Vol.2009-ICS-156 No.1
2009/8/27

遺伝的プログラミングを用いたデータマイニング

野球における打者の最適な評価モデルの探索

福永圭佑[†1]　伊藤昭　寺田和憲

これまで野球というゲームの戦略の形成や選手の評価は，主に経験者のノウハウを重視して行われてきた．しかし近年の計算機の発展も手伝い，実際の試合で得られるデータを客観的に分析し野球というゲームを見直そうという試みが始まった．この試みはセイバーメトリクス (Society for American Baseball Research Metrics) と呼ばれ，実際にメジャーリーグにおけるチームの戦略として重視されている．本研究では，セイバーメトリクスへのアプローチの一例として遺伝的プログラミングを使用したバッターにおける最適な評価モデルの探索を目的としている．また，その結果得られた評価モデルと現在主流とされる評価モデルとを比較し，その有用性と問題点を提議する．

Data Mining with Genetic Programming

Searching the most suitable evaluation model of batters in MLB

Keisuke Fukunaga,[†1] Akira Ito
and Kazunori Terada

On baseball game, the know-how of experienced persons are considered inmortant in making strategies and evaluating players in the game. With the development of the computer, however, the baseball game is being reconsidered by objective analysis using data obtained in a real game, which is now known as SABR-metrics(Society for American Baseball Research Metrics) theory. Many of the Major League Baseball (MLB) team adopt this theory to make a strategy of the team. In this research, we attempt to find the most suitable evaluation model of the batter in MLB using data mining technique with genetic programming. We compare the model we found to that currently employed best in MLB, and discuss the merits and demerits of our model.

1. はじめに

野球は２つのチームが攻撃と守備を交互に繰り返して，結果的に相手チームより多くの得点を記録して勝つことを目的とした多人数ゲームである．野球が生み出された１８００年代から現在に至るまで，数多くの人々によってこのゲームに勝つための最適な戦略や戦術が考え出されてきた．特に野球の勝利条件は「試合終了時点で相手より得点が上回っていること」とされているため，より効率よく得点できるオーダーの作成の目安として選手の正確な評価モデルの作成は非常に重要であり，これまで多くの評価モデルが開発されてきた．

1.1 時代ごとの評価モデルの推移

黎明期（１８００年代）においては、選手を評価するにあたって実際の試合で記録された客観的なデータよりも実際に野球をプレイする人間のノウハウが重視されていた．１８５０年代前後にはイギリスのスポーツジャーナリストであるヘンリー・チャドウィックによって**ボックススコア**[⋆1] が発明され，さらに野球における最もポピュラーな評価モデルである打率 (Batting Average) が作り出されるなど，野球というゲームを統計的に捉える試みが始まりつつあった．その試みが本格的に始まったのは１９７０年代である．当時スポーツライターであったビル・ジェームズが独自に野球の試合のデータを集計し，確率統計学的な観点からそれらの数値の分析を行った．つまり，野球における戦略・戦術に統計学的根拠を持たせようとしたのである．これはセイバーメトリクス (Society for American Baseball Research Metrics) と呼ばれ，今でこそ野球理論の主流とされているが，それまでの野球の伝統的価値観をしばしば覆すものであった．また計算機の発展が著しい１９９０年以降においてはセイバーメトリクス理論の信憑性も増し，以前は保守的であったメジャーリーグの各チームも戦略の一環として重視するようになった．特に１９９７年にオークランド・アスレチックスの**ゼネラルマネージャー**[⋆2] に就任したビリー・ビーンがこの理論を徹底しチームを優勝に導いた．そのチーム運営戦略を紹介したノンフィクション「マネー・ボール」によって，セイバーメトリクスは日本でも一般的に認知されることとなる．

†1 岐阜大学大学院，工学研究科
Faculty of Engineering, Gifu University

⋆1 試合ごとの選手の成績データを表にして記録したもの．

⋆2 スポーツビジネスにおいてチーム運営及び選手補強などの総括を務める役職．

2. 本研究の目的

メジャーリーグにおける過去数年分のデータから遺伝的プログラミングを用いたデータマイニングによる分析を行い，最適な打者の評価モデルを探索・生成する．具体的には後に詳しく紹介する，現時点で最適な打者の評価モデルとされている OPS よりも性能が高いモデルの生成を目的とする．

2.1 メジャーリーグの基礎知識

実験に使用したデータは **SEAN LAHMAN'S BASEBALL ARCHIVE**[*3]より入手できる．ただ，メジャーリーグについてあまり詳しくない方はこれらのデータを見てもイメージが沸きにくいと思われるので，以下にメジャーリーグに関しての基礎知識を記しておく．現在メジャーリーグのチーム数はア・リーグ１４球団，ナ・リーグ１６球団の計３０球団から成る．ただしこれは１９９８年以降の事で，初期メジャーリーグ発足の１８７６年から１９７６年までのア・リーグ１２球団，ナ・リーグ１２球団の計２４球団から始まり，その後の幾度の編成によりチーム数を増やしてきた経緯がある．また，このデータには多くの欠損が存在しており，そのほとんどは１９５０年以前のデータである．特に犠打は１８９５年から，犠飛は１９５４年から等といったように途中から記録を付け始めたパラメータがいくつか存在している．これらの理由から今回の実験では，データの欠損がほとんど無い全３０チーム編成の時代を主に用いる．次に，今回使用するデータは 40-Man Roster という枠に登録された選手のみが記録されている．１チームは選手を最大４０人までベンチに登録することができ，この枠を 40-Man Roster と呼ぶ．また実際に試合に出場できるのはその内の２５人であり，この枠を 25-Man Roster と呼ぶ．もちろんこれらは野手と投手を合わせた人数であり，その内の約半分が野手と考えてよい．ただしデータを見ると１チーム４０人以上の野手のデータが存在しているが，これはシーズン途中にマイナーリーグの選手（40-Man Roster から溢れた選手）を多く入れ替えたためと考えられる．ちなみにメジャーリーグでは１シーズン１６２試合が基本である．ただし日本と異なり消化試合はキャンセルされることもあるため１６２試合に達しないチームもある．そのためチームごとに１，２試合分の差が存在する場合もあるが，それはほとんど影響が無いと見て無視している．ア・リーグとナ・リーグの違いとして，ア・リーグは**指名打者制**[*4]を採用しておりその分理論的には得点効率が良いと考えられるが，本研究ではそれらを一まとめに扱っている．

2.2 最適な評価モデルとは

まず打者の評価モデルを作成するにあたり，何をもって最適な評価モデルとするかを考える．前述したとおり野球の勝利条件とは「試合終了時点で相手より得点が上回っていること」である．そのため最適な評価モデルは，チームの得点に多く貢献する選手を高く評価するべきものと考えられる．ただし，野球というゲームは個人単位の活躍だけでは多くの得点を得ることはできない．チームが多くの得点を得るためには出塁した選手，そのランナーを先の塁に進めた選手，そして本塁に還した選手といった複数の選手による功績を考慮しなければならない．つまり個々の打者の評価モデルを作成するためにも，まずはチーム単位で考える必要がある．以上を踏まえ今回の研究では，最適な打者の評価モデルとはチーム単位の評価モデルとチームの１試合平均得点との相関係数が高いものであると定義しておく．

2.3 既存の評価モデル

古くから重視されてきた打者の評価モデルとしては，打率と**打点**[*5]が挙げられる．前者は打つ能力，後者はランナーを還す能力であり直感的にも分かりやすいといえる．

打率 = 安打/打数

しかしセイバーメトリクスの理論においてこれらの評価モデル，特に打点に関しては打者個人を評価するには不適切であるとしている．その根拠として，打点は選手個人の能力のみでなくチームメイトの能力に大きく影響されるためである．例えばある選手が打席に入るたびにどれだけランナーが出ているかというのは，前の打者の出塁率に左右される．仮に本当にチャンスに強い打者だったとしても，その打者の打席で多くのランナーがいなければ打点を得ることはできず，逆に本当はチャンスに強くない打者だったとしても打席に立つたび多くのランナーがいるのであればそれなりに多くの打点を得ることができる．このように打点は状況による揺らぎが大きすぎることが主な原因である．一方でセイバーメトリクスでは次のような評価モデルを重要視している．一つは出塁率 (On-Base Percentage) と呼ばれ，次式で表される．

出塁率 = (安打 + 四球 + 死球)/(打数 + 四球 + 死球 + 犠飛)

セイバーメトリクスの重視する打者の評価モデルと従来のそれとで最も異なるのは四死球の扱いである．従来の考えでは四死球は投手の能力に起因するものであって，打者の能力と

⋆3 URL は http://www.baseball1.com
⋆4 攻撃時に投手の代わりに打席に立つ打撃専門の選手で守備につくことはない．
⋆5 打者が安打などにより走者を本塁に還した場合，もしくは自身の本塁打によって本塁に還った場合記録される．

は独立であるとされていた．しかしセイバーメトリクスでは，過程はどうあれ結果として出塁することができる選手が得点に多く貢献するものとして高評価を与えている．次に長打率 (Slugging Percentage) と呼ばれるモデルは次式で表される．

長打率 = 塁打/打数

長打率は打率の問題点を改善したモデルである．その問題とはすなわち，打率は単打も本塁打も同じ 1 安打とカウントする点である．単打よりも本塁打を打てる選手の方がチームの得点に多く貢献するであろうことは直感的に理解できると思われる．その改善案として塁打という概念を導入しており次式で表される．

塁打 = 安打 + 二塁打 + 三塁打 × 2 + 本塁打 × 3

もしくは

塁打 = 単打 + 二塁打 × 2 + 三塁打 × 3 + 本塁打 × 4

このように長打率は安打の内容に重み付けをすることで長打を打てる選手を優位に評価できるモデルである．出塁率，長打率ともに得点との相関が比較的高い優秀なモデルではあるが，現在最も優れているとされる評価モデルは別に存在する．それが OPS(On-Base plus Slugging) と呼ばれるモデルである．その式は非常に単純で次式で表される．

OPS = 出塁率 + 長打率

ここで注目すべきは，出塁率と長打率は分母も分子もそれぞれ異なるため，単純に足し合わせたところで何ら意味を持たない値だという点である．にも関わらず，OPS はどの年においても得点との相関が非常に高く，また式も単純であるためメジャーリーグでは打者の優れた指標として公式に認められている．

2.4 データマイニングによる未知なるモデルの発見

何故 OPS のような，それ自体は何ら意味をなさないモデルが開発されたのか？その理由として，計算機の発達による膨大なデータの分析，つまりデータマイニングが可能になったことが挙げられる．データマイニングは人間が考え付かないような最適解を発見することができる．そのためデータマイニングの方法によっては OPS を越える評価モデルを生成することも可能ではないかと考え今回の実験を行うに至った．

3. 遺伝的プログラミング

遺伝的プログラミング (Genetic Programming,GP) は，生物の進化のメカニズムを基にした進化的計算方法である遺伝的アルゴリズム (Genetic Algorithm,GA) の拡張版である．GA が遺伝子データを配列で表現していたのに対し，GP は木構造にすることによって GA では扱えなかった数式やプログラム的手法を表現できる．例えば図 1 のような構造を持つ．

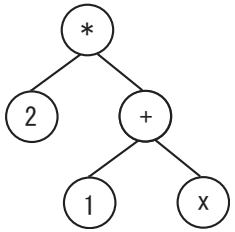

図 1　木構造

木構造においてそれぞれの要素をノードと呼び，その個体の遺伝子を構成する．ノードは終端ノードと非終端ノードに分類され，図 1 の木構造における終端ノードは 1,2 と x、非終端ノードは*と+である．つまり図 1 の木は 2*(1+x) という数式を表している．GP ではこのようなノードの組み合わせによって得られる遺伝子の木（エージェント）を一定数用意し，それらがある環境（解決したい問題）にどれくらいマッチしているかを適応度関数によって判別する．その結果環境に合った優秀な遺伝子を選び出し，交叉 (crossover) や突然変異 (mutation) といった遺伝的オペレータを行うことでより多様性のある遺伝子を生成し，それらを次世代に残す．これらを一定の世代繰り返すことで，その環境での最適な遺伝子 (戦略) を探索，生成することができる．本研究で用いた GP システムの流れを図 2 に示す．

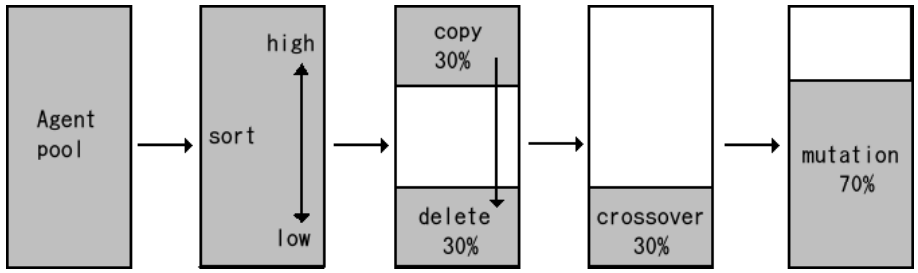

図 2　GP のフローチャート

表 1　オリジナルデータのフォーマット

選手名	年	チーム名	リーグ名	出場試合数	打席数	安打数	二塁打数	…	併殺打数
x_1	x_2	x_3	x_4	p_1	p_2	p_3	p_4	…	p_{17}

表 2　実験に用いるフォーマット

チーム名	年	チーム打席数	チーム安打数	チーム二塁打数	…	チーム併殺打数
T_j	i	$Y_iT_jP_1$	$Y_iT_jP_2$	$Y_iT_jP_3$	…	$Y_iT_jP_{14}$

表 3　打撃パラメータ

P_k	node	meaning		P_k	node	meaning
P_1	AB	打席		P_8	BB	四球
P_2	H	安打		P_9	SO	三振
P_3	twoB	二塁打		P_{10}	IBB	敬遠
P_4	thrB	三塁打		P_{11}	HBP	死球
P_5	HR	本塁打		P_{12}	SH	犠打
P_6	SB	盗塁		P_{13}	SF	犠飛
P_7	CS	盗塁死		P_{14}	GIDP	併殺打

4. 実験の概要

今回の実験で使用するオリジナルデータは図 1 のように，選手名 x_1，年 x_2，チーム名 x_3，リーグ名 x_4 によってパラメータ $p_1(x_1,x_2,x_3,x_4)$，…，$p_{17}(x_1,x_2,x_3,x_4)$ を特定するようなフォーマットである．これから図 2 のようなフォーマットを作成する．[⋆6]

今回の実験で使用するデータや関数について以下のように定義する．

- Y_i：i 年のデータ
- T_j $\{1 \leqq j \leqq 30\}$：チーム名（全３０チーム）
- P_k $\{1 \leqq k \leqq 14\}$：打撃パラメータ（表 3 に P_k に対応する終端ノードとその意味の内訳を示す．）
- S_{T_j}：チーム T_j の 1 試合平均得点 (Average Run：Y_i における T_j の 1 年間の得点をゲーム数で割った値)

ここで，Y_i の年におけるチーム T_j に所属する全選手の P_k の和を，チーム打撃パラメータ $Y_iT_jP_k$ とする．例えば，２００７年における T_1 の合計安打数は $Y_{2007}T_1P_2$ と表される．

⋆6 この際オリジナルデータのパラメータから試合数，得点，打点の３つを削除する．試合数は打撃との関連が薄く，得点と打点は前述したように個人の評価には不向きであると考えているからである．

4.1 従来の評価モデルと 1 試合平均得点との相関

新しい評価モデルを生成する前に，まず従来の評価モデルと 1 試合平均得点との相関を確認する．対象の評価モデルは打率（BA），出塁率（OBP），長打率（SLG），OPS の 4 つとする．Y_{2007} のデータを用いて以下の手順を行う．

(1) 対象の評価モデル M の各パラメータ P_k に対応するチーム打撃パラメータ $Y_{2007}T_jP_k$ を代入し，チーム T_j の評価モデル M に対する値 $V(T_j,M)$ を求める．

(2) $V(T_j,M)$ と 1 試合平均得点 S_{T_j} を $(x_j,y_j)=(V(T_j,M),S_{T_j})(1 \leqq j \leqq 30)$ のように対応付ける．これを全ての j に対して行う．

(3) 得られた３０個の $(x_j,y_j)=(V(T_j,M),S_{T_j})$ について，横軸を x，縦軸を y としてプロットし，その相関係数を $C(M)$ とする．

例えば，T_1 の打率は $V(T_1,\mathrm{BA})=Y_{2007}T_1P_2/Y_{2007}T_1P_1$ と表される．その結果，打率（BA），出塁率（OBP），長打率（SLG），OPS に対してそれぞれ図 3，図 4，図 5，図 6 のグラフを得た．

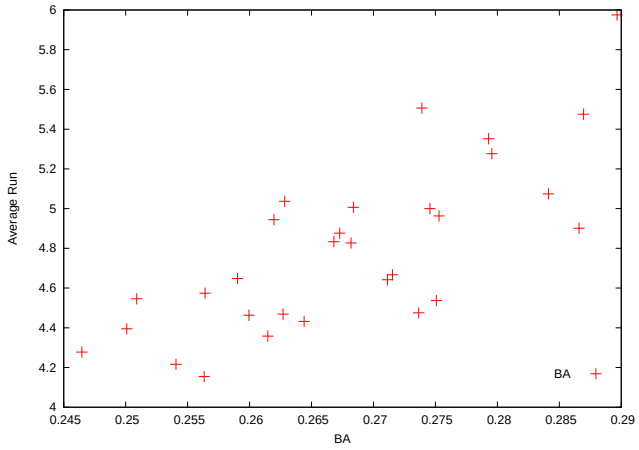

図 3　打率 $C(\mathrm{BA})=0.763$

まず，４つの評価モデルのうち最も歴史が長くかつポピュラーな打率に関して見ると，相関係数は 0.763 と確かに正の相関を持ってはいるが決して高いとはいえない．(図 3) 次に，出塁率（OBP）と長打率（SLG）の相関係数に注目する．(図 4，図 5) 前者では 0.874，後

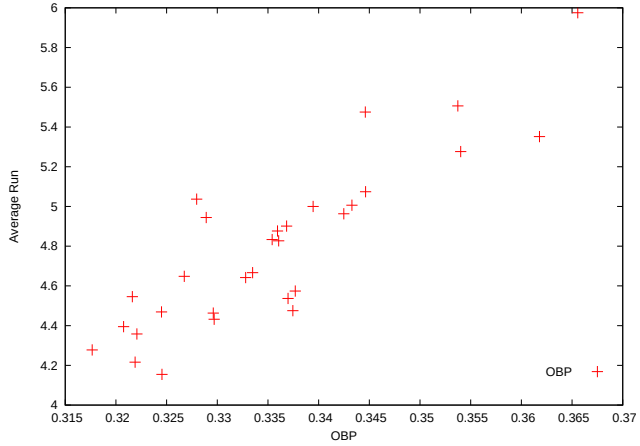

図 4 出塁率 C(OBP)=0.874

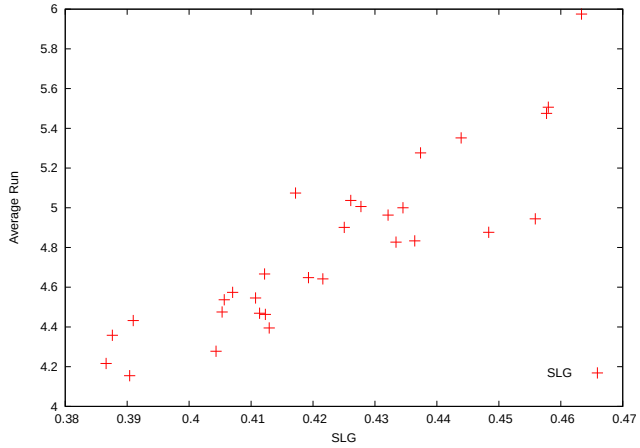

図 5 長打率 C(SLG)=0.885

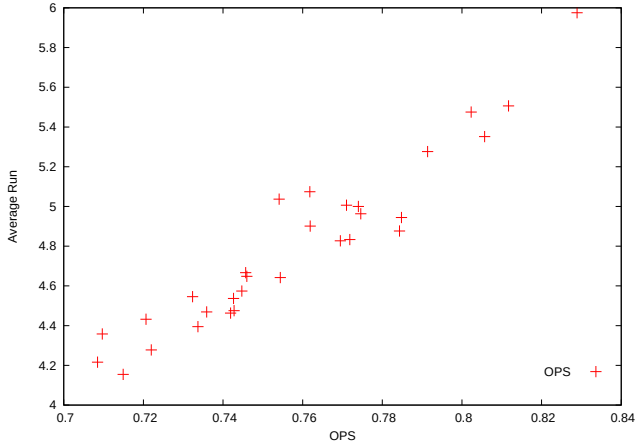

図 6 OPS C(OPS)=0.951

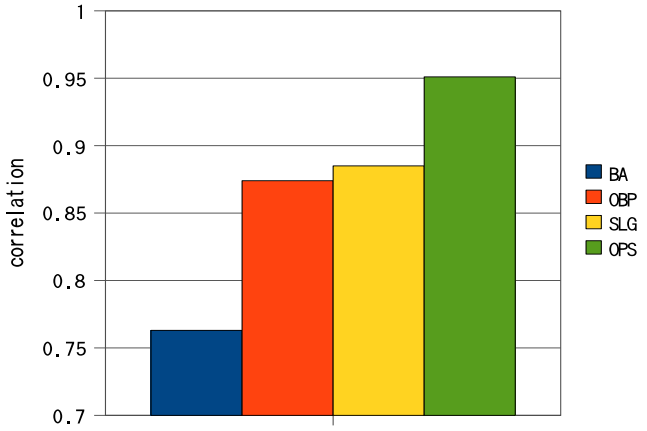

図 7 各モデルの相関係数の比較

者では 0.885 といったように打率のそれと比べて 0.1 以上増えており，明らかに強い相関を持っていることが分かる．さらに OPS は期待通り 0.951 という非常に高い相関係数をはじき出している．(図 6) これらを比較したものを図 7 に示す．

以上の結果，前述したセイバーメトリクス理論の整合性が一定は示されたといえる．

4.2 GP による最適な評価モデルの生成

ここまでの結果を踏まえた上で，本研究の主眼となる GP による最適な評価モデルの生

成を行う．GP では使用するノードや各初期値の設定が非常に重要である．それらを以下に記す．

- 終端ノード：表 1 の１４個，非終端ノード：加算，減算，乗算，除算の４個
- 進化させる世代数：1000
- 初期エージェント数：100
- 突然変異率：0.2
- 木の長さに対するコスト：0.0002
- 木の最大長：1000

ここで以降 GP の生成する評価モデルを GPC(Genetic Programming's Creation) と呼称する．

(1) ランダムな構造の GPC を持つ初期エージェントを１００個生成する．それらのエージェントが持つモデルをそれぞれ $GPC_n(1 \leqq n \leqq 100)$ とする．

(2) 前の実験と同じく，GPC_n にそれぞれの P_k に対応するチーム打撃パラメータ $Y_{2007}T_jP_k$ を代入し，チーム T_j の GPC_n の値 $V(T_j,GPC_n)$ を求める．これを全ての T_j に対して行い，相関係数 $C(GPC_n)$ を求める．

(3) これらを全ての n に対して行い，求まった $C(GPC_n)$ の値を適合度として GP を展開する．(参照：図 2) この場合 $C(GPC_n)$ の値が高いほど次世代に生き残りやすい．

(4) (2),(3) を 1 世代として GP を展開する．

実験の結果，１０００世代の GP において相関係数 0.969 という非常に高い相関を持つ個体を発見した．(図 8) GP の推移を見ると，４０世代ほどで相関係数 0.9 を越え，５００世代にはほぼ解が収束していることが分かる．(図 9) １０００世代目で最も高い相関係数を持つ GPC に注目し，それを従来の評価モデルと比較したところ打率や出塁率，長打率はもちろん，OPS をも越える評価モデルであることが分かる．(図 10)

4.2.1 GPC の木構造の例

図 8 で示した GPC の構造式を以下に示す．

(+ (+ H thrB) (+ (+ (+ (+ (+ (+ IBB (+ thrB BB)) HR) (/ AB IBB)) H) thrB) (+ HR (+ (+ H thrB) (+ HBP (- HR AB))))) twoB))

これを整理すると以下のようになる．

$$3 \times H + twoB + 4 \times thrB + 3 \times HR + BB + HBP + IBB - AB + (AB/IBB)$$

この式を見ると，安打系のパラメータ（H,twoB,thrB,HR）と出塁系のパラメータ（BB,HBP,IBB）がその大半を占めており，それなりに納得のできる式であるといえる．

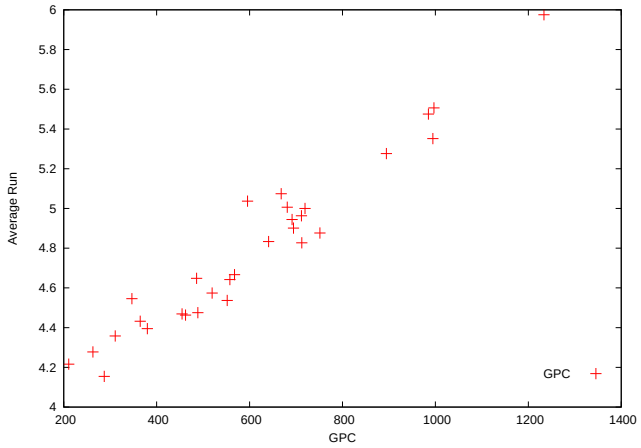

図 8 GPC C(GPC)=0.969

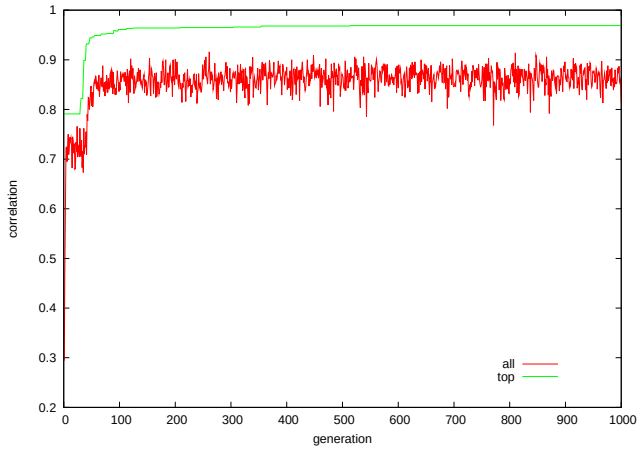

図 9 GPC の相関係数の推移

4.3 ロバスト ネスな評価モデル

特定のデータにおいて OPS を上回る評価モデル GPC の生成に成功したが，それだけで OPS を越える評価モデルかといえば答えは否である．どの年のデータを適用しても満遍な

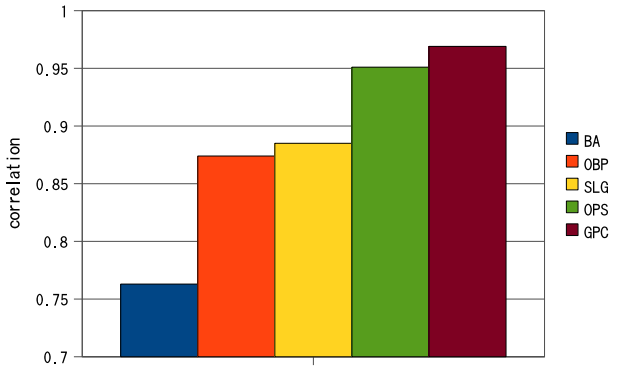

図 10 GPC と各モデルの相関係数の比較

く高い相関を得ることができて，初めて優れた評価モデルといえる．このような環境の変化への適応をロバストネス（頑強性）と呼び，機械学習においては重要な概念である．そこで GPC がロバストネスな評価モデルであるかの確認を以下の手順で行う．

(1) Y_{2000} から Y_{2007} までの８年分のデータを用意する．
(2) Y_i のデータから１０個の初期乱数を用いた**並列 GP 処理**[⋆7]を行う．(生成方法は前の実験と同様)
(3) その内最も相関係数が高い GPC を S_i とする．
(4) S_i に Y_i 以外の年のデータをそれぞれ適応し，それらの相関係数を求める．
(5) (2),(3),(4) を全ての Y_i に関して行う．

図 11 は S_i に加え，打率，出塁率，長打率，OPS に関しても調べた結果である．これを見ると，ほとんどの S_i はどの年のデータを用いた場合においても安定して相関係数 0.9 を越えていることが分かる．また従来のモデルに関して，OPS では同じく相関係数 0.9 を越えているが，その他のモデルは年ごとに相関係数が激しく上下している．また，これらモデルの８年分の相関係数の平均値を図 12 にまとめた．この結果から，GP で生成された GPC は十分にロバストネスなモデルであるということがいえる．これは OPS も同様であるが，その他のモデルは非常に不安定であり優れた評価モデルとは言い難い．

⋆7 GP は初期乱数により進化の方向性が決まるため，異なる初期乱数ではそれぞれ進化の結果に微妙な差異が発生する．よって異なる初期乱数で GP を並列に処理し，その内の最もよい結果を選び出すことで処理時間を短縮できる．

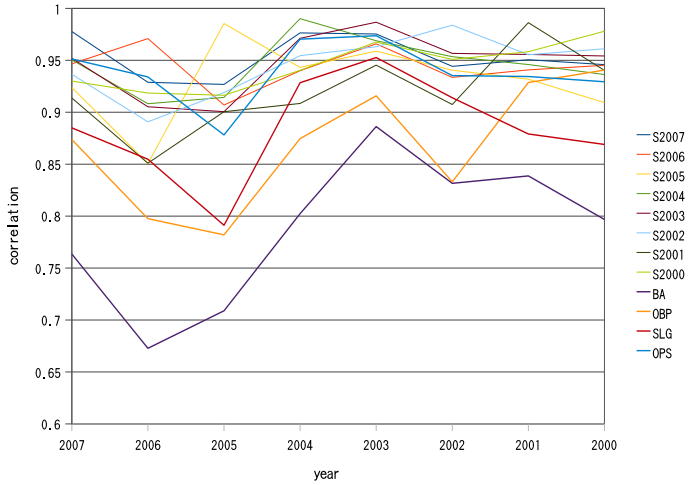

図 11 他の年との相関の比較

5. 考　　察

これまでの評価モデルを次の３つの視点から考える．

(1) 計算式の理解・納得しやすさ
(2) 打者の能力の総合的な評価（いかなるタイプの選手でも優劣無く評価できる）
(3) １試合平均得点との相関の高さ

これら３つはどれも優れた評価モデルに必要なファクターであり，これら全てを兼ね備えたモデルが理想と考えられる．ただ実際にはこれらはトレードオフの関係にある．例えば打率は１に関しては非常に高いが，その反面２，３はかなり低い．OPS に関しても３は高いが，１，２はそれほどでもない．[⋆8] 一方 GPC の場合は，機械学習の性質上多くの１を犠牲にしてその分２と３を重視している．これが良いか悪いかは評価モデルを使用する用途に依存する．つまり評価モデルの使用者が３のみが高ければ後は低くても構わないというのであれば GPC は良いモデルといえるし，そうでないならば OPS の方が優秀なモデルだといえる

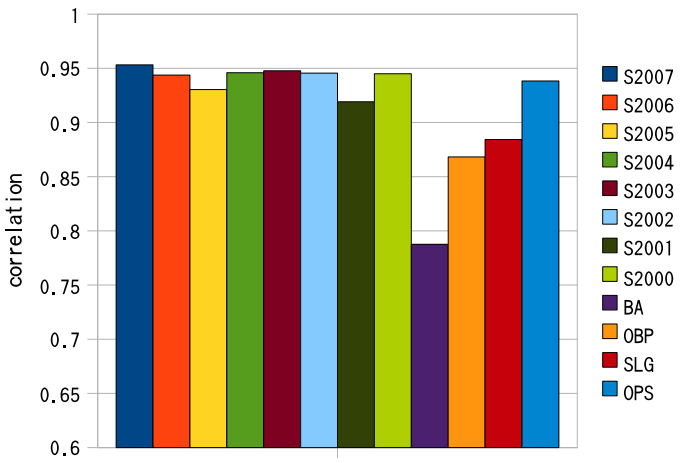

図 12 モデル別の相関係数平均値

のである．ただ今回の実験において GP における生き残るための基準は 3 のみなので，その点でいえば GPC は正しいモデルであるといえる．

6. ま と め

本研究では遺伝的プログラミングを用いてメジャーリーグにおける打者の最適な評価モデルを生成し，従来の評価モデルとの比較を行った．その結果，従来の評価モデルよりも得点との相関が高く，かつ学習データ以外のデータにも適応できるロバストネスな評価モデル GPC を獲得するに至った．ただ同時にいくつかの問題点・改善点も見つかっている．最大の問題としては，GPC は計算式の内容が非常に複雑だという点であり，一概に従来の評価モデルに比べて優れているとはいえないことは事実である．ただそれでもこれまでにない評価モデルを高いレベルで生成できたことは大きな成果だと考えている．今後の展望としては以下のとおりである．まず，ある年のデータにおける評価モデルの生成にその前年の選手データを使用する．それによってある程度未来を予測できるような評価モデルの生成が可能ではないかと推測している．また，打者と同じようにピッチャーの評価モデルを生成することも考えている．さらに打者と投手の評価モデルを併せて考え，チームの勝率との相関を調べることでペナントの順位を予想できるモデルの作成も可能ではないかと考えている．ペナントレースの順位予想は野球ファンにとっての至上命題であり，データマイニングでもってその命題を解決できるとなると非常に面白いのではないだろうか．

★8 2に関していえば，OPS は盗塁や犠打のパラメータが含まれていないためリーディングオフタイプ（1，2番打者）の評価には適していないとされる．

参 考 文 献

1) Koza.J:Genetic Programming II:Automatic Discovery of Reusable Programs,MIT Press,1994
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) J. アルバート／ J. ベネット：メジャーリーグの数理科学　上下，シュプリンガー・フェアラーク東京，2004