グラフィクスとCAD 88－1
（1997．12．18）

スプライン関数を用いたデータあてはめ
―遺伝的アルゴリズムによる節点の自動的な決定―

吉本 富士市　　森山 真光
和歌山大学 システム工学部 情報通信システム学科

スプライン関数を用いたデータあてはめ問題でよい解（近似関数）を得るためには，しばしば節点を変数として扱う必要がある．そのとき，解くべき問題は多変数で多峰性の連続系非線形最適化問題となる．したがって，大域的な最適解を求めることは困難である．本論文では，元の連続系最適化問題を離散系最適化問題へ変換し，それを遺伝的アルゴリズムを用いて解く方法を提案する．離散化した節点の位置の候補を遺伝子と見なして個体を構成することにより，連続系から離散系への変換を行う．また，適応度として赤池の情報量規準*AIC*を用いることにより，統計的に最適なモデルを探索する．提案する方法は，節点の数と位置を自動的かつ同時に求めることができる．また，遺伝的アルゴリズムの特性から，大域的な準最適解を求められる可能性が高い．提案方法の有効性を示すため，数値計算例をあげている．

キーワード：スプライン，節点の配置，最適化，多峰性，遺伝的アルゴリズム

Data Fitting with a Spline Function
-Automatic Knot Placement by a Genetic Algorithm-

Fujiichi YOSHIMOTO and Masamitsu MORIYAMA
Department of Computer and Communication Sciences,
Faculty of Systems Engineering, Wakayama University

In order to obtain a good result (that is, a good approximation) for data fitting with a spline function, frequently we have to deal with knots as variables. Then, the problem to be solved becomes a continuous- nonlinear- and multivariate- optimization problem with many local optima. Therefore, it is difficult to obtain the global optimum. In this paper, we propose a method that we convert the original problem into a discrete combinatorial optimization problem and solve the converted problem by a genetic algorithm. We construct individuals by considering candidates of the positions of knots as genes, and we convert the continuous problem into a discrete problem. We search for the statistically best model among candidate models by using Akaike's Information Criterion (*AIC*) as a fitness function. Our method can determine suboptimal number and positions of knots automatically at the same time. By the characteristics of genetic algorithms, our method has high probability that gives a global suboptimum. Numerical examples are given to show the effectiveness of our method.

Key Words : Spline, knot placement, optimization, multimodal, genetic algorithm

1．まえがき

スプライン関数を用いたデータあてはめは，形状モデリングの重要な要素技術の一つである．よく知られているように，よいスプライン関数（よいモデル）を得るためには，通常は節点の数と位置を適切に決める必要がある．このとき，節点を変数として扱わなければならず，解くべき問題は多変数で多峰性の連続系非線形最適化問題となる[1)~2)]．したがって，大域的な最適解を求めることはきわめて困難である．

このため，いろいろな簡便法が提案されてきている[1)~6)]．しかし，「自動的によいモデルを得る手法」の観点から見るとまだ十分とは言えない．ここで「よいモデル」とは，データの元にある関数(underlying function of data)をできるだけよく近似し，しかもパラメータの数ができるだけ少ないスプライン関数のことを意味する．自動的によいモデルを得るためには，節点の数と位置を自動的に決めるアルゴリズムと，モデルのよさを評価するための客観的な規準が必要である．スプライン関数を用いたデータあてはめで，この両者を兼ね備えた汎用性の高い手法はまだ見あたらない．

ところで，スプライン関数がよい近似関数となるために必要な節点の数と位置は，通常は厳密な意味での最適値でなくてもよく，準最適値であれば十分である[1)，7)]．このため，上記の解くべき連続系の問題をある程度細かく離散化すれば，離散化された問題の最適解を元の問題の最適解の代わりに用いても十分よい結果を得ることができる．

本論文では，元の連続系の最適化問題を，離散系の組合せ最適化問題へ変換し，それを遺伝的アルゴリズム[8)~11)]を用いて解く方法を提案する．以下，遺伝的アルゴリズムを簡単のためGA（Genetic Algorithm）と呼ぶことがある．GAを用いることにより，節点の数と位置を自動的かつ同時に決定することが可能となる．また，GAの評価関数として赤池の情報量規準*AIC*（Akaike's Information Criterion）[12)]を用いることにより，データの元にある関数を最もよく近似するモデルを選択することができる．提案する手法は，多次元格子点データへの拡張が容易である，並列計算への適合性もよい，などの優れた特徴をもっている．

2．スプライン関数によるデータあてはめ

あてはめを行うべきデータは，x軸上の区間$[a,b]$内で与えられ，

$$F_j = f(x_j) + \varepsilon_j \quad (j = 1,2,\cdots,N) \tag{1}$$

と表わされるものとする．ここで，$f(x)$はデータの元にある関数であり，ε_jは平均値0，分散σ^2の正規分布をする互いに独立な誤差であると仮定する．また，$f(x)$は未知の関数であり，そのよい近似関数を作ることがあてはめの目的である．

必要な節点を$\xi_i (i = 1-m, 2-m, \cdots, n+m)$と書くことにする．ここで，$n$は区間$[a,b]$の内部に配置する節点$\xi_i (i = 1,2,\cdots,n)$の数である．また$m$は，式(3)に示すように，$f(x)$の近似関数$S(x)$を表わすために使うB-スプライン$N_{m,i}(x)$の階数（次数＋1）である．両端の$m$個の節点はそれぞれ端点$a, b$に重ね，

$$\left.\begin{aligned} a &= \xi_{1-m} = \cdots = \xi_0 \\ b &= \xi_{n+1} = \cdots = \xi_{n+m} \end{aligned}\right\}. \tag{2}$$

とする．

このとき，近似関数$S(x)$は

$$S(x) = \sum_{i=1}^{n+m} c_i N_{m,i}(x) \tag{3}$$

と表わすことができる．ここで，c_iはB-スプライン係数である．

式(3)に含まれるB-スプラインは，次の漸化式を用いて容易に計算できる[15)]．

$$N_{1,i}(x) = \begin{cases} 1 & (\xi_{i-1} \le x < \xi_i), \\ 0 & (otherwise), \end{cases} \tag{4}$$

$$N_{r,i}(x) = \frac{(x-\xi_{i-r})N_{r-1,i-1}(x)}{\xi_{i-1}-\xi_{i-r}} + \frac{(\xi_i - x)N_{r-1,i}(x)}{\xi_i - \xi_{i-r+1}}, \quad (r = 2,3,\cdots,m). \tag{5}$$

式(5)で，節点が分子および分母の両方に入って

いることに注意したい．すなわち，節点はB-スプラインの非線形パラメータである．

最小二乗法を用いて式(3)を与えられたデータ(1)にあてはめるとき，残差の2乗和Qは

$$Q=\sum_{j=1}^{N} w_j\left\{S(x_j)-F_j\right\}^2 \qquad (6)$$

となる．ここで，w_jはデータの重みであり，$N>n+m$とする．式(6)を最小にする条件からB-スプライン係数$c_i(i=1,2,\cdots,n+m)$を求めることができる．ただし，よい近似を得るためには内部の節点$\xi_i(i=1,2,\cdots,n)$の数と位置を適切に決める必要がある．

以上の議論から分るように，目的関数は式(6)であり，その変数はB-スプライン係数および内部の節点である．ここで，B-スプライン係数は，目的関数の線形パラメータであるが，内部の節点は非線形パラメータであることに注意したい．式(6)を最小化する問題は，多峰性の最適化問題となることが知られている[2)]．

3．遺伝的アルゴリズムの適用

3．1 遺伝的アルゴリズムを用いる理由

スプライン関数へGAを応用した研究はすでにあるが[13)，14)]，データあてはめの節点の決定へ応用したものはまだ見あたらない．一般に，GAは計算量が多いので，従来の解法（Newton法など）で容易に解ける問題には適用すべきではない．本論文で扱う問題に対してGAを用いる理由は，主として次の3つである．

(i) 目的関数は，多峰性の多変数関数であり，内部の節点が非線形パラメータとなっている．

(ii) 目的関数を非線形パラメータで微分することが容易でない．

(iii) 通常は，非線形パラメータ（内部の節点）の数の適切な値は未知であるため，その数も変数として最適化問題を解く必要がある．

これらのことから分かるように，解くべき問題は変数の数が未知で複数個の極値をもつ最適化問題である．このため，その大域的な最適解を求めることは，従来の方法ではきわめて困難である．

この困難さを克服するために，本論文では次の方法を提案する：

(a) 元の連続系の最適化問題を離散系の組合わせ最適化問題へと変換する．

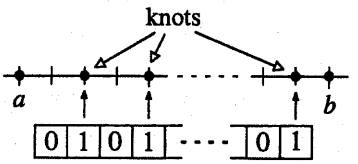
(b) その変換された組合わせ最適化問題をGAを用いて解く．

この方法は，準最適な節点の数と位置を自動的かつ同時に求めることができる特徴がある．

3．2 遺伝的アルゴリズムの構成要素

3.2.1 コード化

節点は連続変数であるので，それを次のようなコード化によって離散変数に変換する．x軸上の区間$[a,b]$を$L+1$等分してその（内部の）分点を染色体上の遺伝子座に対応させる．そして，ある遺伝子座の遺伝子が1ならその遺伝子座に対応する分点を節点とし，そうでなければ（0であれば）節点としないことにする（図1参照）．このようなコード化を行うと，遺伝子長（遺伝子列の長さ）はLとなる．



artificial chromosome (bit string)

図1　コード化

3.2.2 初期集団

染色体（個体）の初期集団としては，各遺伝子座の遺伝子をランダムに0か1とした染色体をK個発生させたものを用いる．ここでKの値は染色体の数（個体数）である．GAでは，この染色体を解候補の初期集団として最適解を大域的に探索する．

ところで，一般にデータの元にある関数$f(x)$が複雑な形をしている場合には，節点を多く必要とする．したがって，遺伝子として1を多く

もつ染色体が有利な個体となる．逆に，$f(x)$が単純な形をしている場合には，節点は少なく，1の少ない染色体が有利な個体となる．

したがって，計算の収束を早めるためには，データの関数形を見て区間$[a,b]$の内部に配置する節点の数nの初期値（すなわち初期集団に含まれる各個体の1の数）を制御できる仕掛けを導入した方がよい．そこで，節点率λを設けた．この値を0から1の間で変化させることによって，内部の節点数nの初期値の平均を制御できる．

3.2.3 評価関数

評価関数としては，赤池の情報量規準AICを用いる[1),12)]．AICは，2章で述べたあてはめ問題の場合には，

$$AIC = N\log_e Q + 2(2n+m) \qquad (7)$$

と表現できる．ここで，Nはデータの数，Qは式(6)で表される残差の2乗和である．また，$(2n+m)$はモデル関数に含まれるパラメータの数である．この中で，$n+m$はB-スプライン係数の数，nは内部の節点の数である．評価関数の値（AICの値）を適応度と呼び，それを最小にするモデルが最も良いモデルであると見なされる．

なお，GAの文献では，“適応度が大きいほど最適値に近い”と表現されている場合が多いが，本論文ではAICをそのまま評価関数として用いているため，“適応度が小さいほど最適値に近い”ことになるので注意されたい．

3.2.4 遺伝的オペレータと制御パラメータ

遺伝的オペレータとしては，選択，交叉，および突然変異の3つがある．本論文では，選択にはトーナメント方式を，交叉には2点交叉を，突然変異には遺伝子を一定の確率で対立遺伝子に置き換える方法を用いる．

また，GAを実行させるためには，いくつかの制御パラメータが必要である．その主なものは，個体数，遺伝子長，交叉の確率，突然変異の確率である．

4．データあてはめのアルゴリズム

GAを用いた節点の決定方法を組み込んだ，スプライン関数によるデータあてはめのアルゴリズムの概要は，次のようになる．

ステップ1：式(1)で表される，あてはめを行うべきデータを入力する．

ステップ2：個体数，遺伝子長，交叉・突然変異などの確率，節点率を設定する．

ステップ3：乱数を用いて染色体の初期集団を生成する．

ステップ4：各個体ごとに，それに対応する節点を用いてスプライン関数によるあてはめを行い，適応度を計算する．

ステップ5：最終世代まで計算したか？YESのとき計算を終了する．NOのとき次のステップ6へ行く．

ステップ6：各個体の適応度に基づいて選択を行う．

ステップ7：選択された個体に対して交叉を行い，次世代の個体候補を生成する．

ステップ8：個体候補に突然変異を行って次世代の染色体集団を作成し，ステップ4へ戻る．

上記のアルゴリズムの中では，ステップ4に計算負荷が集中しているが，必要であればこの部分は並列計算が可能である．

5．数値実験

前章で述べたアルゴリズムの有効性を調べるため，多くの例題を用いて数値実験を行った．その中から2つの結果を報告する．スプライン関数の次数は，最もよく使われている3次の場合について計算したが，本論文で提案する方法は次数には依存しないことに注意したい．

以下に示す例は，個体数＝50，遺伝子長L＝99，交叉率＝0.9，突然変異率μ＝0.01，節点率λ＝0.3（節点数nの初期値の平均が約30）とした場合である．なお，各図の「あてはめの結果」のx軸上にある黒丸は，節点の位置を表してい

る．

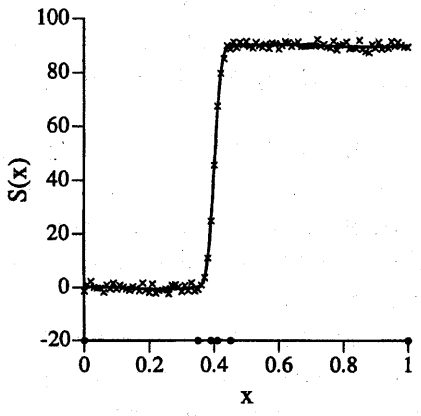
例１：ロジスティック曲線データの場合

あてはめるべきデータを次の式

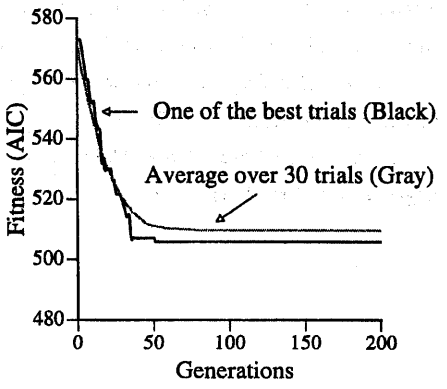
$$F_j = 90/(1+e^{-100(x_j-0.4)}) + \varepsilon_j \quad (j=1,2,\cdots,N) \qquad (8)$$

で作成した．ここで，ε_jは期待値０，分散１の正規分布をする誤差である．x_jの値は，0.0, 0.01, …, 1.0の101個とした．また，あてはめを行う区間は$[a,b]=[0,1]$にした．

計算結果を図２示す．図２(a)を見ると，節点がデータの元にある関数の変化が大きいところに集中している．このことは，専門家の経験的な知識[3),6)]とよく一致している．図２(b)に示すように，本例題では51世代で収束している．なお，図２(b)の灰色の線は，初期集団を変えて30回の試行を行い，その平均を取ったものである．また，黒い線は，それらの中でよい結果を与えたものの一つである．



（ａ）あてはめの結果



（ｂ）適応度

図２　ロジスティック曲線データの場合

例２：複合有理式曲線データの場合

あてはめるべきデータを次の式

$$F_j = 1.0/(0.01+(x_j-0.3)^2) + 1.0/(0.02+(x_j-0.6)^2) + \varepsilon_j \quad (j=1,2,\cdots,N) \qquad (9)$$

で作成した．ここで，ε_j，x_jの値，およびあてはめを行う区間は例１と同じである．

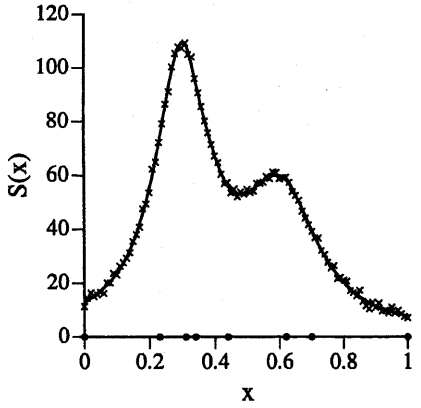
図３は計算結果であるが，この例でも節点はデータの元にある関数の変化が大きいところに集中しており，専門家の経験的な知識とよく一致している．図３(b)に示すように，本例題では59世代で収束している．

６．あとがき

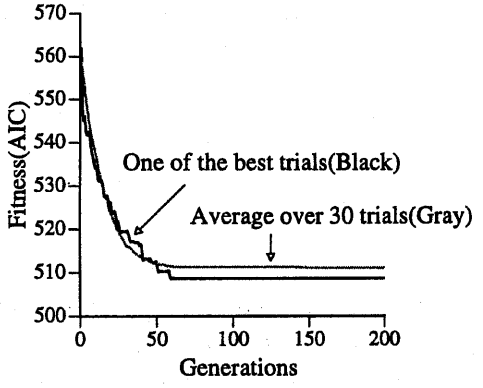
本論文では，スプライン関数を用いたデータあてはめの節点を，GAによって決定する方法を提案した．あてはめ問題の節点の決定は，多変数で多峰性の非線形最適化問題であり，それをまともに解くことはきわめて困難である．

しかし，この問題は実用上十分な精度で離散的な組合せ最適化問題へ簡単に変換できる．また，変換された組合せ最適化問題は，GAを用いて解くことができ，自動的に準最適な節点の数と位置を同時に決めることができる．得られる節点の数と位置は，GAの性質から，大域的な準最適解である可能性が高い．本研究により，GAがスプライン関数の節点の決定問題に対して有効であることがわかった．

今後の課題としては，計算量の軽減，初期収束の回避，適切な制御パラメータの決定方法などがある．これらの問題は，本論文で提案した方法に限らず，GA全般の課題である．なお，多次元問題への拡張，並列計算への適合性などについての研究結果は，別の機会に報告する．



（a）あてはめの結果



（b）適応度

図3　複合有理式曲線データの結果

参考文献

1)市田浩三，吉本富士市：スプライン関数とその応用，p.220，教育出版，東京，1979.

2)Dierckx, P.: *Curve and Surface Fitting with Splines*, p.285, Clarendon Press-Oxford, 1993.

3)Cox, M. G.:A survey of numerical methods for data and function approximation, In *The State of Art in Numerical Analysis* (Ed. by D. A. H. Jacobs), Academic Press, New York, pp. 627-668, 1977.

4)Jupp, D. L. B.: Approximation to data by splines with free knots, *SIAM J. Numer. Anal.*, pp. 328-343, 1978.

5)Anthony, H. M., Cox, M. G. and Harris, P. M. : The use of local polynomial approximations in a knot-placement strategy for least-squares spline fitting, *NPL Report*, DITC 148/89, 1989.

6)吉本富士市：ファジィ概念を用いたスプライン関数の節点の決定，情報処理学会論文誌，第35巻第9号，pp.1682-1690，1994.

7)Rice, J. R.: *Numerical Methods, Software, and Analysis*, Second Ed., p.720, Academic Press, San Diego, 1993.

8)Goldberg, D. E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, p. 412, Addison-Wesley, 1989.

9)坂和正敏，田中雅博：遺伝的アルゴリズム，p.203，朝倉書店，東京，1995.

10)樋口哲也，北野宏明：遺伝的アルゴリズムとその応用，情報処理，第34巻第7号，pp.871-883,1993.

11)伊庭斉志：遺伝的アルゴリズムの基礎－GAの謎を解く－，p.254，オーム社，東京，1994.

12)Akaike, H.: A new look at the statistical model identification, *IEEE Trans. Automatic Control*, Vol. AC-19, pp. 716-723, 1974.

13)Manela, M., Thornhill, N. and Campbell, J. A. : Fitting spline functions to noisy data using a genetic algorithm, *Proc. of the Fifth Int. Conf. on Genetic Algorithms*, pp. 549-556, Morgan Kaufmann, 1993.

14)Markus, A., Renner, G. and Vancza, J.: Spline interpolation with genetic algorithms, *Proc. 1997 Int. Conf. on Shape Modeling and Applications*, IEEE Computer Society Press, pp. 47-54, 1997.

15)de Boor, C.: *A Practical Guide to Splines*, p. 392, Springer-Verlag, New York, 1978.