GAによる教師データの統計量の算出*

2E－1

星　仰　石黒　一哉†
茨城大学‡

1　はじめに

リモートセンシング画像データの利用でもっとも基本的、かつ広い用途を持つものに土地被覆分類図がある。これは、与えられた画像データを植生、水域、人工構造物などの類似データでまとめて一定の記号や色調で表現するものである。従来の統計的処理である教師付き分類では、画像データの中から各クラスの特徴を示すトレーニングデータを抽出し、その平均値、標準偏差、共分散行列などを用いてパターンの分類を行っている。このトレーニングデータの性状は正規分布と仮定しているが、これに反するときには平均値と極大値に差を生じる可能性が高い。ここで、極大値とは各クラスの頻度分布においてもっとも頻度の大きいデータ値のことである。

そこで、遺伝的アルゴリズム (Genetic Algorithms) を用い、トレーニングデータの極大値を求め、平均値と極大値の比較を行い、それぞれの値を用いてクラスタリングを行った。

2　GAによる極大点探索法

実験には、LANDSAT・TMによる衛星画像データのバンド４～７の４つのバンドを用いる。まず、衛星画像データのトレーニングデータを抽出する。このとき、各トレーニングデータのデータ量の差からの弊害をなくすため、トレーニングデータのデータサンプリング数をほぼ同じ値にするかもしくは、正規化を行う。このトレーニングデータをもとに８×８×８×８ビットの頻度分布を作り、GA（遺伝的アルゴリズム）により極大点を探索するが、探索空間が多きすぎるので探索空間を主成分分析により縮小する操作を行う。主成分分析とは、相関のある多くの変数の値を、少数個の合成変量（主成分）で表す方法であるが、ここでは、LANDSAT・TMによる衛星画像データのバンド４～７の４つのバンドを第１主成分と第２主成分の２変量を用いて表す。第１主成分 u_1 と第２主成分 u_2 は、８×８×８×８ビットの頻度分布の座標を (b_1, b_2, b_3, b_4) とすると、

$$u_1 = a_{11}b_1 + a_{21}b_2 + a_{31}b_3 + a_{41}b_4 \qquad (1)$$

$$u_2 = a_{12}b_1 + a_{22}b_2 + a_{32}b_3 + a_{42}b_4 \qquad (2)$$

と表される。ここで係数ａの決定は、各クラスのトレーニングデータの分散共分散行列の固有値を求め、この固有値に対応する固有ベクトルを第１主成分と第２主成分の係数としている。

これにより、第１主成分と第２主成分の９×９ビットの頻度分布に直す。この際、第１主成分と第２主成分の累積寄与率が重要になるが、目安として累積寄与率が95％以上のものを採用し、それ以下のものに対してはもう一度トレーニングデータを取り直すか、このクラスに対しては極大値の抽出は行わず、平均値で置き換えるなどの操作をする。

次に、頻度分布内に存在する極大点を遺伝的アルゴリズムによって探索するが、頻度分布内には複数の極大点が存在し、また、クラス数分の極大点を抽出を行いたいが、通常の遺伝的アルゴリズムでは解が１つしか求められない。そこで、極大点の抽出を２段階の操作にわけ、２種類の遺伝的アルゴリズムを用いることによって複数の極大点を抽出することにする。

まず、最初に極大点の候補点を探索する。候補点の探索は星、山本[1]によって研究された多峰サーフェスモデルの極大抽出用GAにより探索する。この極大抽出用のGAは保存世代間隔を設け、その間隔毎にそのとき１番評価の高い個体を保存する。さらに、保存された個体の近傍に対してペナルティ半径を設定し、この半径内にある個体の評価に対してペナルティをかける操作を行う。

次に、この保存された各個体の近傍６×６ビットの範囲に探索区間を絞り込み、通常のGAにより極大点を抽出する。しかし、極大点の候補点を抽出する際に、ある保存世代間隔毎にそのとき１番評価の高い個体を保存するため、必ずしも設定したクラス数と同数の極大候補点を抽出できるとは限らず、クラス数より多くの極大候補点が抽出される。

そこで、各極大点は極大候補点が密集しているところに存在する可能性が高く、密集する部分をそれぞれ６×６ビットの範囲として定義し、この６×６ビットの各範囲を通常のGAにより探索し、極大点を抽出する。

また、これにより抽出された極大値は、どのクラスの極大点になるのかという対応付けがなされていないが、これは各クラスの平均値との差と極大点の評価値を考慮して対応付けを行う。

---

*"Estimation of statistics of training data using Genetic Algorithms"

†Takashi Hoshi and Kazuya Ishiguro
‡Ibaraki University
4-12-1 Nakanarusawa,Hitachi,Ibaraki 316,Japan

3 クラスタリング手法

パターン認識には、教師付き分類の中で最も一般的な最大尤度法を用いる。ここで、最大尤度法のアルゴリズム[3]は、画像データのバンド数をn、グランドトゥルースデータのクラス数をl、第kクラスのグランドトゥルースデータの個数をmとすると、グランドトゥルースデータ x_{ik} は、次式のように書き表される。

$$x_{ik} = \{x_{ijk} | i = 1, m; j = 1, n; k = 1, l\} \qquad (3)$$

第kクラスの平均を $\bar{x_k}$、分散共分散行列を S_k とすると各画像データ $x = \{x_j | j = 1, n\}$ に対する第kクラスの尤度 $f_k(x)$ は、次の式で表される。

$$f_k(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |S_k|^{\frac{1}{2}}} exp| - \frac{1}{2} d_k^2| \qquad (4)$$

ここで、$|S_k|$ は S_k の行列式であり、d_k は式(5)となる。

$$d_k^2 = (x - \bar{x_k})^t S_k^{-1} (x - \bar{x_k}) \qquad (5)$$

そして、未知データは尤度 $f_1(x), f_2(x), \cdots, f_l(x)$ のうち最大尤度 $f_k(x)$ をもつ第kクラスへと分類される。このように最大尤度法とは各クラスの尤度関数（確率密度関数）を多次元正規分布と仮定し、未知のデータに対し各クラスの尤度を計算し、最大尤度をもつクラスに分類するものである。

4 実験結果

LANDSAT・TMのトレーニングデータにおけるバンドの平均値、標準偏差、GAにより抽出された極大値を表1に示す。さらに、表1のデータの中でデータNo. 1と6のヒストグラムをそれぞれ図1と図2に示す。

図1は、海のトレーニングデータとして抽出したデータのヒストグラムであるが、標準偏差が2.4と小さく、平均値と極大値もほぼ一致している。図2は、雲のトレーニングデータとして抽出したデータのヒストグラムであるが、標準偏差が海のトレーニングデータのそれと比べると、62.8とかなり大きなものとなっている。性状も正規分布のものとかなり異なり、平均値と極大値においてもかなり差がでている。これは、雲のトレーニングデータを抽出する際に、雲のみのデータを抽出するのが困難なため、平均値が他のクラスのデータの影響を受けているものと思われる。また、データNo. 2は湖のトレーニングデータを抽出したもので、データNo. 7は、湖のトレーニングデータに裸地域のデータがノイズとして混ざったものであるが、極大値はデータNo. 2と同じ値を抽出しているのに対し、平均値はノイズの影響を受けてしまっている。

表1 平均値と極大値の比較

データNo.	平均値	極大値	標準偏差
1	44.9	44	2.4
2	63.2	64	3.5
3	87.3	88	6.4
4	49.9	48	6.1
5	41.3	36	6.4
6	220.9	255	62.8
7	70.6	64	11.0

5 おわりに

このように極大値は、トレーニングデータの性状が、正規分布と仮定できない場合やデータにノイズが入る場合などに有効であるといえる。今後、平均値と極大値をどのように使い分けるかが課題となろう。

参考文献

[1] 星 仰、山本 真靖、庄野 誠二：" 遺伝的アルゴリズムによるクラスタリングへの応用" 、情報処理学会第５０回全国大会、5Q-4、pp.2-263 - 2-264 (1995)

[2] 北野 宏明：「遺伝的アルゴリズム」、産業図書、pp.4-46(1993)

[3] 星 仰：「地形情報処理学」、森北出版、pp.165-188(1991)

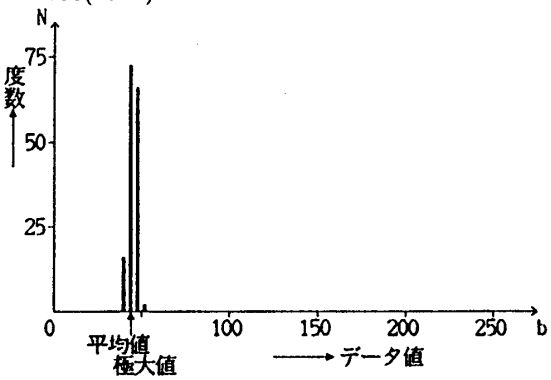

図1 データNo.1のヒストグラム

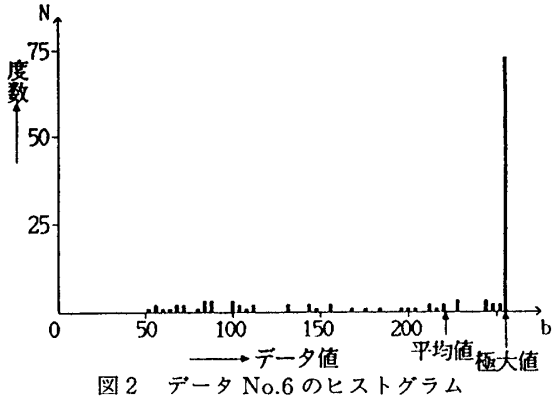

図2 データNo.6のヒストグラム