7C-6 赤池統計量基準による自動クラスタリング最適化の一手法

張 紹星
(筑波大学)

1．はじめに

クラスタリング分類法は、パターンに関してどのようなクラスが存在するのか分からないときに使われるパターン認識分類法であり、リモートセンシング画像処理、医学用画像処理などに適用される。従来の手法では、パターン間の類似度の閾値、あるいはクラスタ数などのパラメータを指定する必要があった。すなわち、パラメータの定義の仕方によって、同じデータセットに対しても、異なった結果が得られ、また、分類結果を人為的に選択しなければならなかった。

本稿では、赤池情報量基準AICを用いた自動クラスタリング最適化手法の処理概要、および画像処理結果について報告する。

2．自動クラスタリング最適化手法

クラスタリング分類法のパラメータの選択の任意性問題に対して、処理されたデータセットから、数学的な基準によりパラメータを自動的に抽出することを考えた。画像処理のような大容量データセットに適する非階層的なクラスタリング分類法について、クラスタ数パラメータの可能な範囲、およびこれらのクラスタ数の最適分類結果を赤池情報量基準によって決定するアルゴリズムを筆者らは提案した。

2．1 赤池情報量基準

情報量基準 AICは、最尤法で当てはめられたモデルが複数個あるときに、モデルの適応性を評価する基準である。

AIC＝-2（モデルの最大対数尤度
　　ーモデルの自由パラメータ数） (1)

がモデル選択の基準となる。AICを最小とするモデルが最適なモデルと考えられる[1,2]。

2．2 処理手法の概要

（1）最適ヒストグラムの作成　データの各次元の最適ヒストグラムを情報量基準AIC_{his}によって作成する。ヒストグラムの度数分布表を多項分布とみなすことにし、度数分布表の階級数をc、各階級の確率をp(i)、観測度数をn(i)(i=1,...,c)とすると、多項分布関数は、

$$P(\{n(i)\}\mid\{p(i)\})=\{n!/\prod_{i=1}^{c} n(i)!\}\prod_{i=1}^{c} p(i)^{n(i)} \quad (2)$$

で与えられる。情報量基準AIC_{his}は、

$$AIC_{his}=-2[\log\{n!/\prod_{i=1}^{c}n(i)!\}+\sum_{i=1}^{c}n(i)\log\{n(i)/n\}]+2(c-1) \quad (3)$$

になる。最小のAIC_{his}をもつ階級数は最適なヒストグラムを作ると思われる。

（2）クラスタ数の範囲の決定　得られた各次元のヒストグラムのピーク数k_iに基づいて、クラスタ数の範囲NC_{min}、NC_{max}を決定する。

$$NC_{min}=\mathrm{Max}\{k_i\} \quad (4)$$

$$NC_{max}=\prod_{i=1}^{D}\{k_i\} \quad (Dは次元数) \quad (5)$$

（3）クラスタリング分類　最小クラスタ数NC_{min}から最大クラスタ数NC_{max}まで、K－means法というクラスタ分類法を用いてパターンを分類する。k－means法では、最初に、クラスタ数を指定し、初期分割を行う。そして、すべてのパターンを最短距離のクラスタへ分類する。この処理は各クラスタ間のメンバー交換がなくなるまで繰り返される。

（4）最適分類結果の評価　分類された各クラスタの分布は正規分布に従うことと仮定した上で、赤池情報量基準AIC_{nor}によって分類結果を評価する。正規分布の密度関数は、平均値μと標準偏差σ^2を自由パラメータとして、

$$f(x\mid\mu,\sigma^2)=(1/\sqrt{2\pi\sigma^2})\exp(-(x-\mu)^2/2\sigma^2) \quad (6)$$

で与えられる。したがって、

$$AIC_{nor}=n\log 2\pi+\sum_{i=1}^{K}n(i)\log\sigma_i^2+n+4k \quad (7)$$

となり，最小のAIC_{nor}をもつ分類結果が最適なクラスタリング分類結果と考えられる。

Automatical Optimization of Clustering Using Akaike Information Criterion

Zhang Shaoxing
Egineer Department, University of Tsukuba

３．アルゴリズム検証実験の結果

アルゴリズムを検証するために正規ランダムデータに適用したところ、良い結果が得られた[3]。ここでは、このアルゴリズムを衛星画像に適用することを試みた。この目的は土地被覆状況に対してアルゴリズムの適切性を調べることである。ここで使用した衛星画像データは１９７９年５月２１日に撮られたＬＡＮＤＳＡＴ・ＭＳＳ関東シーン（Path 115, Row 35）の東京都中心地域（512x512 pixels）である（図１）。このような大容量データセットでは最適化のための反復処理に膨大な時間がかかる。そこで、256x256pixels の地域を対象地域として抽出する。Ｂａｎｄ４、５、７のデータのヒストグラム（図２、図３、図４）から、可能なクラスタ数の範囲は NC_{min}=5、NC_{max}=20と判定された。土地被覆は通常10個以上のクラスタに分けられるので、最小クラスタ数は11とした。11個から20個までの分類結果に対するAIC_{nor}の値（表１）を見ると、クラスタ数が13になるとき、AIC_{nor}は最小になった。図５にＮＣ＝ 13 ときの分類結果を示す。これらのすべての処理に要したＣＰＵ時間は約40分（FACOM M780/20）であった。

４．おわりに

提案された自動クラスタリング最適化手法によって、入力されるデータセット自身から情報を抽出し、自動的に最適分類することができた。ただし、用いられる分類結果の評価基準は正規分布関数を適用した基準である。今後は、分類結果に従う分布関数を判明し、各種の分布と対応する評価基準も加えていく予定である。

参考文献

１）H. Akaike : A New Look at the Statistical Model Identification, IEEE. Trans. Autom. Contr., AC-19, pp.716-723, 1974.

２）坂元慶行、石黒真木夫、北川源四郎： 情報量統計学、共立出版株式会社、pp.42-43, 1982.

３）星　仰、張　紹星：Ｋ－ｍｅａｎｓクラスタ分類法における赤池情報量基準の適用、日本写真測量学会秋季学術講演会発表論文集、pp.37-42、1988.

図１．東京都中心地域のＭＳＳ画像

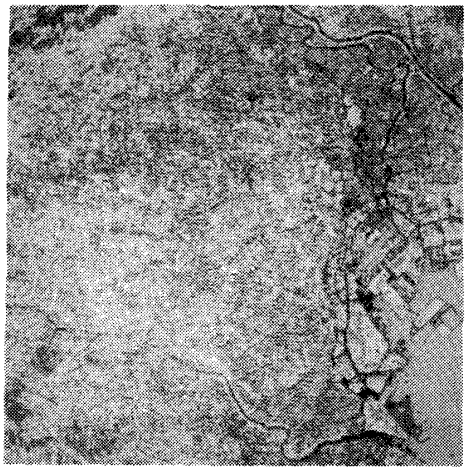

図５．最適分類結果（ＮＣ＝１３）

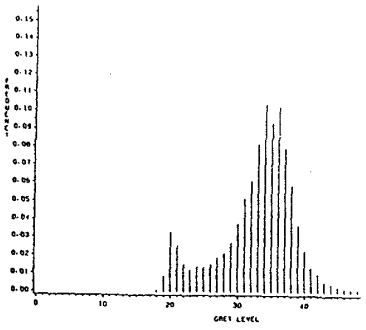

図２．バンド４のヒストグラム

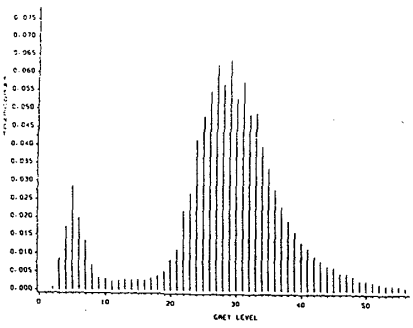

図３．バンド５のヒストグラム

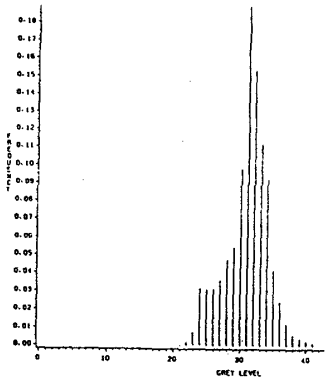

図４．バンド６のヒストグラム

表１．分類結果を評価するAIC_{nor}値

NC	11	12	13	14	15
AIC_{nor}	508235	509280	506172	508986	512600
NC	16	17	18	19	20
AIC_{nor}	522123	513747	523211	523502	523667