_i(\mathbf{x})} \tag{7}$$

と定義する．ここでFは入力特徴量の集合である.分母はクラスiに属する教師データのクラスiの成立度の和で，分子はクラスiに属する教師データのクラスjの成立度の和で，領域の重なりをはかる指

Input Feature Selection by Analysis of Ellipsoidal Regions

Shigeo Abe1 and Ruck Thawonmas2, 1:Hitachi, Ltd., 2: RIKEN

標と考えることができる．しかしながら，クラス i のデータがクラス j に誤認識しなければ，(7)式の値が高くても分離が難しいと考える必要はない．これを反映するために，入力領域の複雑度を

$$O(F) = \sum_{i,j=1}^{n} p_{ij}\, o_{ij}(F) \tag{8}$$

で定義する．ここで

p_{ij} =(クラス j に間違ったクラス i のデータ数)
／(学習データの数)

である．

(8) 式の複雑度は文献[1] の複雑度と似た定義になっているが，大きな違いは本論文では各クラスの領域を一つの楕円で近似しているのに対して文献[1] では入れ子構造の超直方体で近似していることである．

4． 特徴選択

m次元の入力の番号 1,..., m を要素として含む集合を Iとして，m個全ての入力を用いたときの複雑度を求めこれを $O(I)$とする．次に集合 Iを集合 Fに設定して，添字 i を 1 に初期化する．集合 Fから i 番目の変数を削除した集合を $F(i)$として，入力のうち集合 $F(i)$に対応する複雑度を計算し，$O(F(1))$,..., $O(F(|F|))$の最小値を求めそれを $O(F(k))$とする．ただし|・|は集合の要素数を示す．最小値をとるのは複雑度が最も増加しない入力を削除するためである．ここで

$$|O(I) - O(F(k))| / O(I) \geq \alpha \tag{9}$$

$$|O(F) - O(F(k))| / O(I) \geq \beta \tag{10}$$

のどちらかが成立したときに入力変数の選択を終了する． このとき，α, β は入力の削除処理を打ち切るための正の数である．(9)式の左辺は全ての入力変数を使った場合からどの程度複雑度が相対的に増加したかを示し，(10)式は複雑度の変化率を求めるものである．

5． 特徴量削減の効果

あやめのデータ，数字認識データ，サイロイドデータ，血球データで，方式の評価を行った．図１に数字認識の場合を示す．認識率はテストデータに対するもので，文献[1]の方法の場合に比べて，ロバストな選択が行なわれていることが分かる．

４種のデータで文献[1]の方法と同数以上の入力変数を削除できることが確かめられた．特に血球データのようにデータの分布が入力軸に平行でない場合には，文献[1]よりもよい結果がえられた．また，サイロイドデータのように，データがガウス分布に従わない場合でも汎化能力を低下することなく特徴量が削減できた．詳細は講演時に示す．

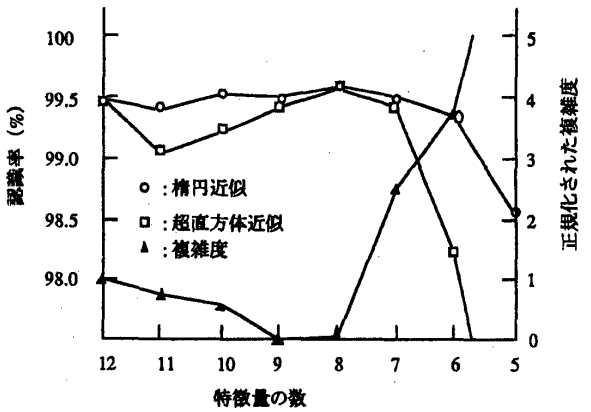

図１ 数字認識における特徴量の選択の比較

6． 結言

クラスの存在領域を楕円で近似して領域の重なりの複雑度を定量化して，複雑度が低下しない範囲で特徴量を選択する方式を述べた．

参考文献

[1] R. Thawonmas and S. Abe, "A Novel Approach to Feature Selection Based on Analysis of Fuzzy Regions," *IEEE Trans. Systems, Man, and Cybernetics-Part B*, vol. 27, no. 2, April 1997.

[2] S. Abe and R. Thawonmas, "Fast Training of a Fuzzy Classifier with Ellipsoidal Regions," *Fifth IEEE International Conference on Fuzzy Systems*, September 1996.