遺伝アルゴリズムによるクラスタリングへの応用

5Q－4

星　仰，山本　真靖，庄野　誠二
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最も基本的かつ広い用途をもつものに土地被覆分類図または、土地利用分類図の作成がある。これらを作成する際に、与えられた画像の各画素をヒストグラムでデータの性状を調べたりして、植生、裸地、人工構造物などのいくつかの分類項目にクラスタリングすることが必要となる。クラスタリング手法には、距離を基本として考えたユークリッド距離法や、平均・分散を基本とした最尤法などある。しかし、標準偏差のみを考えた場合、解析するデータが正規分布に類似していなければ精度の悪いものとなる。また、頻度のみを考慮した場合に頻度が高くても、分散が小さければ精度が必ずしも高くならない。

そこで、これらの問題点を解決するため本研究では、まず最初に極大候補点を抽出するために多峰性関数探索用につくられたGAを用いる。次に極大点を抽出するために一般的GAを用いる。そして、求められた極大点から標準偏差を求め、　距離、標準偏差、頻度の３要素を全て考慮した手法でクラスタリングを行う。

2　遺伝アルゴリズムによる極大点探索

極大点とは、Landsat TMによる衛星画像データのバンド１（分類項目は植生、人工構造物、裸地、表面水、雪・雲）とバンド４（分類項目はバンド１に同じ）を８×８ bitの頻度分布に直し、８×８ bitの範囲内に存在する各分布の頻度が最も大きい点と定義する。極大点を求めるにあたり、まず最初に、極大点の候補点を探索する。候補点は、山本、星により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。画像データは多峰関数的要素を含むため、極大抽出用GAを用いる。極大抽出用のGAは、個体群が初期の段階で局所解に陥ってしまい、その後、大域最適解がもとめにくいといった通常のGAの問題点を逆手にとり、多数の局所解を求めることを目的としたGAである。保存世代間隔を設け、その間隔毎に１番評価の高い個体を抽出したり、最大値近傍をベナルティー半径により設定し、その内にある個体の評価に対しベナルティー をかけるといった最大値近傍の再評価などの特徴がある。

各極大点は、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しているところをそれぞれ６×６ bitの範囲として定義する。範囲の決定は、任意候補点を基準点P'(x',y')とした場合、(x+32 , y+32) , (x-31 , y+32) , (x+32 , y-31) , (x-31 , y-31) の４点により範囲αが決定され、範囲α内の点において最上、最下、最右、最左の４点$(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)$をとり、４点の中心$((x_3+x_4)/2,(y_1+y_2)/2)$を基準点$P_1$(x,y)とし、もう１度範囲$\beta$をとり直す。この操作を２回繰り返すことにより、範囲γが候補点の密集地へ移動し、範囲γ内に極大点が入る確率が高くなっていくことが推測される。以上の操作を密集地がなくなるまで繰り返し範囲を決定する。決定された各範囲内でGA探索を行い、最終的に極大点P(x,y)が範囲の数だけ決定される。

3　標準偏差の検出方法

各頻度分布が正規分布であると仮定し、２点$P_1(x_i , y_j)$, $P_2(x_i , y_j)$を用いて標準偏差の検出をする。ただし、２点を結んだ直線上に極大点Pを存在させるようにP_1,P_2を決定させる。）

正規分布の確率密度関数の式より、

$$f(x_i,y_i)=\frac{h_1}{S}=\frac{1}{\sqrt{2\pi}\sigma}\epsilon^{-\frac{x_i^2+y_i^2}{2\sigma^2}}$$

$$g(x_j,y_j)=\frac{h_2}{S}=\frac{1}{\sqrt{2\pi}\sigma}\epsilon^{-\frac{x_j^2+y_j^2}{2\sigma^2}}$$

h_1：P_1における密度、h_2：P_2における密度

$$S=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx$$

以上の２式の変形より、

$$\sigma=\sqrt{\frac{x_2^2-x_1^2+y_2^2-y_1^2}{2(\log h_1-\log h_2)}}\qquad(h_1>0,h_2>0)$$

The Application of Clusterring using Genetic Algorithm
Takashi Hoshi , Naoyasu Yamamoto , Seiji Shono
Ibaraki University

となる。上記の式を用い、標準偏差を求めたが、各分布には複数の凹凸があるため2点(1組)だけでは、厳密な標準偏差の値に比べ誤差を生じる可能性がある。そこで、精度を高めるために2 n点(n組)をとり、精度不良のe組を削り、n-e組での平均により厳密値に近づけている。

4 クラスタリング手法

標準偏差のみでクラスタリングを行う場合、頻度分布データに大きな凹凸がある時や頻度分布データが正規分布に類似しない時、クラスタリングの精度が悪くなるという問題がある。また、極大点の頻度を用いてクラスタリングを行う場合には、頻度は高いが、分散が小さいといった時、精度が悪くなるという問題がある。上記のような問題点を解決するため標準偏差と極大点の頻度を共に用いたクラスタリングを行う。

まず、標準偏差と極大点の頻度のクラスタリングに影響を及ぼす比率を統一する。統一することにより、標準偏差や頻度に偏ったクラスタリングを避けることができる。比率を一定にするため、求められた各標準偏差$\Sigma(\sigma_1,\cdots\sigma_i,\cdots\sigma_n)$と各頻度$H(h_1,\cdots h_j,\cdots h_n)$の最大の$(\sigma_i,h_j)$を共に1.0とし、双法の比率をあわせる。つまり、それぞれ$\Sigma'(\sigma_1/\sigma_i,\cdots 1,\cdots\sigma_n/\sigma_i)$, $H'(h_1/h_j,\cdots 1,\ \cdots h_n/h_j)$に変換する。その後、実際にクラスタリングを行う。

極大点P_1と極大点P_2において各頻度をh_1, h_2とした時、直線P_1P_2間の各点は、直線P_1P_2間の距離を$\frac{h_1}{h_1+h_2}:\frac{h_2}{h_1+h_2}$に内分する点Tにより、$P_1$の領域$h_1\sim t$に含まれるか$P_2$の領域$t\sim h_2$に含まれるかが決定される。$h_1\sim t$までの距離をa、$t\sim h_2$までの距離をbと定義すると$\frac{h_1}{a}=\frac{h_2}{b}$が成り立つ。

以上より、距離a、b(a+b=一定)を変化させていき、$\frac{h_1}{a}>\frac{h_2}{b}$の点は$P_1$の領域に、$\frac{h_1}{a}<\frac{h_2}{b}$の点は$P_2$の領域にあるといえる。標準偏差や(標準偏差+頻度)にも同じことがいえる。また、2クラスタの場合だけではなく、3クラスタ以上のときにも標準偏差、頻度、極大点からの距離が分かっていれば各点の領域決定は(標準偏差+頻度)/(距離)が最大となる領域に決定される。

5 実験結果

Landsat・TMの2バンドの頻度分布を図1に示す。これを頻度によるクラスタリングをしたを結果を図2に示す。図3は標準偏差によるクラスタリング結果である。さらに、図4は標準偏差と頻度を用いたクラスタリング結果である。これらの結果をさらに検討する予定である。

参考文献

[1] 星　仰、山本　真靖:"遺伝アルゴリズムによる衛星画像データ分類への試み"、土木学会48回大会、IV-217、1994.9.

[2] 山本、星:"GAによるサーフェスモデルの極大抽出"、電子情報通信春季大会、D-468、1994.3.

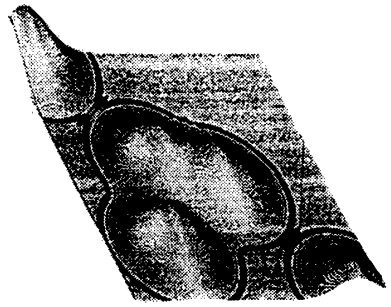

図1．原データの頻度分布

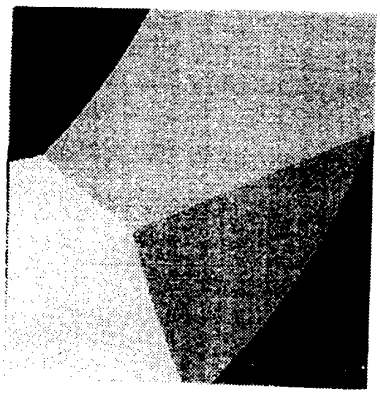

図2．頻度によるクラスタリング

図3．標準偏差によるクラスタリング

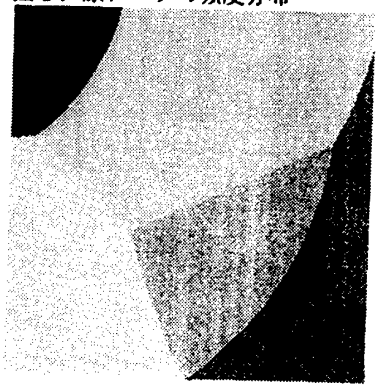

図4．本手法を用いたクラスタリング