The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

実数値遺伝的アルゴリズムの遺伝子分布情報と解探索進捗率を活用した変数選択手法の提案

A proposal of a new variable selection method utilizing gene's distribution information and solutions search progress rate of Real-coded genetic algorithms

小畑崇弘　　倉橋節也
Takahiro Obata　　Setsuya Kurahashi

*1筑波大学大学院ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba

Recently variable selection and parameter optimization are getting more and more important. Regarding parameter optimization, much attention has been paid to Real-coded Genetic Algorithms (RCGA) because of their good searching ability and high flexibility. As for variable selection, traditionally Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) are used quite often as selection criteria. These criteria estimate the relative quality of analysis models for a given set of data, but do not evaluate the importance of the variables themselves. This paper proposes a new variable selection method applying RCGA. This new variable selection method consists of 2 main components. The one is a new variable selection criterion utilizing the variances of genes in RCGA and the other is an estimation method of how far is in progress of RCGA optimization. The effectiveness of this new variable selection method is confirmed through application to a multiple linear regression model.

1. はじめに

近年，計算機システムや計測・測定技術の発展により，様々な分野で大量かつ複雑なデータの獲得と蓄積が進み，それに伴い，多数の分析データから重要なものを選ぶ変数選択手法の重要性が高まっている．また，分析モデルの複雑化も進んでおり，モデル内のパラメータ推定の難易度が高まっている．これらの問題に対し，本稿では実数値遺伝的アルゴリズム（RCGA）の同一世代内における遺伝子の分散を活用した変数選択指標とRCGA による解探索進捗率の推計値を用いた変数選択手法を提案する．当手法を用いることで，遺伝的アルゴリズム（GA）の枠組みの中でパラメータ推定と変数選択の両方に対応できる可能性を示す．

本稿の構成は次の通りである．2 章で，関連研究として GA および変数選択について概略する．3 章では新たな変数選択指標として，遺伝子の分散を活用した I 値について説明する．4 章では，変数選択のタイミングを判定するために GA による解探索の進捗率という考えを導入し，その推計方法について説明する．5 章は I 値と進捗率を用いた変数選択手法の実行結果についてで，6 章はまとめである．

2. 関連研究

2.1 実数値遺伝的アルゴリズム

遺伝的アルゴリズムは進化計算の一種で，環境に適応した生物が生き残り，適応できなかった生物が死滅する自然淘汰の考え方が取り入れられており [goldberg 89]，(1) 目的関数の微分可能性を仮定する必要がない，(2) 大域的探索が可能，といった特徴を持つ．

GA では遺伝子型として古くから 0 と 1 によるビットコーディングが使われてきたが，実数値パラメータを扱う最適化問題を解く場合，遺伝子型空間の位相構造は表現型空間である実数空間の位相構造とは異なっており，表現型空間において互いに近い二つの親個体から生成された子個体が，遺伝子型空間では親個体の近傍にあっても表現型空間上で親個体の近傍に生成されるとは限らない．従って GA による探索の後半で有望な領域を重点的に探索したい場合に探索範囲が大きく移動してしまう恐れがあり，探索に無駄が生じる可能性が指摘されてきた．

こうした指摘に対応したのが，遺伝子型として実数値ベクトルを扱う RCGA である [wright 91]．RCGA は，子個体生成（これを交叉と呼ぶ）の際に実数値ベクトルを直接操作するので，表現型空間における親個体群の近傍に子個体群を生成することが出来る．そのため，従来のバイナリコーディングに較べて実数値を扱う問題では解の探索効率が格段に向上する．

RCGA の解の探索能力は，交叉モデルと世代交代モデルにどのような手法を組み合わせるかによって決まる．本稿では RCGA の世代交代モデルとして Just Generation Gap (JGG) を，交叉モデルとして AREX を用いた [秋本 07][秋本 09-1][秋本 09-2]．JGG と AREX の組み合わせは解の探索能力に優れた代表的な組み合わせである．簡単に各モデルを紹介すると，JGG は多親交叉に適したモデルで，生存選択の対象は子個体に限定され交叉に参加した親固体は全て淘汰される．AREX は多親交叉の一般的枠組みである REX[小林 09] に初期収束を回避するための仕組みを取り入れた手法である．その一つは交叉の拡張率を適応的に調節する仕組みで，集団が重心移動中に交叉による子個体生成領域を広げるようにコントロールすることで局所解への初期収束を回避する．もう一つは交叉中心降下というもので，子個体の生成分布の中心を親個体群の重心から最適解のあると思われる方向にシフトさせる仕組みであり，この概念を導入したことにより，最適解が初期化領域外にある場合でも解に到達する能力が向上している．

2.2 変数選択

複数のモデルの中からデータをうまく説明するモデルを選択することをモデル選択という．モデル選択の中でも，多数の説明変数の中から目的変数の変動を説明するのに適した変数を取捨選択する問題を，特に変数選択という．

変数選択を行う際にはモデルを比較する順番を決める変数選択アルゴリズムと，モデルの優劣を判断する選択基準の二つが重要である．

連絡先: 氏名，所属，住所，電話番号，Fax 番号，電子メールアドレスなど

2.2.1 変数選択アルゴリズム

変数選択アルゴリズムの中で最も確実な方法は，可能な変数の組み合わせを全て比較する総当たり法である．しかし説明変数の数を p とすると組み合わせの総数は $2^p - 1$ となり，説明変数が 10 個の場合は組み合わせ数は 1,023，説明変数が 20 個になると 1,048,575 といった具合に説明変数の増加によって計算コストが爆発的に増大してしまう．そこで準最適な方法として，変数増加法，変数減少法，変数増減法などが用いられている．

2.2.2 変数選択基準

変数選択の基準は様々なものが用いられてきた．古くは残差平方和が利用されてきたが，説明変数の数が増えれば残差平方和は必ず小さくなるため残差平方和をそのまま用いるのではなく，残差平方和に修正を加えた自由度調整済み決定係数や Mallows の提案した C_p が利用されるようになった．その後，真の分布に対する予測分布の良さという視点から赤池情報量規準（AIC）が提唱され，広く活用されるている．さらにベイズ情報量規準（BIC）や最小記述長（MDL）原理などの基準も提案され，状況に応じて各基準が用いられている．

情報量基準による変数選択の場合，変数の重要性に関する情報が陽には示されず，最終的に選ばれたモデルの変数を受け入れるしかない．特に，非連続・非線形なモデルの場合は，変数の重要性をどのように評価するかが課題となる．

2.2.3 変数選択に GA を応用した事例

変数選択に GA を応用した既存研究としては [Broadhursta 97][栗田 94] などがある．これらの研究では GA を変数選択アルゴリズムの替わりに用いており，本稿で目指すものとは軸を異にしている．

3. RCGA の遺伝子の分散を活用した変数選択指標

3.1 遺伝子の分散と標準誤差

分析モデル内のあるパラメータの変化がモデルの評価値に与える影響の大きさはパラメータによって異なり，その相違が RCGA でパラメータ推定をする際の遺伝子の分散の大小に反映されていると考えられる．[小畑 18] はこの点を検証するため，(1) 式のような線形回帰モデルを用いて実験を行った．

$$y = a_0 + a_1 \cdot x_1 + \cdots + a_p \cdot x_p \tag{1}$$

(1) 式の y は目的変数，x_1〜x_p は説明変数，a_0 は定数項，a_1〜a_p は回帰係数，p は説明変数の数である．この式の定数項および回帰係数を RCGA によって推定し，その際の遺伝子の分散について分析した．2 章で説明した通り，利用した RCGA の交叉モデルは AREX，子個体生成モデルは Just Generation Gap (JGG) である．世代内の個体数や子個体生成数は [秋本 09-1] の推奨値を採用した．すなわち，個体数は問題の次元数の 10 倍，子個体生成数は次元数の 4 倍である．

各個体は定数項および各回帰係数に対応する遺伝子を持つ．遺伝子の分散についてはパラメータ毎に各個体の対応する遺伝子のみを取り出し，その遺伝子群ごとに分散を計算した．以後，遺伝子の分散を Vg と表す．例えば回帰係数 a_1 に対応する遺伝子の分散は Vg_{a1} と表す．

実験結果の分析から，RCGA によるパラメータ推定が収束した状態では，あるパラメータに対応する Vg の値と重回帰分析による当パラメータの標準誤差がほぼ比例するという関係を見出した．この関係は (2) 式のように表せる．

$$\frac{Vg_a}{STD_a} \simeq \frac{Vg_{a1}}{STD_{a1}} \simeq \cdots \simeq \frac{Vg_{ap}}{STD_{ap}} \simeq K \tag{2}$$

ここで STD_i はパラメータ i の標準誤差を表し，K はある定数を表す．

3.2 I 値の導入と変数選択

ここで，(3) 式で定義する I 値という新しい指標を導入する．

$$I_i = \frac{\nu_i^2}{V_{gi}} \tag{3}$$

I_i はパラメータ a_i に対する I 値である．ν_i はパラメータ a_i の推定値，V_{gi} は同一世代内の個体のパラメータ a_i に対応する遺伝子の分散を表す．

線形回帰モデルにおいて T 値が回帰係数÷標準誤差で表されることをを考慮すると，I 値が T 値の 2 乗（つまり F 値）と比例関係にあると期待できる．実際の分析結果でも I 値と F 値との比例関係が確認できている．T 値は説明変数の係数の有意水準の判定に用いられる．この T 値の 2 乗と比例関係にあるのであれば，I 値も係数の有意水準が反映されていると考えらえ，I 値の大小は変数選択指標として活用できる可能性がある．[小畑 18] では I 値を変数選択基準として用いた変数選択を実行し、AIC を選択基準としたステップワイズ変数選択と似た結果が得られたと報告している．

4. 最適解探索進捗率の推計

3 章で遺伝子の分散を活用した変数選択指標を導入した．しかし，どういったタイミングで変数選択を行えばいいのかは未検討であった．そこで，この章ではソフトウェア信頼度成長モデル（SRGM）の手法を応用して RCGA の最適解探索の進捗率を推計し，その進捗率に基づいて変数選択を実施することを検討する．

4.1 SRGM を応用した RCGA の最適解探索進捗率の推計

SRGM とは，ソフトウェア開発において生じる累積バグ数を推定する手法の一つで，テスト工程で検出されるバグの累積値の傾向から残存バグ数を推定する方法である．これまで様々な SRGM が論じられているが，多くが指数型や S 字型などの各モデル固有の典型的な特性を持った曲線をもとにしたものであり，あるモデルを固定的に使用すると，対象データの特性によっては残存バグ数の推定精度が低くなることが指摘されている．こうした問題を解決するひとつの手段として代表的なモデルを包含する統合モデルが提案され，特に [古山 96] は統合モデルを表す微分方程式の対数をとることにより，得られたデータ系列から対数誤差の二乗和を最小にする方法で解析的にパラメータを推定することができることを示した．統合モデルは次の微分方程式で表される．

$$\frac{d(y+\delta)}{dt} \cdot (y+\delta)^{\gamma-1} = \alpha \cdot \mathrm{e}^{-\beta t} \tag{4}$$

ここで y は時刻 t における累積バグ数である． 式 (4) は $\delta = 0$ のとき次式のように表せる．y' は時刻 t で発見されたバグ数である。

$$y' \cdot y^{\gamma-1} = \alpha \cdot \mathrm{e}^{-\beta t} \tag{5}$$

(5) 式の両辺の対数を取ると, 次のようになる.

$$\ln y' + (\gamma - 1)\ln y = \ln \alpha - \beta t \tag{6}$$

式 (6) から対数誤差の二乗和を最小にする方法で α，β，γ を推定し，そこから累積バグ数を算出するのが [古山 96] の手法であり，これを RCGA による解探索の進捗率推定に応用する．そのため，y' および y を次のように置き換える．

y ': t-1 世代の個体の評価値の平均値 − t 世代の個体の評価値の平均値（つまり評価値の平均値の改善幅）

y: y ' の累計

RCGA の解探索が順調に進めば，世代交代が進むにつれて評価値の改善幅は逓減していき，やがてゼロになることが期待できる．これはソフトウェア開発のテスト工程が進むにつれて時点ごとのバグ発見数が逓減していくのと類似している．SRGM では t 時点における最終累積バグ数予測と実際の累積バグ数の比をとることでテスト工程の進捗状況を管理するが，それと同様に t 世代における最終累計評価値改善幅の予測と実際の改善幅累計の比をとることで RCGA による解探索状況の進捗率の推計が可能になると考えられる．

前述のように，SRGM では様々な曲線をもとにしたモデルがあり，指数型，ゴンペルツ曲線，超指数型の 3 タイプを試したところ，ゴンペルツ曲線と超指数型では進捗率の推定が不安定だったことから本稿では以後，指数型モデルを用いて進捗率を推計する．指数型モデルは (7) 式で表される．N は最終累積バグ数の予測値である．RCGA の解探索の文脈では，N は評価値の平均値の改善幅の最終累計値となる．

$$y = N(1 - \mathrm{e}^{-bt}) \tag{7}$$
$$\alpha = Nb, \beta = b, \gamma = 1$$

5. I 値と最適解探索進捗率を用いた変数選択

5 章では，3 章で検討した変数選択指標，および 4 章で説明した解探索進捗率を用いた変数選択手法の実験結果について報告する．

分析に用いたデータセットはワイン品質（0〜10 までの 11 段階スコア）とワイン品質に影響する 11 項目の数値（実数値）からなる．データはカリフォルニア大学アーバイン校が提供している機械学習用データセット集から取得した．このデータセットを用いた重回帰モデルに対して提案手法による変数選択を実行した．

ワイン品質データによる重回帰分析，およびステップワイズによる変数選択結果を表 1 に示した．ステップワイズの結果，T 値の絶対値の小さい変数が変数削除されていることが分かる．

次に，同じ重回帰モデルの回帰係数を RCGA により推定し，探索が収束した段階で I 値の最も低い変数を削除することを繰り返す変数選択法を実行した．5 回の試行の結果，表 1 の変数 no でみて，9 → 2 → 5 → 4 → 7 → 10 の変数削除順は 5 回の試行全てで共通となり，その後は 6 → 8 → 11 → 3 → 1 の順が 2 回，8 → 6 → 11 → 3 → 1 が 2 回，8 → 6 → 11 → 3 → 1 が 1 回となった．この変数削除の順番は重回帰分析の T 値の絶対値の小さい順，もしくは P 値の大きい順に概ね一致しており，変数削除を 4 変数で止めた場合は AIC によるステップワイズで変数削除されたものと完全に一致する．

表 1: ワイン品質データに関する回帰分析結果

変数 no	変数名	重回帰分析 回帰係数	T 値	P 値	ステップワイズ 回帰係数	T 値	P 値
1	(Intercept)	22.0	1.04	0.300	4.43	11.00	< 0.001
2	fixed.acidity	0.0250	0.96	0.336	–		–
3	volatile.acidity	-1.08	-8.95	< 0.001	-1.01	-10.04	< 0.001
4	citric.acid	-0.183	-1.24	0.215	–		–
5	residual.sugar	0.0163	1.09	0.277	–		–
6	chlorides	-1.87	-4.47	< 0.001	-2.02	-5.08	< 0.001
7	free.sulfur	0.00436	2.01	0.0447	0.00508	2.39	0.017
8	total.sulfur	-0.00327	-4.48	< 0.001	-0.00348	-5.07	< 0.001
9	density	-17.9	-0.83	0.409	–		–
10	pH	-0.414	-2.16	0.0310	-0.483	-4.11	< 0.001
11	sulphates	0.916	8.01	< 0.001	0.883	8.03	< 0.001
12	alcohol	0.276	10.43	< 0.001	0.289	17.23	< 0.001

*灰色部分はステップワイズで削除された変数

表 2: ワイン品質データについての変数選択結果

変数 no	進捗率の閾値 0.999 試行 NO 1	2	3	4	5	進捗率の閾値 0.9999 試行 NO 1	2	3	4	5	進捗率の閾値 0.99999 試行 NO 1	2	3	4	5
1					3				1						
2				2	2		3	3	2	3	3				
3															
4															
5	2	2	2	3	1	3		1		1		1	1	1	1
6			1												
7	4	4					1				1				
8	1	1				1									
9	3	3	3	1		2	2	2		2	2	2	2	2	2
10															
11															
12															

*灰色部分はステップワイズで削除された変数

さらに，進捗率が閾値に達した段階で変数削除する変数選択法を，閾値を様々に変えて試した．このとき，GA の解探索終了条件として（1）評価値の改善幅がゼロになる，(2）進捗率が 1 に達する，(3）残った変数の数が 1 になる，のどれかに該当した場合に RCGA を終了するようにした．その結果，閾値を 0.9999999 以上に設定した場合は前述の解探索が収束した段階で変数削除する変数選択法と当初の 2〜3 変数までは変数削除順が同様だった．但し，その後は先に述べた RCGA 探索終了条件 (1) に該当して RCGA の処理が終了した．さらに進捗率の閾値を 10 倍ずつ変化させて変数選択を試したところ，0.99999 以上の閾値では変数削除順は，若干異なるケースがあるものの，概ね T 値の絶対値の小さい変数を先に削除する傾向がみられたが，進捗率の閾値を 0.9999 以下に設定すると変数削除順が不安定になった．

なお，進捗率 0.99999 に達するのは 200 世代目前後であり，一方で解探索が収束するのは 650 世代前後である．計算コストの面では進捗率の閾値を出来るだけ低い水準に設定するのが望ましいと言えるが，一方で閾値が低すぎると重要な変数を削除してしまう可能性が高まる．このトレードオフの関係に対して適当な水準を見出す方法が求められるがその点は今後の課題としたい．

6. まとめ

本稿では，RCGA の枠組みで変数選択を行う手法を提案し，各分野の分析で広く用いられている AIC による変数選択と類似の結果が得られることを示した．この変数選択手法のために次の 2 つの仕組みを導入した．

（1）RCGA の遺伝子の分散の大きさを活用した I 値という変数選択指標.

（2）SRGM を応用した，RCGA の解探索の進捗率推計法.

一方，本稿の今後の課題も多い．一つ目は変数選択を実行する進捗率の閾値の設定方法である．進捗率の進み具合やその水準と解探索状況の関係は分析に用いるモデルによって異なると考えられ，閾値の設定水準を検討する作業負担は小さくない．この閾値の設定方法に関する基準や計算方法を見出すことが今後の大きな課題と考えている．また，評価値改善幅累計の形状が異なった場合にその形状に適した進捗率計算モデルを選択できるようにすることも課題の一つである．

これらの点を改善することで，非線形や非連続などのより複雑なモデルでも変数選択を実行できるようにすることが今後の大きな目標である．

参考文献

[Broadhursta 97] Broadhursta, Goodacrea, Jonesa, Rowlandb, Kell, : Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry : Analytica Chimica Acta Vol. 348, 1-3, pp. 71-86, (1997)

[栗田 94] 栗田 : 遺伝的アルゴリズムによる線形回帰分析における説明変数の選択の試み，情報処理学会 全国大会講演論文集　第 48 回平成 6 年前期 (2) pp.253-254（1994）

[goldberg 89] Goldberg :Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley Longman Publishing Co. Inc., (1989)

[wright 91] Wright : Genetic Algorithms for Real Parmter Optimization, Foundations of Genetic Algorithms, pp. 205-218 (1991)

[秋本 07] 秋本，羽佐田，佐久間，小野，小林 : 多親を用いた実数値 GA のための世代交代モデル〜 Just Generation Gap (JGG) の提案と評価〜，SICE 第 19 回自律分散システムシンポジウム資料，pp. 341-346 (2007)

[秋本 09-1] 秋本，永田，佐久間，小野，小林 : 適応的実数値交叉 AREX の提案と評価，人工知能学会誌 24(6), pp. 446–458,（2009）

[秋本 09-2] Akimoto, Sakuma, Ono, and Kobayashi : Adaptation of expansion rate for real-coded crossovers, Proc. of the 11th Annual Conf. on Genetic and Evolutionary Computation (GECCO'09), pp. 739-746 (2009)

[小林 09] 小林 : 実数値 GA のフロンティア，人工知能学会誌 24(1) pp.128-143（2009）

[小畑 18] Obata and Kurahashi:A Study of Variable Selection within A Framework of Real-coded Genetic Algorithm, 2018 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2018)

[古山 96] 古山 : ソフトウェア信頼度成長モデルに関する統合モデルの解析的パラメータ推定法 情報処理学会論文誌, Vol.37 No.12, pp. 2326-2333, (1996)