社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

プロジェクト類似性に基づく工数見積もりのための変数選択

瀧 進也† 柿元 健† 角田 雅照† 大杉 直樹† 門田 暁人† 松本 健一†

†奈良先端科学技術大学院大学情報科学研究科 〒630-0192 奈良県生駒市高山町 8916-5

E-mail: †{shinya-t, takesi-k, masate-t, naoki-o, akito-m, matumoto}@is.naist.jp

あらまし ソフトウェア開発プロジェクトの完遂に要する工数を正確に見積もるため，プロジェクト間の類似性に基づく見積もり方法の研究が進んでいる．しかし，少ない計算量で効率的に変数選択を行う方法はまだ提案されていない．本稿では，工数と相関が小さい変数を除去する Correlation Based Selection (CBS) を提案する．提案手法は，データ内の各変数について相関係数を計算するのみであるため，必要となる計算量が非常に少ない．ソフトウェア開発企業で収集されたデータを用いた評価実験の結果，提案手法は従来手法(変数減少法)より効率的に，より大きく精度を改善できた．実験で用いた精度評価尺度の値に着目すると，従来手法を用いた場合は MMRE (Mean Magnitude of Relative Error)が 0.047，相対誤差分散が 0.124 の改善にとどまり，PRED(25)は 0.039 悪化した．一方で，提案手法を用いた場合は MMRE が 0.165，相対誤差分散が 0.243，PRED(25)が 0.039 と，より大きく改善した．

キーワード 実証的ソフトウェア工学，コスト予測，事例ベース推論，プロジェクト管理

Variable Selection for Effort Estimation based on Similarity of Projects

Shinya TAKI† Takeshi KAKIMOTO† Masateru TSUNODA† Naoki OHSUGI†
Akito MONDEN† and Ken-ichi MATSUMOTO†

†Nara Institute of Science and Technology 8916-5 Takayama, Ikoma-shi, Nara, 630-0192 Japan

E-mail: †{shinya-t, takesi-k, masate-t, naoki-o, akito-m, matumoto}@is.naist.jp

Abstract Analogy-based estimation methods have been researched for deriving accurate effort estimate required to complete a project; however, there were few activities to develop efficient variable selection methods. In this paper, we propose *Correlation Based Selection* (*CBS*) excluding variables that do not correlate with the total effort. The proposed method needs few calculation amounts because it calculates only the correlation coefficients between the total effort and each variable in the data. We experimentally evaluated accuracy of the proposed method by using the data collected from a software development company. The result suggested the improvement of the accuracy with the proposed method was larger and efficient than the conventional method (*Step-down Variable Selection*). The result suggested the proposed method made larger improvement efficiently than the conventional method. In terms of the accuracy evaluation criteria, the proposed method improved 0.165 of *Mean Magnitude of Relative Error* (*MMRE*), 0.243 of *Variance of MRE*, and 0.039 of *PRED(25)*. On the other hand, the improvement of the conventional method was only 0.047 of MMRE, and Variance of MRE of 0.124; in addition PRED(25) was worse 0.039.

Keywords Empirical software engineering, cost prediction, case-based reasoning, project management

1. まえがき

ソフトウェア開発プロジェクトの完遂に要する工数を正確に見積もることを目的とし，工数見積もり手法が数多く研究されている[10]．例えば，COCOMO[2]や SLIM[11]などの手法では，ソフトウェアのソースコード行数を手法発案者が定義したモデルに入力することで，開発工数，期間，要員数，生産性を見積もれる．しかし，これらの手法ではモデル定義の際に考慮されていない組織やプロジェクトに固有の特徴（例えば，要因のスキル，顧客の業務分野，顧客との信頼関係など）を見積もりに反映できないため，精度が低くなってしまうことがある[3]．

これに対し，重回帰分析やニューラルネット[13]などを用い，実績データから組織独自の工数見積もりモデルを構築することもできる．実績データとは，プロジェクト毎に，当該プロジェクトの特徴を表す変数（工数，開発規模，検出バグ数など）の値を記録，蓄積したものである．この方法を用いると，組織固有の特徴を見積もりに反映できる．一方で，プロジェクト毎にモデルを構築するわけではないため，組織内でも珍しい特徴を持つプロジェクト（他と比べて規模が非常に大きい，多くの新人が参加しているなど）精度が低くなってしまう．

これらの問題を解決する手段として，プロジェクト

間の類似性に基づく見積もり方法が注目されている[12][14]．これら手法では，実績データから見積もり対象と似たプロジェクトを探し出し，過去プロジェクトでの実績から類推して見積もる．これは現場のプロジェクト管理者が，過去に手がけた類似案件をもとに見積もる方法を系統的に実行する方法である．これら手法を用いると，過去に似たプロジェクトが行われたのであれば，珍しい特徴を持つプロジェクトも高い精度で見積もれる．

一方で，プロジェクト類似性に基づく見積もりには，少ない計算量で効率的に変数選択を行う方法は提案されていない．変数選択法とは，見積もりの役に立たない変数を実績データから除去し，精度を改善する方法である．例えば，重回帰分析に対してはステップワイズ法などの有効な変数選択法が確立されている．一方，プロジェクト類似性に基づく見積もりに対しては，総当りで考えられる全ての組み合わせを試行するなどの単純な方法が提案されているのみである[12]．

そこで，本稿ではプロジェクト類似性に基づく工数見積もりに対する効率的な変数選択法として，Correlation Based Selection (CBS)を提案する．提案手法では，工数との相関係数が閾値より小さい変数を除去する．提案手法では，データに含まれる各変数について相関係数を計算するのみであるため，必要となる計算量が非常に少ない．本稿では，提案手法の有効性を確認するため，ソフトウェア開発企業で収集された実績データを用いた評価実験についても報告する．評価実験では，変数選択前と後の見積もり精度を比較し，精度改善の程度を観察した．精度改善の態度が従来の変数選択法と同程度か大きければ，効率性の面から提案手法の方が有効であると考えることができる．

以降，2 章では研究の背景として，主にプロジェクト類似性に基づく工数見積もりについて説明する．3 章では提案手法，および，評価実験で用いた従来手法について説明する．4 章で評価実験の方法として，用いたデータ，評価基準，実験手順を説明する．5 章で実験結果を述べ，6 章で結果について考察する．7 章で関連研究について紹介し，7 章でまとめと今後の課題を述べる．

2. 背景

ソフトウェア開発プロジェクトにおいてより正確に工数見積もりを行うことを目的として，プロジェクト類似性に基づく工数見積もりの研究が行われてきた[12][14]．この見積もり方法は，工数以外の特徴が互いに似たプロジェクトは，工数も互いに似た値を取るであろうという仮定に基づいている．まず，データに含まれる値の値域を [0.0, 1.0] に正規化する．次に，見積もり対象のプロジェクトと，過去に行われた各プロジェクトとの間の類似度を既に収集された変数の値から計算する．次に，類似度が高いプロジェクトの工数から，見積もり対象プロジェクトの見積工数を算出する．

この手法で，精度に大きく影響を与える類似度計算の方法として，ベクトル計算を用いる手法[14]とユークリッド距離を用いる手法[12]が提案されている．これら手法では欠損値(データに含まれる未記録の値)の取り扱い方が異なる．前者は欠損値を含んだまま類似度を計算できるのに対し，後者は平均値挿入法[9]などの前処理で欠損値を含まないデータを作る必要がある．本稿の評価実験では，類似度計算にベクトル計算を用いる手法を採用した．

プロジェクト類似性に基づく手法では，変数選択法が確立されていない．変数選択を行うことで，正確な見積もりを行う上で妨げとなっている変数を排除し，見積もり精度の向上が期待できる変数のみを利用して工数を見積もることができ，見積もりのための見積もりモデルを効果的に最適化することができる．

Shepperd ら[12]は，プロジェクト類似性に基づく見積もり手法に対して，総当りに基づく変数選択法を適用し，複数のデータセットを用いた実験において，ステップワイズ変数選択を適用した重回帰分析よりも高い見積もり精度を得ている．総当りに基づく変数選択は，変数のあらゆる組み合わせの中から最も見積もり精度が高くなるものを選ぶというアルゴリズムで行われる．総当りに基づく変数選択法は，理論上最も高い見積もり精度が期待できる手法であるが，計算に膨大な時間がかかるという問題がある．総当りに基づく変数選択では，変数のあらゆる組み合わせに対して見積もり精度を算出する．このため，変数の数を n とすると計算量はオーダ $O(2^n)$となる．このため，変数の数 n が大きくなると変数選択に膨大な計算が必要となる．

本稿ではプロジェクト類似性に基づく工数見積もりに対し，少ない計算量で，総当りに基づく変数選択法と同程度の精度改善を達成できる変数選択法を提案する．

3. 変数選択法

変数減少法

変数減少法は，変数選択法として広く用いられている．総当りによる結果と同程度の精度改善を，より少ない計算量で得られることが知られている．ある変数を取り除いたときに見積もり精度が向上した場合，それは不要な変数である，という考え方に基づいており，有用でないと考えられる変数を段階的に一つずつ削除

表 1. 実験に用いたデータセット

変数	平均値	中央値	最大値	最小値	総工数との間の相関係数
期間	11.30	10.00	36.00	1.00	0.65
要員の経験年数	2.30	2.00	4.00	0.00	0.26
プロジェクト管理者の経験年数	2.65	3.00	7.00	0.00	0.16
トランザクション数	177.47	134.00	886.00	9.00	0.58
調整済ファンクションポイント	282.39	247.00	1116.00	62.00	0.73
開発環境 (非数値)	–	–	–	–	–
開発完了年 (非数値)	–	–	–	–	–
エンティティ数	120.55	96.00	387.00	7.00	0.50
総工数	4833.91	3542.00	23940.00	546.00	–

していくことで変数選択を行う．段階的に変数を削除していくため，得られる変数選択結果は，複数の説明変数間に存在する複雑な関係も考慮した結果になっていると考えられる．

変数減少法に基づく変数選択は，以下の手順で行われる．

1. 説明変数を一つ削除して目的変数の見積もりを行い，見積もり精度を算出する．
2. 1 を全ての説明変数に関して行い，見積もり精度がもっとも高くなった変数を削除する．
3. 2 を見積もり精度が向上しなくなるまで繰り返す．

簡易変数減少法

簡易変数減少法は，変数減少法と同じ考え方に基づいた手法であるが，より簡略化し，さらに計算時間を小さくした手法である．変数減少法では段階的に変数を削除していくが，簡易変数減少法ではその変数が有用かどうかの判断を一回だけ行い，不要と判断された変数を一度に削除する．見積もり精度を算出する回数が少なくなるため，その分計算時間も小さくなる．一方，説明変数間に存在する関係について十分な考慮がされていないため，見積もり精度は変数減少法よりも低下する．

簡易変数減少法に基づく変数選択は，以下の手順で行われる．

1. 説明変数を一つ削除して目的変数の見積もりを行い，見積もり精度を算出する．
2. 1 を全ての説明変数に関して行い，削除前と比べて見積もり精度が高くなった変数を全て削除する．

CBS

CBS は本稿において新たに提案する変数選択法であり，適用手法の中で最も計算時間が小さい手法である．変数の値が似ているプロジェクト同士は工数も似た値を取る，というプロジェクト類似性に基づく手法の考え方を変数選択にも適用した手法であり，目的変数との間の相関係数が低い変数を除去する．CBS では，変数選択の際に見積もり精度を算出する必要がないため計算時間を小さく抑えることができる．CBS 変数選択は，以下の手順で行われる．

1. 全ての説明変数に関して，目的変数との間の相関係数を算出する．
2. 予め閾値を設定しておき，1 で求めた相関係数が閾値より小さかった変数を全て削除する．

相関係数は，データ件数を n，変数 X において i 番目のプロジェクトが取っている値を x_i，変数 X の平均値を $\bar{x}$，変数 Y において i 番目のプロジェクトが取っている値を y_i，変数 Y の平均値を $\bar{y}$ として，(1)式のように定義される．値が高い程，該当する変数間の関連が強いことを表す．

$$\text{相関係数} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i-1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

4. 評価実験

4.1. 目的

プロジェクト類似性に基づく工数見積もり手法に適用した 3 つの変数選択法と変数選択を行わなかった場合の見積もり精度を確認し，各変数選択法の性能を比較することを目的として評価実験を行った．以降，実験に用いたデータ，用いた精度評価尺度，精度評価のための実験手順について説明する．

4.2. 用いたデータ

実験に用いたデータセットは，Desharnais によって収集されたカナダのソフトウェア開発企業における 80 年代のデータである[5][7]．データセットには 77 件のプロジェクトについて，10 種類の変数が記録されている．データに欠損は含まれない．データの詳細を表 1

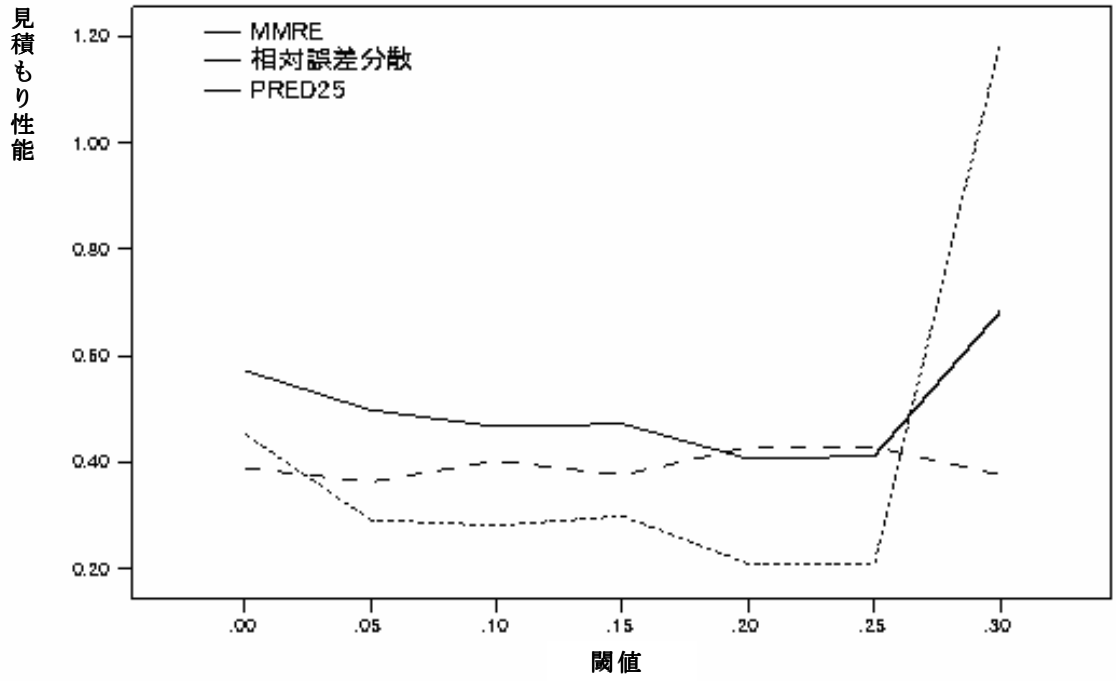

図 1. CBS の閾値を変化させたときの見積もり性能の変化

表 2. 各手法の見積もり性能

変数選択法	MMRE	相対誤差分散	PRED(25)
変数選択無し	0.572	0.453	0.390
変数減少法	0.525	0.329	0.351
簡易変数減少法	0.564	0.430	0.390
CBS(0.1)	0.467	0.282	0.403
CBS(0.2)	0.407	0.210	0.429

に示す．表中の開発期間からエンティティまでを説明変数とし，総工数を目的変数として見積もった．

評価実験では，見積もりを行う工程として，開発対象ソフトウェアのファンクションポイントの計測が終了した時点，すなわち，概要設計もしくは基本設計の完了時を想定している．また，開発期間（納期）は，開発初期に予め決定されていることを想定している．

4.3. 精度評価尺度

評価実験で見積もり性能の比較に Mean Magnitude of Relative Error (MMRE)，相対誤差分散，PRED(25)の 3 種類の精度評価尺度を用いた．以降で各評価尺度について説明する．

MMRE：*i* 番目のプロジェクトの相対誤差 MRE_i[4]を，実績工数を e_i，見積もり工数を $\hat{e}_i$ として(2)式のように定義すると，*n* 件の見積もりを行った場合の MMRE は(3)式で定義される．

$$MRE_i = \left| \frac{e_i - \hat{e}_i}{e_i} \right| \tag{2}$$

$$MMRE = \frac{1}{n} \sum_{i=1}^{i=n} MRE_i \tag{3}$$

MMRE は実績値と見積もり値の誤差から計算される．MMRE の値が小さいほど誤差が小さい，即ち，見積もり精度が高いことを示す．

相対誤差分散：相対誤差分散は次式で定義される．

$$\text{相対誤差分散} = \frac{1}{n-1} \sum_{i=1}^{i=n} MRE_i{}^2 \tag{4}$$

相対誤差分散は誤差のばらつきから計算される．相対誤差分散が小さいほど誤差のばらつきが小さく，より安定して見積もりができたことを示す．

PRED(25)：MRE_i が 0.25 以下となった件数を m とすると，PRED(25)は(5)式のように定義される．

$$PRED(25) = 100 \times \frac{m}{n} \tag{5}$$

PRED(25)は全プロジェクト中，誤差 25%以下で見積もれたプロジェクトの占める割合を表す．PRED(25)が大きいほど正しく見積もった割合が大きいことを示す．

4.4. 実験手順

本稿では，leave-one-out 法によって各プロジェクトの工数を見積もった．見積もった工数を実績値と比較することで精度評価尺度を算出した．実験で用いた leave-one-out 法の手順を以下に示す．

1. データ中のプロジェクトを 1 つ選ぶ．
2. 選んだプロジェクトにおける工数の実績値を隠し，見積もり対象のデータ(テストデータ)とする．
3. その他のプロジェクトを見積もりの根拠として使用するデータ(フィットデータ)とする．
4. フィットデータを用いて変数選択を行う．
5. 変数選択したフィットデータを用い，テストデータの工数を見積もる．
6. 見積もった工数と実績値を比較する．
7. 上記，1 から 6 をデータに含まれる全プロジェクトに対して行い，精度評価指標を算出する．

5. 結果

5.1. CBS の閾値による見積もり性能の変化

CBS を用いる場合，変数選択の基準となる閾値を決定しなければならない．CBS の閾値を 0 から 0.3 まで 0.05 刻みで変化させた場合の，各精度評価指標の変化を図 1に示す．グラフは，縦軸に各評価指標の値，横軸に変数選択の際の閾値を取っており，MMRE，相対誤差分散は値が小さいほど精度が高く，PRED(25)は値が大きいほど精度が高いことを表す．閾値を 0 から上げていくと，閾値 0.15 を除いて，0.25 までは MMRE，相対誤差分散が共に向上しており，見積もり精度，見積もり結果の信頼性ともに向上しているのがわかる．PRED(25)については，大きな変化は見られなかった．

5.2. 手法間の見積もり性能の比較

各変数選択法を適用した場合と，変数選択を行わなかった場合の見積もり性能の各評価指標の値を表 2に示す．左から順に，適用した変数選択法，MMRE，相対誤差分散，PRED(25)を表しており，CBS の後の括弧内の数字は閾値を表し，ここでは，閾値 0.1 と 0.2 の結果について示す．全ての変数選択法において，変数選択を行わない場合と比較して，見積もり精度の向上が確認できた．また，変数選択法の中では，閾値を 0.2 に設定したときの CBS（CBS(0.2)）が最も精度が高く，以下 CBS(0.1)，変数減少法，簡易変数減少法の順に精度が高かった．相対誤差分散についても同様に，変数選択によって精度が向上することが確認できた．一方，PRED(25)に関しては，大きな値の変化は見られず，変数減少法では，変数選択を行わない場合よりも低下した．

6. 考察

実験の結果，3 つの変数選択法によって見積もり精度が改善されることが確認できた．このことから，全ての変数が見積もりに有用なわけではなく，有用でない変数を除去することが精度向上に繋がっていることがわかる．また，今回の結果では新たに提案した変数選択法である CBS が最も高い精度を示した．

従って，相関係数に基づいて変数選択を行うことに，ある程度の妥当性があると考えられる．閾値を変化させた場合，閾値が 0.25 に達するまでほぼ常に見積もり精度が改善されていき，閾値が 0.3 になったときに急激に精度が悪化している．このことから，目的変数との間の相関係数が 0～0.25 程度であるような低い値を取っている変数は，見積もりの妨げとなる傾向が強いと考えられる．また，閾値が 0.15 になったときに精度が若干低下していることから，相関係数が低い変数の中にも，見積もりに有用な変数が存在していると考えられる．

見積もり精度は，データセットの持つ特性やデータ件数に大きく影響を受ける．従って，一つのデータセットでしか評価実験を行っていない今回の結果だけでは，CBS がプロジェクト類似性に基づく工数見積もり手法に対して最適な変数選択法であるとは一概には言えない．そのため，他のデータセットによる実験も行う必要があると考えられる．

7. 関連研究

Kirsopp らは，類似性に基づく工数見積もりに対して，ランダム変数選択，山登り法，変数増加法の 3 つの変数選択法を適用し，変数選択前よりも高い精度で見積もることができることを確認している．特に，ランダム変数選択でも見積もり精度が向上したことは興味深く，データセットの中に，見積もりに有用でない変数が多数存在することを示唆していると考えられる．

また，Auer らは，類似性に基づく工数見積もりに対して総当りでの重み付けを行う手法を提案している[1]．見積もりに用いられる全ての変数が等しく見積もり性能に寄与しているわけではなく，強く影響している変数や弱く影響している変数が存在していると考えられるため，適切な重み付けを行うことは重要である．しかし，総当りに基づく重み付けは非常に大きな計算時間を要する．大きな計算時間を必要としないように手法を改良することで，さらに実用性を高めることができる．

8. まとめ

本稿では，プロジェクト類似性に基づく工数見積もりに対する効率的な変数選択法として，Correlation Based Selection (CBS)を提案した．ソフトウェア開発企業で収集された実績データを用いた評価実験の結果，用いたデータに対しては提案手法が従来手法(変数減少法)より有効であると結論付けることができる．従来手法を用いた場合，MMRE (Mean Magnitude of Relative Error)が 0.047，相対誤差分散が 0.124 の改善にとどまり，PRED(25)は 0.039 悪化したのに対し，提案手法では MMRE が 0.165，相対誤差分散が 0.243，PRED(25)が 0.039 と，いずれもより大きく改善できた．つまり，提案手法は従来手法より効率的に，より大きく精度を改善できた．

今後は，実験結果の信頼性を高めるため，さらに多くの実績データを用いて本稿と同様の実験を行う予定である．様々なデータセットを用いて実験を重ねることで，各変数選択法が有効に機能する条件を明らかにすることができると考えている．また，提案手法による精度改善の効果をさらに向上するため，閾値の求め方，非数値変数の扱いについても検討を進める予定である．

謝　辞

本研究の一部は，情報処理推進機構ソフトウェア・エンジニアリング・センターとの共同研究に基づいて行われた．本研究の一部は，文部科学省「e-Society 基盤ソフトウェアの総合開発」の委託に基づいて行われた．

参　考　文　献

[1] M. Auer, A. Trendowicz, B. Graser, E. Haunschmid, and S. Biffl, “Optimal Project Feature Weights in Analogy-Based Cost Estimation: Improvement and Limitations,” IEEE Trans. on Software Eng., Vol.32, No.2, pp.83-92, 2006.

[2] B. Boehm, Software Engineering Economics, Prentice Hall, 1981.

[3] S. Chulani, B. Boehm, and B. Steece, “Bayesian Analysis of Empirical Software Engineering Cost Models,” IEEE Trans. on Software Eng., Vol25, No.4, pp.573-583, 1999.

[4] S. D. Conte, H. E. Dunsmore, and V. Y. Shen, Software Engineering Metrics and Models, The Benjamin/Cummings Publishing Company, Inc., 1986.

[5] J. M. Desharnais, “Analyse statistique de la productivitie des projets informatique a partie de la technique des point des function,” Masters Thesis, University of Montreal, 1989.

[6] 柿元健，角田雅照，大杉直樹，門田暁人，松本健一，“協調フィルタリングに基づく工数見積もりのロバスト性評価,” ソフトウェア工学の基礎 XI，日本ソフトウェア科学会 FOSE2004，pp.73-84，2004.

[7] C. Mair, G. Kadoda, M. Lefley, K. Phalp, C. Schofield, M. Shepperd, and S. Webster, “An Investigation of Machine Learning Based Prediction Systems,” J. Systems and Software, Vol.53, Issue 1, pp.23-29, 2000.

[8] I. Myrtveit, and E. Stensrud, “A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models,” IEEE Trans. on Software Eng., Vol.25, No.4, pp.510-525, 1999.

[9] I. Myrtveit, E. Stensrud, and U. H. Olsson, “Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods,” IEEE Trans. on Software Eng., Vol.27, No.11, pp.999-1013, 2001.

[10] Project Management Institute, A Guide To The Project Management Body Of Knowledge (PMBOK Guides), Project Management Institute, 2004.

[11] L. H. Putnam, “A General Empirical Solution to the Macro Sizing and Estimating Probrem,” IEEE Trans. on Software Eng., Vol.4, Mo.4, pp.345-361, 1971.

[12] M. Shepperd, and C. Schofield, “Estimating Software Project Effort Using Analogies,” IEEE Trans. on Software Eng., Vol.23, No.12, pp.736-743, 1997.

[13] K. Srinivasan, and D. Fisher, “Machine Learning Approaches to Estimating Software Development Effort,” IEEE Trans. on Software Eng., Vol.21, No.2, pp.126-137, 1995.

[14] 角田雅照，大杉直樹，門田暁人，松本健一，”協調フィルタリングを用いたソフトウェア開発工数予測方法 ,” 情報処理学会論文誌，Vol.46，No.5，pp.1155-1164，2005.