

進化論的手法を用いた金融データの予測

Financial Data Prediction by means of Evolutionary Computation

伊庭 斉志*

Hitoshi IBA

要 旨： 進化型計算を用いた金融データ予測の研究が近年盛んに行われている。本稿では、遺伝的プログラミング (GP, Genetic Programming) とその拡張手法である STROGANOFF を用いて、日経平均株価とオプション価格を予測する研究について説明する。実験の結果、GP と STROGANOFF による予測はおおむね両方のデータにおいて安定した成績を残すことができた。ニューラルネットワークとの比較実験では、ニューラルネットによる探索には山登り法の性質が良く出ているが、最良値は GP 系の手法より低く局所値への陥りやすさがしばしば観察された。

キーワード： 進化型計算, 遺伝的プログラミング, 金融データ予測, オプション価格, 日経平均株価

Abstract : There have been several applications of Evolutionary Computation (EC) to the financial problems, such as portfolio optimization, bankruptcy prediction, financial forecasting, fraud detection and scheduling. This paper presents the application of Genetic Programming (GP) to the prediction of the price data in Japanese stock market. The goal of this task is to choose the best stocks when making an investment and to decide when and how many stocks to sell or buy. We describe how successfully GP and its variant, i.e., STROGANOFF, are employed to predicting the stock data so as to gain the high profit. The comparative experiments are conducted with neural networks to show the effectiveness of the GP-based approach.

1 はじめに

インターネット上での証券取引の普及などにより多くの個人投資家が生まれている。一日のうちのわずかな株価変動を通して (空売りや空買いなどの) 信用取引により莫大な儲けを出すデイトレーダーも増えてきている。そのため株価や為替などの金融データ予測の実用的価値が高くなり、システム同定のアプローチはより重要な課題になりつつある。しかしながら、金融データは非線形で非常に複雑な動きをするため正確な予測は容易ではない。

金融データのシステム同定にはさまざまな人工知能の手法が使われている。その中でも遺伝的アルゴリズム (Genetic Algorithm, GA) や遺伝的プログラミング (Genetic Programming, 遺伝的プ

ログラミング)[2] などの進化論的計算手法は、探索性能、並列高速化、頑強性などにおいて優れている点から、頻繁に利用されるに至っている [5]。例えばオプション価格を扱った初期の GP 研究 [4] では、S&P 500 index (アメリカの主要業種を代表する 500 銘柄の時価総額で算出される株価指数) のオプションを用いて、ブラック・ショールズの式よりも一般的な関数でのオプション価格の導出を目指している。また、ブラック・ショールズの式で生成した仮想市場と jump-diffusion process で生成した仮想市場に GP を適用し、頑強な売買ルールの構築を試みる研究 [6] もある。

本稿では GP とその拡張である STROGANOFF [10] を金融データのシステム同定問題に応用する。第一の金融データとして

* 東京大学大学院 新領域創成科学研究科基盤情報学専攻 〒277-8563 千葉県柏市柏の葉 5-1-5 Tel&Fax: (04)7136-3875, email: iba@iba.k.u-tokyo.ac.jp

日経平均株価を用いて株価動向を予測し、一株売買のロジックにより売買を行ってどのくらいの利益があげられるかを調べる。次に日経平均株価オプションを用いて価格の予測式を生成し、ブラック・ショールズ式との比較を行う。さらにGP手法の有効性を調べるためにニューラルネットワークによる予測との比較実験を行う。

2 GPの株価予想への応用

株価指数とは株式市場全体の株価動向を把握するための指数である。東京証券取引場の第一部上場銘柄のうち市場を代表する225銘柄を対象に平均をとって算出される日経225や、第一部上場全銘柄の現在の時価総額を基準時価総額（1968年1月4日の時価総額）で割ったものを100倍して算出されるTOPIX等がその代表的なものである。図1は1993年の上半期における日経225の1分毎のデータである（合計33177分）。この時系列データを見ると、多少の振動はするが全体としては上り調子、下り調子の傾向が一定の期間続いている。例えば12000分位から18000分位の間は全体として上り調子であり、28000分位から30000分位の間は下がり調子である。このような全体傾向を長期トレンドという。長期トレンドの予測はそれほど難しくないとされている。それに対して難しいのは、短期間に大幅に変動する株価の予測である。

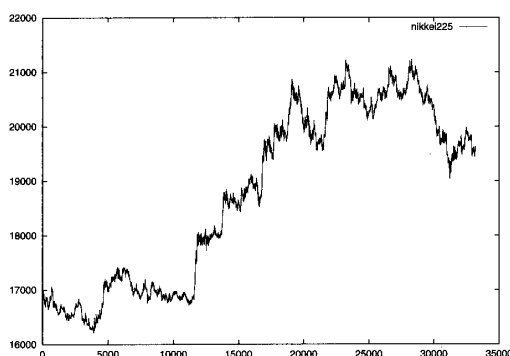


図1: 日経平均株価の推移

GPを用いて過去のデータをもとに未来の日経225の値を予測する近似関数を見つけ出してみよう。まず、1993年の上半期における日経225の1分毎のデータを0と1の間をとるように次式で正

規化する。

$$sdata = \frac{data - \min}{\max - \min} \quad (1)$$

ただし、正規化する前の日経225のデータをdata、正規化された後の日経225のデータをsdataとする。また正規化する前の日経225の最大値、最小値をそれぞれmax、minとしている。さらに正規化されたデータを訓練データ（最初の3000個）とテストデータ（残りの30177個）に分ける。これは過学習を避けて一般的なデータに適合する近似関数が得られるかをテストするためである。近似関数は訓練データ内での当てはめの成績が良くなるように進化していく。そして、訓練データ内で最も適合度の良い近似関数がテストデータで評価される。

GPシステムで用いた終端記号、非終端記号、近似関数の評価方法、および実行条件を以下で説明する。なお本稿ではspgpc1.1という汎用のGPシステムを実験で使用した。そのパラメータの詳細や動作方法については[1]を参照してほしい。

GPの終端記号としては y_i , ave_i , m_i , d_i , v_i , r_i , $\&$ を用意した。ここで y_i は*i*分だけ前の日経225の値を表している。つまり、時刻*t*分における日経225の値を $x(t)$ とすると

$$y_i = x(t - i) \quad (2)$$

である。 ave_i は日経225の10分毎の平均値を表しており、次式で計算される。

$$ave_i = \frac{\sum_{k=1}^{10} x(t - 10 * (i - 1) - k)}{10} \quad (3)$$

m_i は日経225の10分毎の分散を表しており、次式で計算される。

$$m_i = \frac{\sum_{k=1}^{10} (x(t - 10 * (i - 1) - k) - ave_i)^2}{10} \quad (4)$$

また、 d_i , v_i , r_i はそれぞれ

$$d_i = x(t - i) - x(t - i - 1) \quad (5)$$

$$v_i = |x(t - i) - x(t - i - 1)| \quad (6)$$

$$r_i = \frac{x(t - i) - x(t - i - 1)}{x(t - i - 1)} \quad (7)$$

で計算される値である。 $\&$ はランダムに定数値を発生させる終端記号である。また、非終端記号としては $\{+, -, *, \%\}$ を用いた。 $+$, $-$, $*$, $\%$ は二つの引数を持ち四則演算を計算する。

訓練データにおける近似関数の評価方法としては、実際の値と近似関数から得られる予測値との

平均二乗誤差を用いた。平均二乗誤差は次式で与えられる。

$$E = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_i)^2 \quad (8)$$

ここで y_i は正しい出力で、 \bar{y}_i は近似関数の予測値である。ただし $i = 1, 2, \dots, N$ とする (N はデータの数)。この値はそのまま適合度として扱われる。従って適合度は 0 に近いほど良い。

テストデータにおける近似関数の評価方法として株売買シミュレーションを行う。このシミュレーションでは、100 万円を元手に近似関数から得られる予測値に従って株の売買をし、日経 225 の増減の予測の的中率と売買利益を求める。つまり、一分後の予測値が現在の株価より上昇(下降)しており、かつ現在株を保有していない(している)なら、株を買う(売る)というロジックである。予測の的中率は以下のように定義される。

$$\text{hit} = \frac{N_{u,u} + N_{d,d}}{N_{u,u} + N_{u,d} + N_{d,u} + N_{d,d}} \quad (9)$$

$$= \frac{N_{u,u} + N_{d,d}}{30,177}, \quad (10)$$

ここで $N_{u,u}$ は予測値が株価の値上がりを示し、実際に株価が値上がりした回数である。また、 $N_{d,u}$ は予測値が株価の値上がりを示し、実際には株価が値下がりした回数である。他の変数も同様に定義される。予測回数の合計は 30,177 であり、テストデータの数に等しい。売買利益は、株の売値から買値を引いた額である。なお、簡単のため売買では一株しか扱わず、市場への自分の売買の影響は考えない。また売買の手数料も考慮していない。この株売買のロジックをまとめると、次のようになる。ただし時刻 t での株価を $Pr(t)$ とする。

Step1 はじめに手持ちの元手 BG を定義する (例えば 1,000,000 円)。タイムステップ t を 3000 とする。これはテストフェーズの開始時点である。株保有フラグ ST を 0 にセットする。

Step2 GP の木構造を用いて出力 (日経 225 の予測値) を求める。この予測値を $\tilde{Pr}(t)$ とする。

Step3 もし $Pr(t-1) < \tilde{Pr}(t)$ かつ $ST = 0$ ならば、株を買う。そして ST を 1 にセットする。

Step4 さもなければ、もし $Pr(t-1) > \tilde{Pr}(t)$ かつ $ST = 1$ ならば株を売る。そして ST を 0 にセットする。

Step5 もし $ST = 1$ ならば、 $BG := BG + Pr(t) - Pr(t-1)$ とする。

Step6 $BG < 0$ なら 0 を返して終了する (損切り)。

Step7 もし $t < 33,177$ ならば、 $t := t+1$ として **Step2** に戻る。さもなければ売買利益 (つまり $BG - 1,000,000$ 円) を返して終了する。

株保有フラグ ST は、現時点での株の保有状態を示す変数である。つまり $ST = 0$ なら株を保有しておらず、 $ST = 1$ なら株を保持していることを示す。**Step5** では、全資産を新しく得られた株価 (の実測値) をもとにして求めている。**Step6** は破産してシミュレーションが終了する条件である。

株価指数予測を行うのに A~G の 7 通りの終端記号のもとで GP を実行した。

- 条件 A: $\{y_1, \dots, y_{10}, \mathcal{R}\}$
- 条件 B: $\{\text{ave1}, \dots, \text{ave10}, \mathcal{R}\}$
- 条件 C: $\{m_1, \dots, m_{10}, \mathcal{R}\}$
- 条件 D: $\{\text{ave1}, \dots, \text{ave10}, m_1, \dots, m_{10}, \mathcal{R}\}$
- 条件 E: $\{y_1, \dots, y_{10}, \text{ave1}, \dots, \text{ave10}, m_1, \dots, m_{10}, \mathcal{R}\}$
- 条件 F: $\{d_1, \dots, d_{10}, \mathcal{R}\}$
- 条件 G: $\{v_1, \dots, v_{10}, r_1, \dots, r_{10}, \mathcal{R}\}$

A から E に対しては予測するのは現在の値であり、F と G では現在の値と 1 分前の値の差である。

上に述べた各条件において GP を 10 回実行した。どの条件でも、集団数 1000、世代数 100 のパラメータを用いた (他のパラメータは表 1 を参照)。その実行結果を表 2 に示す。

表 2 から、わずかな的中率の差が利益の大きな差につながることが見てとれる。ランダムな予測を行った場合の的中率は 50% であることに注意しよう。また、訓練データ内での平均二乗誤差が小さいとテストデータにおける的中率や利益も良い。このため GP による方法は過学習の問題をある程度回避していると考えられる。

条件 B, C, D, E では予測の平均的中率と最良的中率の両方とも 50% 程度であり、利益もほとんどあがっていない。一方、条件 A, F, G では平均的中率は 55% 以上、最良的中率においては 60% にも達し、その結果最適利益も 3 万円近くになっている。

表 1: sgpc1.1 の GP パラメータ (パラメータの詳細は [1] を参照)

max_generation	100	max_depth_after_crossover	17
population_size	1000	max_depth_for_new_trees	6
steady_state	0	max_mutant_depth	4
grow_method	GROW	crossover_any_pt_fraction	0.2
tournament_K	6	crossover_func_pt_fraction	0.7
selection_method	TOURNAMENT	fitness_prop_repro_fraction	0.1
function set	{+, -, *, %, sin, cos, exp}		

表 2: GP の実行結果

条件	訓練データ内での 平均二乗誤差	的中率 (%)		利益 (円)	
		平均値	最良値	平均値	最良値
A	1.79e-06	55.02	62.78	12411.01	31256.06
B	1.22e-05	47.47	48.17	-4093.22	-2341.50
C	5.82e-04	50.42	51.00	127.03	305.13
D	1.28e-05	41.09	51.64	-19727.52	-3811.19
E	1.94e-06	49.45	50.34	-806.33	1061.38
F	1.64e-06	55.02	62.49	14276.06	30722.81
G	1.80e-06	61.38	62.56	28942.03	30896.56

特に条件 G では平均的中率 61.38% という好成績である。

条件 A のもとで得られた GP の予測木もある程度の利益をあげている。しかしながら、この成績は必ずしも安定したものではなく、平均の収益を見ればかなり劣っていることが分かる。条件 A は E の部分集合なので、E で得られた成績が A より劣るのは矛盾していると思われるかもしれない。これは、GP による探索効率が終端記号の大きさにより左右されることから説明される。つまり適切な終端記号を選ぶことが GP の探索では非常に重要である。条件 A と G で獲得された木構造の例を表 3 に示す。

ここで GP の実行結果を簡単にまとめると

- 平均や分散の終端記号は結果がマイナスになるなどあまり有効ではなかった
- 全体的に平均値と最良値の差が大きく結果が不安定であった
- 木の記述長が長くなり利益が出ないパターンが多かった
- 条件 E の終端記号が有効であった

条件 E の結果から分かるように、訓練例の予測の好成績 (つまり MSE 値の小ささ) は必ずしも

表 3: GP で獲得された木構造

条件	木の表現
A:	(+ y1 (* y7 (* (* (* (* (* (- y1 y3) (% (* y4 y10) (* (+ y6 y10) (* (+ y8 y4) (- (+ y6 y10) (- y9 y6)))))) (% (* y3 y10) (* y2 y9))) (% (* y4 y10) (* (* (- y9 y9) (% (+ y8 y4) (% (- (- (* y2 y9) (- y9 y6)) (* (* (- y9 y6) (+ y6 y10)) (* y3 y10))) (* y10 (+ y6 y10))) (- y9 y6))) (* y4 y10))) (% (* y2 y9) (* (- (+ y6 y10) (- y9 y6)) (* (+ y8 y4) (+ y8 y4)))) (% (* y3 y10) (* y2 y9))) (% (* y4 y10) (* (* (- y9 y9) (% (* (- y9 y9) (% (- y9 y6) (* y3 y10))) (% (- (- (* y3 y10) (* y4 y10)) (* (* (* y3 y10) (+ y8 y4) (- (* (- y9 y9) (% (- y9 y6) (+ y6 y10)) (- y9 y6)))) (* y3 y10) (- y9 y6)))) (* y4 y10)))) (% (* y4 y10) (* y2 y9))))
G:	(+ (+ (+ (+ (* r1 v1) (* r1 v5)) (* r1 v5)) (+ (* r1 v5) (+ (* r1 v1) (* v4 r2)))) (+ (* r1 v1) (* r1 v5)))

高収益を保証しない。効果的な売買ルールでは、MSE 値を小さくすることも大切であるが、むしろ上がり傾向 (下がり傾向) を的確に予測することが重要になる。とくに今まで上がっていた (下がっていた) のが急に下がる (上がる) ときに、この傾向の変化を知ることが高収益につながる。差に基づく方法 (条件 F, G) では、こうした定性的な傾向の予測がある程度可能であると思われる。

3 STROGANOFF による株価データ予測

従来の GP の問題点として結果が不安定なことがあげられる。GP の探索に GMDH (Group Method of Data Handling, [9]) と呼ばれる多重回帰分析の統計的手法を追加することで、より頑強で安定した探索を目指す。GMDH と GP を統合したシステムとして STROGANOFF (STructured Representation On Genetic Algorithms for Non-linear Function Fitting) が提案されている [7][10]。STROGANOFF では適合度計算には MDL 基準を用いる。これにより GP で問題となっていた過学習やブロート (木の過大な成長) を回避することができる。

以下では STROGANOFF による株価データの予測実験について説明する。STROGANOFF の実行に用いる終端ノードは GP と同じものとする。STROGANOFF のパラメータは集団数が 100 であることを除いて GP と同じである (表 1)。また STROGANOFF 固有のパラメータとして MDL における重み (WEIGHT) を何種類か変えて実行する。以上の条件で STROGANOFF を 10 回ずつ実行した結果を表 4 に示す。ただし売買ルールは前述の GP と同じものを用いた。

表 4 から STROGANOFF では条件 A で良い結果が出ているのがわかる。条件 A で最良の結果を得た木構造を表 5 と図 2 に示す。図 3 と図 4 はこの個体の予測値と利益のグラフである。図 3 では緑の線 (上方の線) が予測値、赤の線 (下方の線) が実際の値である。図 4 では緑の線 (下方の線) が予測により売買をして得た利益、赤の線 (上方の線) が最適条件 (予測が 100% 成功したときの利益) である。

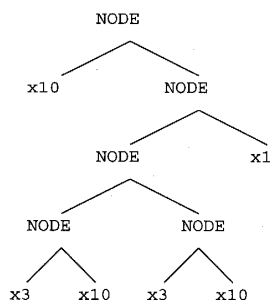


図 2: 条件 A の最良個体の木構造

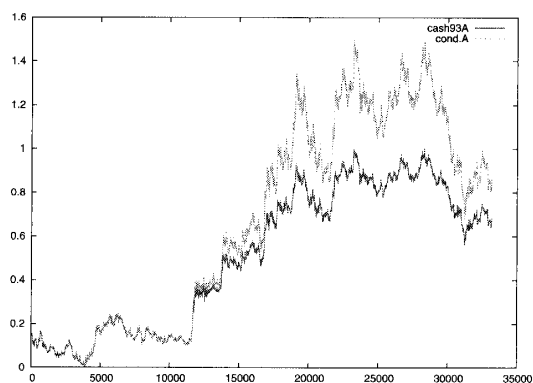


図 3: 最良個体の予測値と実際の値の比較

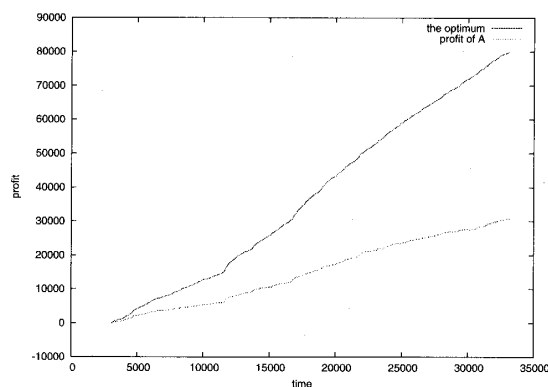


図 4: 条件 A の最良個体と最適条件の比較

以上の実験結果から、GP の場合とは異なり STROGANOFF では過去の値や平均の終端記号が予測に有効に働いていたことが分かる。また、STROGANOFF による予測は GP に比べると最良値はほとんど一定であるが、どの試行でも非常に安定しており平均がマイナスになることはなかった。STROGANOFF の訓練データについての平均二乗誤差は GP に比べて大きい、利益の最良値はほとんど同じで平均は高い。その原因は、STROGANOFF はより広範なデータに適合するために特定のデータに対する精度が落ちたからと考えられる。以上から、訓練データとして異なる期間や数種類の時系列を使えばさらに頑強な予測が可能になると考えられる。

表 4: STROGANOFF の実行結果

条件	訓練データ内での 平均二乗誤差	的中率 (%)		利益 (円)	
		平均値	最良値	平均値	最良値
A	9.37E-06	62.2	62.3	30516	30823
B	1.25E-05	57.3	57.7	18594	19194
C	6.57E-04	50.0	58.2	841	4044
D	1.25E-05	57.2	57.7	18390	19194
E	7.25E-04	51.2	51.3	5785	6071

表 5: 条件 A の最良個体の S 式と係数

条件	木の表現	NODE の係数
A	(NODE139224304 X10 (NODE145246112 (NODE145246144 (NODE145248224 X3 X10) (NODE157573488 X3 X10)) X1))	2.848737e-03 -8.345704e-03 9.464328e-01 6.843502e-02 3.943218e-02 2.024200e-01

4 ニューラルネットワークとの比較

GP と STROGANOFF の有効性の確認のためにニューラルネットワークで株価の予測実験を試みる。以下で述べるニューラルネットを使った実験は、WWW 上の”Neural Networks at your Fingertips”[11]で入手可能なプログラムをもとに実装した。実行条件を表6に示す。ニューラルネットワークの入力は GP と STROGANOFF の条件 A, すなわち $x(t-1), x(t-2), \dots, x(t-10)$ の 10 個を用いた。また中間層のユニット数を 5,10,15 の三つにわけ各々 10 回ずつ実行した。その結果を表7に示す。この表から、ニューラルネットによる予測は安定しているが、GP や STROGANOFF のような高収益は得られないことが分かる。この結果は、ニューラルネットによる局所的山登り探索の性質を顕著に示している。

表 6: ニューラルネットワークの実行条件 (パラメータの詳細は [11] を参照)

中間層の数	1
中間ノードの数	5,10,15
入力ノード数	10
出力ノード数	1
入力値	$x(t-1), x(t-2), \dots, x(t-10)$
出力値	$x(t)$
訓練サイクル数	100
エポック数	1000

5 オプション価格の予測

オプションはある品物を売ったり買ったりする権利を売買の対象にするものである。株券を扱う場合には、コールオプションは“ある日時にある価格で株を買う権利”となる。この“ある価格”のことを権利行使価格と言う。一般にオプションはレバレッジ (てこの作用) の効く商品である。レバレッジの効く商品には先物取引などハイリスク、ハイリターンのもが多いが、オプションはレバレッジが効くにもかかわらず損害が限定できるという利点がある。オプション取引はヘッジ (将来の商品価格変動の危険) の回避や投資にも使えるため、ますます重要性が認識されている。

日経平均株価オプションは日経平均株価指数を原資産とするオプションである。満期日に価値のあるオプションは自動的に権利行使され、日経平均株価指数と権利行使価格の差額を利益として受け取る。逆に満期日に価値の無いオプションは自動的に放棄される。日経平均株価オプションのデータはホームページで公開されている。例えば図5のように、日経平均株価、各限月のコールオプションとプットオプションの権利行使価格ごとのオプション価格の現在値や理論価格、権利行使価格などがまとめられている。オプション価格 (プレミアム) を算出する式で有名なものにブラック・ショールズの式があり、次のように表される。

$$P = S \cdot N\left(\frac{u}{\sigma\sqrt{x}} + \sigma\sqrt{x}\right) - Xe^{-rx} \cdot N\left(\frac{u}{\sigma\sqrt{x}}\right)$$

表 7: ニューラルネットワークの実行結果

隠れノード の数	平均 2 乗誤差 (訓練データ内)	的中率 (%)		利益 (円)	
		平均値	最良値	平均値	最良値
5	2.92e-06	58.2	60.2	23682	27586
10	2.70e-06	58.7	59.4	24725	26427
15	2.73e-06	58.3	59.5	23990	26245

The screenshot shows a software window titled 'オプション価格表' (Options Price Table). It displays two main tables: 'CALL' (Call Options) and 'PUT' (Put Options). Each table has columns for '株価' (Stock Price), '行使価' (Strike Price), '権利行使価' (Exercise Price), '期日' (Expiration Date), and '利益' (Profit). The data is organized in a grid format, showing how option prices change with different stock prices and strike prices.

図 5: オプション価格表のホームページの例

$$u = \log \frac{S}{X} + (r - \frac{\sigma^2}{2})x$$

ここで S は株価, X は権利行使価格, x はオプションの期間, σ はボラティリティ, r は非危険利子率, さらに $N(d)$ は標準正規分布 $N(0, 1)$ における $-\infty$ から d までの面積を表す. すなわち

$$N(d) = \int_{-\infty}^d \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz$$

で計算される. ブラック・ショールズ式の S 式での表現を表 8 に示す.

ブラック・ショールズの式でオプション価格が計算できるはずだが, この式から計算された理論価格は実際の市場価格とは異なる. これはブラック・ショールズの式を導出するために用いた種々の仮定 (株価がウィーナー過程に従うという条件やポートフォリオの定義など) に原因があると考えられる. そこで, GP を用いてより精度の高いオプション価格算出式を求めてみる.

表 8: ブラック・ショールズ式の S 式表現

```
(-(* S normal(+(%(+ log om (*(- rf (%(* hv hv 2)) x)) (* hv sqrt(x)))) (* hv sqrt(x)))) (*(% x exp(* rf x)) normal(%(+ log om (*(- rf (%(* hv hv 2)) x)) (* hv sqrt(x))))))
```

実験に用いるデータと条件を次のように設定する.

- 十分な流動性がある日経 225 オプションのデータ (2005 年 6 月の 624 個と 7 月の 508 個) を用いる.
- 6 月のデータを訓練に, 7 月のデータをテストに用いる.
- 取得するデータは, 株価, ボラティリティ, 権利行使価格, オプション価格, オプション期間とする.

表 9: 実験に用いる終端記号

S	株価
X	権利行使価格
S/X	株価/権利行使価格
τ	満期までの期間
rf	非危険利子率
σ	ボラティリティ

表 11: STROGANOFF のパラメータ

population_size	5000
crossover_any_pt_fraction	0.399
grow_method	GROW
crossover_func_pt_fraction	0.5
WEIGHT	0.2, 0.15, 0.10, 0.05

- 終端記号 (ニューラルネットワークでは入力) には表 9 の 6 個を使う。単純化のため非危険利子率は一定 (0.009) とする。

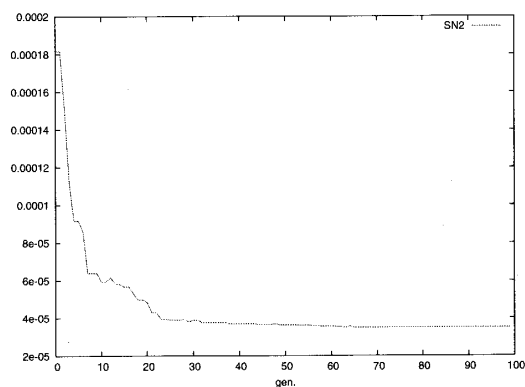
表 10 のパラメータで GP を 10 回実行した。ここで $\sqrt{}$ は平方根, $normal$ は累積標準正規分布関数を表している。STROGANOFF は表 11 のパラメータ (表に無いものは GP と同じ) で各 WEIGHT ごとに 10 回ずつを実行した。表 6 はニューラルネットワークのパラメータであり, 入力ノードは GP の終端記号と同じものとする。

以上の各条件での実行結果を表 12 に示す。表の値は平均二乗誤差を示している。また, HP はホームページ (図 5) に示される理論価格, BS はブラック・ショールズの式に基づく理論価格である。この表から STROGANOFF による実行結果が平均値, 最良値ともに一番低いのが分かる。ニューラルネットワークによる予測は平均値では GP に勝っているが, 最良値では GP の方が良くなった。また, 三つの手法 (GP, STROGANOFF, NN) は全てブラック・ショールズの式よりも誤差が小さくなった。

GP の実行結果は試行によって多少バラツキがあるが, 最良個体の平均誤差は訓練とテスト共に低い値となりニューラルネットワークより良くなった。これは, ニューラルネットワークが山登り的な探索のため局所解に陥ってしまうことが多いのに対して, GP の大域的探索は不安定ではあるが

しばしば飛び抜けて良い個体を見つけ出すことによる。しかしパラメータによっては GP では世代を経るに従って木のサイズが急に大きくなっていき, 探索の効率が悪くなることもあった。

STROGANOFF の実行結果は平均値, 最良値ともに三つの手法の中で一番良くなっており, GP の大域的探索に統計的手法を統合した特長が顕著に現れている。最良個体 (WEIGHT=0.05 のもの) の平均二乗誤差と MDL の推移 (世代毎) を図 6 と図 7 に示す。GP の結果と比べると, STROGANOFF の方が記述長は小さくなり成績も良かった。したがって, GP 木の過大な成長 (ブロー現象 [2]) をうまく抑えることができており, それにより効率の良い探索が行われ好成績が得られたと考えられる。また, WEIGHT を小さくしていくと訓練データでの成績は良くなっていくが, テストデータでの平均誤差が増えていく過学習の傾向が見られた。これは訓練データに特化した記述長の大きい木が選択され易くなるのが原因であろう。このように STROGANOFF では, WEIGHT の調整によって予測精度と汎用性のトレードオフを考慮したより頑強な探索が可能である。

図 6: 最良個体の S_N^2

6 おわりに

本稿では二種類の金融データの予測に対して進化論的手法 (GP と STROGANOFF) を適用した結果を説明した。GP による予測は結果が不安定な場合もあり, 木のサイズも大きくなりがちであった。一方で, 統計的手法により GP を拡張した

表 10: GP のパラメータ (パラメータの詳細は [1] を参照)

max_generation	100	max_depth_after_crossover	12
population_size	25000	max_depth_for_new_trees	6
steady_state	0	max_mutant_depth	4
grow_method	GROW	crossover_any_pt_fraction	0.3
tournament_K	6	crossover_func_pt_fraction	0.5
selection_method	TOURNAMENT	fitness_prop_repro_fraction	0.1
function set	+, -, ×, ÷, exp, log, sqrt, normal		

表 12: オプション予測の実行結果

	平均誤差 (訓練例)	平均誤差 (テスト例)	最良値 (訓練例)	最良値 (テスト例)
HP	26.9	54.1	-	-
BS	26.0	54.0	-	-
GP	24.3	34.1	19.0	20.8
STROGANOFF(0.2)	18.7	26.9	16.6	21.2
STROGANOFF(0.15)	17.9	27.1	16.5	23.9
STROGANOFF(0.1)	17.8	27.9	15.9	19.1
STROGANOFF(0.05)	17.0	26.5	16.0	17.6
NN(15)	19.7	28.8	18.5	22.9
NN(10)	20.0	29.3	19.0	26.8
NN(5)	24.6	32.9	19.1	25.8

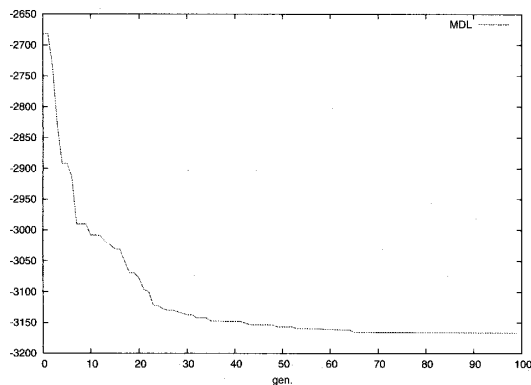


図 7: 最良個体の MDL

STROGANOFF は両方のデータにおいて安定した成績を残すことができた。またニューラルネットワークとの比較実験では、ニューラルネットでは山登り法の性質が良く出ているが、最良値は GP 系の手法より悪くなり、局所値への陥りやすさがしばしば観察された。以上の研究の詳細は [3][10] を参照してほしい。

筆者らのグループでは、GP と STROGANOFF を用いた実際運用のための株価予想システム (図

8) を構築中である。本稿で説明した GP を用いた時系列予測のソフトウェアは、筆者の研究室のウェブサイト <http://www.iba.k.u-tokyo.ac.jp/> から入手可能であるので、興味のある読者は参照してほしい。

参考文献

- [1] 伊庭斉志: "遺伝的プログラミング入門", 東京電気大学出版局, 1996.
- [2] 伊庭斉志: "遺伝的プログラミング入門", 東京大学出版会, 2001.
- [3] 大野勝彦, 伊庭斉志: "遺伝的プログラミングを用いた金融データ予測", 情報処理学会第 68 回全国大会 (学会創立 45 周年記念大会), 2006.
- [4] Chen, S.-H. and Lee, W.: "Option Pricing with Genetic Programming", in *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pp.32-37, 1998
- [5] Chen, S.-H.(ed.): "Evolutionary Computation in Economics and Finance", Springer, 2002.

