識別問題におけるMDL基準を用いたクラスタリング法

4Q−1

天元 宏　　工藤峰一　　新保 勝

北海道大学大学院 工学研究科 システム情報工学専攻

1 はじめに

パターン認識において、サンプルの分布に正規分布などの単純なパラメトリックモデルを仮定できない場合、k 最近隣法や区分的線形識別規則などのノンパラメトリックな識別規則が有効である。

これらの識別規則では、訓練サンプル集合に対する過学習を回避するためや計算コストの削減のために、訓練サンプル集合上の各クラスターの重心である代表点を訓練サンプル集合の代用とすることが多い。ほとんどの応用例では、このクラスター数を主観による評価で発見的に決定し、識別問題に利用している。しかし、それでは必ずしも最適な識別性能を発揮できているとは限らず、客観的な評価基準によるクラスター数の推定法が必要となっている。

この問題に対し、文献 [1] では、サンプルの分布に混合正規分布モデルを仮定し、その特徴ベクトルの発生確率に関する尤度と、分布を記述するパラメータ数のトレードオフを MDL 基準 [2] により評価し、最適なクラスター数の推定を試みている。しかし、この方法は教師なし学習であるため、これを識別問題に適用する場合には各クラス毎に尤度を単独で評価する必要があり、必ずしも識別に最適なクラスター数は求められない。

そこで本研究では、混合正規分布モデルにおいて、文献 [3] の「確率的規則の学習問題」の枠組に基づき、MDL 基準を利用した、識別を前提とした最適なクラスター数の推定を試みる。

2 混合正規分布モデル

本研究では、特徴ベクトル $\mathbf{x} \in R^d$ の分布に、次の分割数 K の混合正規分布モデルを仮定する。

$$f(\mathbf{x}|p,\theta) = \sum_{k=1}^{K} p_k \cdot g(\mathbf{x}|\theta_k) \tag{1}$$

$$p = (p_1, \ldots, p_k, \ldots, p_K)$$

$$\theta = (\theta_1, \ldots, \theta_k, \ldots, \theta_K)$$

ここで、p_k は混合比で、$\sum_{k=1}^{K} p_k = 1$ を満たす分岐確率である。また、θ_k は k 番めの d 次元正規分布 $g(\cdot)$ を記述するパラメータ (平均ベクトル, 共分散行列) である。

この分布に対し、本研究では、EM アルゴリズム [4] によりそのパラメータを推定する。EM アルゴリズムでは、各訓練サンプル $\mathbf{x}_i$ に各クラスターへの帰属度を表すパラメータ z_{ik} ($\sum_{k=1}^{K} z_{ik} = 1$) を考える。ここで、$k = \arg\max_k z_{ik}$ をそのサンプルの所属するクラスターと判定することで、一つのクラスタリング結果を得る。

式 (1) の確率密度関数とベイズ規則により、特徴ベクトル $\mathbf{x}$ でクラスラベル $y = c$ の発生する確率は次式で与えられる。

$$P(y = c|\mathbf{x}) = \frac{f(\mathbf{x}|p_c, \theta_c)}{\sum_{c'=1}^{C} f(\mathbf{x}|p_{c'}, \theta_{c'})} \tag{2}$$

ここで、C はクラス数、p_c, θ_c は c 番目のクラスの混合パラメータである。

なお、本研究では簡単のため、全クラスに同じ分割数 (クラスター数)K を与え、K の最適値を求める。

3 MDL 基準による最適なクラスター数の推定

前節の混合正規分布モデルでは、クラスター数 K を大きくすると、訓練サンプル集合に対する識別性能は向上する。しかし、これは一般に、未知サンプルに対する高い識別性能を意味しない。

そこで本研究では、MDL 基準 [2] を用いて、訓練サンプル集合に対する識別性能と、混合正規分布モデルを記述するパラメータの数のトレードオフを評価することにより、未知サンプルに対して高い識別性能を発揮するクラスター数 K の推定を試みる。

長さ N の訓練サンプル列を X^N、各訓練サンプルの特徴ベクトルを $\mathbf{x}_i$、そのクラスラベルを y_i とすると、式 (2) の確率分布 P を用いて、次式のように Kraft の不等式を満たすビット長で評価できる [3]。

$$L = -\sum_{i=1}^{N} \log_2 P(y_i|\mathbf{x}_i) + L(P) \tag{3}$$

ここで、右辺第一項は訓練サンプル列 X^N に対するクラスラベルに関する P の尤度の評価である。また、第二

A Clustering Method in Classification Problems Using the MDL criterion
Hiroshi Tenmoto, Mineichi Kudo and Masaru Shimbo
tenmo@huie.hokudai.ac.jp
Division of Systems and Information Engineering,
Graduate School of Engineering,
Hokkaido University, Sapporo 060, Japan.

項 $L(P)$ は確率分布 P の記述パラメータ数の評価であり、本研究では次式で定義する。

$$L(P) = \frac{m}{2}(\log_2 N + \log_2 e)$$

ここで、m は次式で求めるパラメータ数、e は自然対数の底である。

$$m = C\{(K-1) + 2K\} \quad (4)$$

式4において、それぞれ $(K-1)$ は p の、$2K$ は θ のパラメータ数である。

以上より、式(3)を最小とする K を最適なクラスター数の推定値として決定する。

4　実験と考察

2次元2クラスの弧状に入り組んだ人工データに対して計算機実験を行った。訓練サンプル数は各クラス100個、検査サンプル数は各クラス1000個である。図1にクラスター数 K を増加させた場合の記述長の変化、図2に識別率の変化を示す。図1より K の推定値は2となる。これに対し、図2より実際の最適値は4である。また、$K=2$ と $K=4$ の場合の混合正規分布を図3と図4に示す。

K の値を実際に識別率が最大となる値よりも小さく推定してしまったものの、これはMDL基準の一般的な傾向と考えられ、訓練サンプル数が多くなるに従って最適値に一致すると予想できる(MDL基準の一致性)。

5　おわりに

混合正規分布モデルを用いたクラスタリングにおいて、MDL基準を利用して、識別を前提とした最適なクラスター数の推定を試みた。実験により、訓練サンプルへの過学習を回避し、未知サンプルに対する高い識別性能を期待できるクラスター数を推定できることを確認した。

今後は、他のクラスのサンプルの発生のしにくさを最尤法で積極的に評価する方法について、EMアルゴリズムを用いて実現したい。

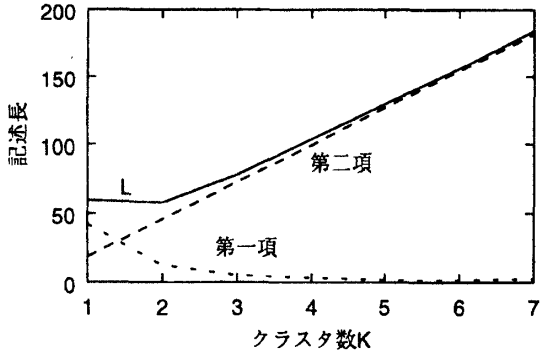

図 1: クラスター数を増加させた場合の L の変化

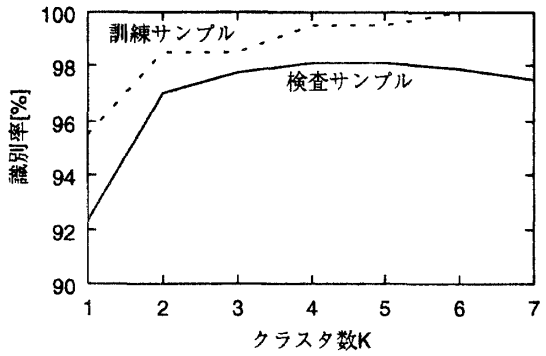

図 2: クラスター数を増加させた場合の識別率の変化

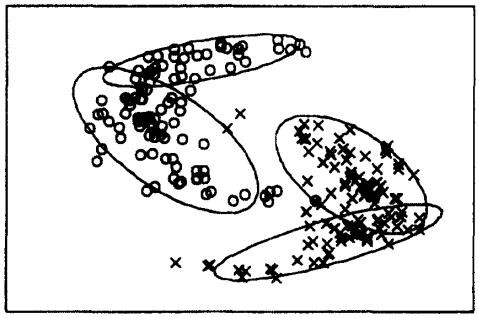

図 3: $K=2$(推定値) の場合の混合正規分布 (標準偏差の2倍の楕円)

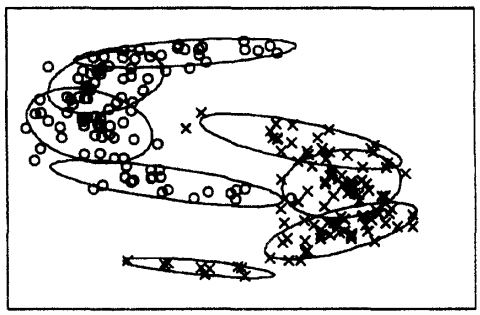

図 4: $K=4$(実際の最適値) の場合の混合正規分布 (標準偏差の2倍の楕円)

文献

[1] 市村直幸, クラスタ数推定のための最ゆう法に基づくロバストクラスタリング. 電子情報通信学会論文誌, D-II, J78, 8(1995), 1184-1195.

[2] J. Rissanen, A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics*, **11**(1983), 416-431.

[3] K. Yamanishi, A Learning Criterion for Stochastic Rules. To appear in *Machine Learning*, An extended abstract in *Proceedings of the Third Annual Workshop on Computational Learning Theory*, 1990, 67-81.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, **39**(1977), 1-38.