主成分分析を用いた特徴抽出が容易なデータセットの作成

宮部 洋太

1　はじめに

多変量解析とは複数の変数からなる多変量データを統計的に扱うことで複雑な情報を的確に判断するための手法である．その多変量解析の手法の一つとして主成分分析がある．

主成分分析とは変数間の関係から主成分という合成変数を求め，少ない主成分でデータの特徴を代表させる手法である．

本研究では 2，3 個の主成分に情報をより集約させることでデータの特徴を捉えやすいデータセットを作成できるのではないかと考えた．本研究では元データから一定割合以上のデータを選択するという制約の下，情報が集約するデータの組合せを探索することで特徴を抽出しやすいデータセットを作成する手法を提案する．

今回はこの問題を最適化手法の一つである遺伝的アルゴリズムに適用させ，その動作確認を行った．

2　主成分分析を用いた特徴抽出が容易なデータセットの作成

2.1　主成分分析

主成分分析 (Principal Component Analysis:PCA) とは多変量データを要約し少ない情報で多変量データの特徴を表す手法である．

PCA を行うことで主成分という新しい指標（直線の式，軸）が求められ，変数やデータの類似性を可視化したり，影響力が大きい変数を発見したりできる．

2.1.1　主成分

P 個の変数 x_p を持つデータに PCA を行うと P 個の主成分が得られる．各主成分 $Z_p(p = 1, 2, \cdots, P)$ は各変数の値 x_p と各変数の重み $a_{pi}(i = 1, 2, \cdots, P)$ の合成変数で表される．各主成分 Z_p は式 (1) のように表せる．

$$Z_p = \sum_{i=1}^{P} a_{pi} x_i (i = 1, 2, \cdots, P) \tag{1}$$

これらの主成分は相関行列の固有値問題の解として得られる．

2.1.2　寄与率

寄与率とは一つの主成分がどの程度データの特徴を表しているかを示す指標である．寄与率は 0 以上 1 以下の値をとり，$V_p(p = 1, 2, \cdots, P)$ を主成分の分散とすると主成分 Z_p の寄与率 C_p は式 (2) から求められる．

$$C_p = \frac{V_p}{\sum_{i=1}^{P} V_i} \tag{2}$$

寄与率は第 1 主成分がもっとも高く，第 2 主成分，第 3 主成分，$\cdots$ と徐々に低くなるため，少ない主成分にデータが集約し，少ない次元でデータの特徴を表すことができる．

2.2　合計寄与率の最小化

本研究では第 2 主成分までに情報を集約させ，データの 80% 以上を選択するという制約を考える．先に述べたようにひとつの主成分がどの程度データの特徴を表しているかは寄与率から分かる．よって第 3 主成分以降の合計寄与率を目的関数とし，本研究の問題をこの目的関数を最小化する問題として定式化を行う．このような制約条件における最小化問題は式 (3) 式 (4) のように定式化される．

$$f = \frac{\sum_{i=3}^{P} V_i}{\sum_{i=1}^{P} V_i} \tag{3}$$

$$g_i = \sum_{i=1}^{N} x_i \geq 0.8N \tag{4}$$

式 (3)，式 (4) では N 個のデータがあり，各データ $i(i = 1, 2, \cdots, N)$ は P 個の変数を持っているとする．また $x_i = 1$ はデータ i を組合せに入れるときを，$x_i = 0$ はいれないときを表現している．V_i は x_i が 1 であるデータ集合の主成分の分散を表している．

3　GA の合計寄与率の最小化問題への適用

3.1　遺伝的アルゴリズム

遺伝的アルゴリズム (Genetic Algorithm:GA) とは生物の進化過程を模倣した最適化アルゴリズムである．

GA では何らかの方法でたくさんの個体を生成し，その集合を初期世代とする．一つ一つの個体がもつ遺伝子から目的にあっているかどうかの評価を行い，適合度を算出する．初期世代から適合度が高いものが多くなるように選択し，交叉や突然変異という操作を行い次世代の個体群を生成する．この処理を満足がゆく適合度を持つ個体が発生するまで繰り返す．

3.2　遺伝子表現

データの母数を N とする．長さ N の遺伝子によってデータの組合せを表現する．データの組合せに $i(i = 1, 2, \cdots, N)$ 番目のデータを使う場合は i 番目の遺伝子座を 1，使わない場合は 0 とする．ただし 1 の数が制約条件を満たすように，つまり母数 N の 80% 以上となるように遺伝子を生成する．

3.3 評価

遺伝子から表現されるデータの集合を A，変数の数を P とする．A に対して主成分分析を行い得られた各主成分の分散を $V_i(i = 1, 2, \cdots, P)$ とする．式 (5) から求められる第 3 主成分以降の合計寄与率の逆数を適合度関数とする．合計寄与率の逆数を適合度関数としているのは選択方法としてルーレット選択を採用するためである．

$$fitness = \frac{\sum_{i=1}^{P} V_i}{\sum_{i=3}^{P} V_i} \tag{5}$$

3.4 交叉

本研究が扱う遺伝子モデルでは制約条件として母数の 80% を使うという制約条件が定められている．しかしながら通常の一点交叉を適用すると，制約条件から外れて致死遺伝子が発生する可能性がある．そのため本研究では以下に示す交叉手法を用いる．

1. 一点交叉を行う
2. 新たに生成された個体の遺伝子が制約条件を満たさない場合は次の処理を行う．
3. 制約条件を満たすまでランダムに選ばれた 0 を 1 に反転させる．

3.5 突然変異

突然変異率に従って遺伝子を 1 ビット反転させる．なお突然変異率は 1/遺伝子長 とする．

4 動作確認

4.1 動作確認に用いたデータ

今回の手法によって第二主成分までに情報が集約することを確認するため UCI のデータベース [2] から公開されている wine regognition data というイタリアのワイン関するデータを用いて自作プログラムの動作確認を行った．これは 178 個の個体から構成される 13 変数のデータである．

4.2 パラメータ

動作確認に使用したパラメータを Table 1 に示す．

Table1　パラメータ

	GA
個体数	200
最大世代数	400
交叉率	0.9
突然変異率	1/遺伝子長
遺伝子長	178
選択方法	ルーレット選択
エリート数	1

4.3 結果

Fig. 1 に GA の各世代における最良の評価値の推移を示す．Fig. 2 に GA を行う前と行った後の寄与率の変化を示す．

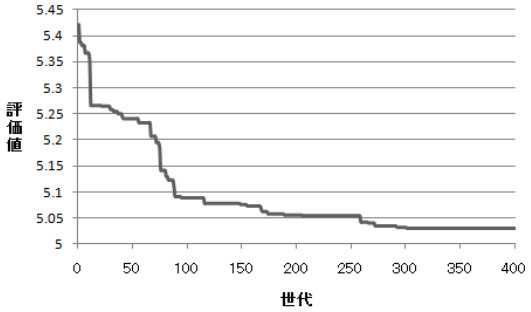

Fig.1　GA の解探索の推移

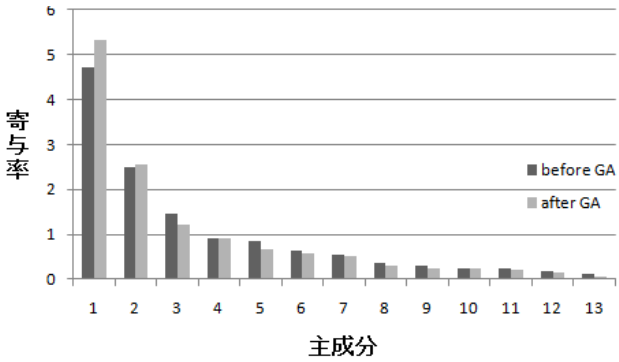

Fig.2　寄与率の変化

Fig. 2 より第一主成分，第二主成分の寄与率が探索前と比較して増加していることが確認できる．

5 まとめと今後の課題

本研究では，少ない主成分に情報を集約することでデータの特徴抽出をしやすいデータセットを作成できると考え，元のデータから一定割合以上を選択するという制約の下，第三主成分以降の合計寄与率を最小化するデータの組合せを探索する手法を提案した．

自作したプログラムの動作確認を行った結果，探索前と比べて第二主成分までに情報が集約していることがわかった．しかしながら本当に特徴を抽出しやすいデータセットを作成できたかについては検証が必要である．また動作確認によって得られた評価値は最適解ではないため，GA アルゴリズムの交叉，突然変異法や遺伝子表現，パラメータ等についても再考する必要がある．

参考文献

1) 管民朗：多変量解析の実践 (上)，現代数学社，2005.
2) UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/