プロジェクト間の類似性に基づくソフトウェアメトリクスの欠損値の補完

田村 晃一　　柿元 健　　戸田 航史　　角田 雅照　　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

E-mail: {koichi-t,takesi-k,koji-to,masate-t,akito-m,matumoto}@is.naist.jp

概要

従来，過去のソフトウェア開発プロジェクトで計測された多数のソフトウェアメトリクス値を用いて，開発中もしくは将来のプロジェクトの信頼性や工数を予測し，計画立案に役立てることが行われている．ただし，予測モデルの構築時には，メトリクスに未記録の値（欠損値）が存在しないことが前提となっているため，モデル構築に先立って，欠損値を含むプロジェクトを除外したり，欠損値を当該メトリクスの平均値で補完することが行われる．しかし，データ全体に対する欠損値の割合が30%を越えることもしばしばあり，そのような場合には，たとえ欠損値を補完したとしても，性能のよいモデルの構築は期待できない．本稿では，欠損値が多い場合にもそれなりの予測性能をもったモデルを構築するために，欠損値を単に平均値で埋めるのではなく，プロジェクト間の類似性に基づいて欠損値を推定し，補完を行う．また，その効果を実験的に評価する．

1. はじめに

ソフトウェア開発プロジェクトにおける工数予測は，プロジェクト完遂に必要な資源，及びスケージュール管理を行う上で重要である．必要な工数を過不足なく予測することで，納期遅れ，コスト超過といったプロジェクトの失敗を防ぐことができる．そのため，工数予測に関する数多くの研究が行われている [1][7][8][10].

ソフトウェア信頼性の観点からは，ソフトウェア開発工程の終盤の工程であるテスト工数を予測し，テストに関わる人員と時間を決定することが重要となる．特に，テスト工数を過小に予測した場合には，テストが十分に行われず，納品後に故障が多発する危険性が増すため，大きな問題となる．

データに基づく工数予測手法では，過去のプロジェクトで収集されたデータを用いて予測を行うが，過去のプロジェクトのデータには，欠損値が数多く含まれている．一方で，重回帰分析などの工数予測手法は，モデルを構築するためのデータセットに欠損値が含まれていないことが前提となっている．そこで，モデル構築に先立って，欠損値を含むプロジェクトを除外したり，欠損値を当該メトリクスの平均値で補完することが行われる．しかし，データ全体に対する欠損値の割合が 30 % を越えることもしばしばあり，そのような場合には，たとえ欠損値を補完したとしても，性能のよいモデルの構築は期待できない [5].

本稿では，データセットに含まれるデータ欠損を補完する方法として，プロジェクト間の類似性に基づいて欠損値を推定し，補完を行う手法を提案する．プロジェクト間の類似性に基づいて欠損値を推定する方法は，協調フィルタリングを応用したアルゴリズム [10] を用いた．協調フィルタリングの特徴として，欠損値が多いデータを入力とした場合でも予測が行える特徴があり，情報検索の分野においてさかんに研究が行われてきた．協調フィルタリングを過去に行われたソフトウェア開発プロジェクトの欠損値補完に利用することで，より適切な欠損値補完が行えるようになり，高い精度で工数予測が可能になると期待される．

本稿では，提案手法の有効性を実データを用いて実験的に評価し明らかにする．実験では，実データの欠損値を提案手法によって補完し，補完したデータセットを用いてテスト工数を予測した．また，比較対象として，従来の欠損値処理である，ペアワイズ除去法，平均値挿入

法を用いた場合でも予測した．

以降，2 章では，本稿の実験で用いた工数予測手法であるステップワイズ重回帰分析について述べ，3 章では，提案手法であるプロジェクト間の類似性に基づく欠損値補完，および，従来の欠損値処理について述べる．4 章では，提案手法の有効性を示すための評価実験について説明し，5 章で評価実験の結果について述べる．6 章で関連研究について述べ，最後に 7 章で本稿の結論について述べる．

2　ステップワイズ重回帰分析による工数予測

重回帰分析は多変量解析の一手法であり，ソフトウェア開発に要する工数を予測するために広く用いられている．本稿では，工数予測手法として重回帰分析の一手法であるステップワイズ重回帰分析を用いた．

重回帰分析では，予測対象の変数 (目的変数) と，目的変数に影響を与える複数の変数 (説明変数) との関係を表した一次式 (回帰式) を作成する．回帰式中の各係数と定数は，予測値の絶対誤差（残差）の 2 乗和が最小になるように決定される．作成された回帰式に，現行プロジェクトで計測した説明変数を与えることで，目的変数を予測することが可能となる．

重回帰分析では，予測精度を向上させるために，多数の説明変数候補の中から，予測精度の向上に寄与すると予測される変数を選択して回帰式を作成する方法がとられる．ステップワイズ重回帰分析は，ステップワイズ変数選択法により採用する変数を決定し，重回帰分析を行う手法である．ステップワイズ変数選択は次の手順で行われる．

手順 1． 変数を全く含まないモデルを初期モデルとして作成する．

手順 2． 作成されたモデルに対して，各説明変数の係数が 0 でないかの検定を行い，指定した有意水準 (本稿の評価実験では，偏 F 値の有意水準を $p_{in} = 0.05$, $p_{out} = 0.1$ とした) で棄却されない場合に変数を採択する．ただし，多重共線性を回避するために，採択する変数の分散拡大要因 (VIF) が一定値以上の場合，またはその変数を採択することによって，他の変数の VIF が一定値以上となる場合，その変数は採択しない．

手順 3． 検定により適切な変数が選択されたと判断されるまで手順 2 を繰り返す．

3　メトリクスの欠損値の補完

3.1　提案手法

本稿では，プロジェクト間の類似性に基づく予測手法を，データセットに含まれる欠損値の補完に適用することを提案する．プロジェクト間の類似性に基づく予測手法は，メトリクス値が類似したプロジェクトの工数から，予測対象のプロジェクトの工数を予測する．プロジェクト間の類似性に基づく予測手法は，欠損値が多数含まれるデータセットを用いても，高い精度で工数を予測できることが報告されている．欠損値の補完においても，メトリクス値が類似したプロジェクトの値を用いることで，より適切な値の補完が可能であると考えられる．

プロジェクト間の類似性に基づく予測手法は 3 つの手順（正規化，類似度計算，補完値計算）から構成される [10]．各手順の詳細とアルゴリズムについて以下で述べる．

手順 1．メトリクス値の正規化 各メトリクスは値域に大きなばらつきがあるため，値域をそろえるための正規化を行い，値域を [0,1] にする．プロジェクト p_i のメトリクス m_j の値 v_{ij} を正規化した値 v'_{ij} は式 (1) で定義される．

$$v'_{ij} = \frac{v_{ij} - \min(p_j)}{\max(P_j) - \min(P_j)} \tag{1}$$

ここで，P_j はメトリクス m_j が計測されているプロジェクトの集合，$\max(P_j)$ と $\min(P_j)$ はそれぞれ $\{v_{x,j} | p_x \in P_j\}$ の最大値，最小値を表す．

手順 2．プロジェクト間の類似度計算 メトリクス値を補完するプロジェクトと類似した他のプロジェクトを見つけるため，プロジェクト間の類似度を算出する．メトリクス値を補完するプロジェクト p_a と他の各プロジェクト p_i との類似度 $sim(p_a, p_i)$ は式 (2) で定義される．

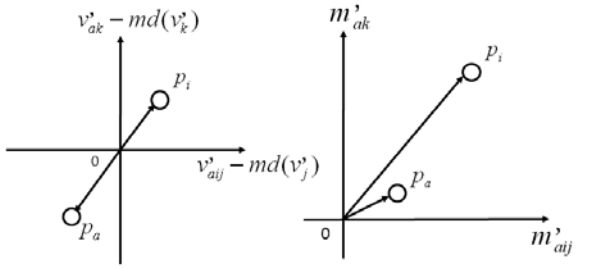

図 1. プロジェクト間の類似度計算例

$$sim_{p_a,p_i}=\frac{\sum_{j\in M_a\cap M_i}\left(v'_{aj}-md(v'_j)\right)\left(v'_{ij}-md(v'_j)\right)}{\sqrt{\sum_{j\in M_a\cap M_i}\left(v'_{aj}-md(v'_j)\right)^2}\sqrt{\sum_{j\in M_a\cap M_i}\left(v'_{ij}-md(v'_j)\right)^2}} \tag{2}$$

ここで，M_a と M_i はそれぞれプロジェクト p_a と p_i に関して記録されている（欠損していない）メトリクスの集合を表し，$md(v'_j)$ は j 番目のメトリクスの中央値を表す．

v_{ij} から $md(v'_j)$ を減算することで，中央値よりも大きなメトリクス値は正の値をとり，小さい値は負の値をとるようになる．したがって，類似度 $sim(p_a,p_i)$ の値域が [-1,1] をとるようになり，大きくメトリクス値が離れたプロジェクト間の類似度が小さくなる（図 1）．

手順 3．類似度に基づく補完値の算出 補完対象となる欠損値について，その補完値の算出に類似したプロジェクトの対応するメトリクスの実測値を用いる．手順 2. の類似度計算では，(ベクトルのなす角を用い，ベクトルの大きさを用いないため）プロジェクトの規模が考慮されず，規模が異なるプロジェクト同士の類似度が高くなる場合がある．そこで，提案手法では，補完値の算出において，類似度 $sim(p_a,p_i)$ を重みとして，プロジェクト p_a と類似したプロジェクトのメトリクス値 v_{ib} に，プロジェクトの規模を補正する $amp(p_a,p_i)$ を乗じた値で加重平均を行う．プロジェクト p_a のメトリクス m_b の補完値 $\hat{v}_{ab}$ は式 (3) で定義される．

$$\hat{v}_{ab}=\frac{\sum_{i\in k\text{-}nearestProjects}(v_{ib}\times amp(p_a,p_i)\times sim(p_a,p_i))}{\sum_{i\in k\text{-}nearestProjects}sim(p_a,p_i)} \tag{3}$$

ここで，k-$nearestProjects$ は，メトリクス m_b が欠損しておらず，かつ，プロジェクト p_a と類似度の高い上位 k 個のプロジェクトの集合を表す．k の値は実験的に別途求める必要がある．

また，$amp(p_a,p_i)$ は式 (4) で定義される．

$$amp(p_a,p_i)=\begin{cases} r_n \quad \cdots \quad h=\text{奇数} \\ (r_1\le r_2\le\cdots\le r_n\le\cdots\le r_{2n-1}) \\ \frac{r_n+r_{n+1}}{2} \quad \cdots \quad h=\text{偶数} \\ (r_1\le r_2\le\cdots\le r_n\le\cdots\le r_{2n}) \end{cases} \tag{4}$$

ここで，$h=|M_a\cap M_i|$，$r_i=\frac{v'_{aj}}{v'_{ij}}$ である．

amp は，プロジェクト p_a の規模が p_i の規模のおよそ何倍になっているかを，正規化されたメトリクスの比 r_j の中央値により求めている．これは，多くのソフトウェアメトリクスが，プロジェクトの規模と相関が高いことを利用している．この amp により，多様な規模のプロジェクトを含むデータセットを用いた場合にも補完が可能となる．

3.2 欠損値処理

欠損値を含むデータセットに対してステップワイズ重回帰分析を適用する手法として，欠損値処理が従来使われている．欠損値処理とは，多変量解析を可能とするために，与えられたデータセットから欠損値を含むプロジェクトを除外したり，欠損値を何らかの値で補完する，といった前処理を行う方法である．重回帰分析に対しては，リストワイズ除去法，ペアワイズ除去法，平均値挿入法の３種類の手法が広く用いられる [5][9].

リストワイズ除去法 欠損値を１つでも含むプロジェクトを全て除去する．

ペアワイズ除去法 重回帰分析に特化した手法で，重回帰分析の過程においてメトリクス間の相関を求め

る際に，相関を求めるメトリクスのいずれかが欠損しているプロジェクトを除外して相関を求める手法である [9].

平均値挿入法　欠損値に対して，当該メトリクスの平均値を挿入することで，欠損値を補完する．

4. 評価実験

4.1　利用データ

実験で利用したデータは，ISBSG(International Software Benchmarking Standards Group) が収集した，20ヶ国のソフトウェア開発企業の実績データ [3] である．データに含まれるメトリクスのうち Effort Test を目的変数とし，13 個のメトリクスを説明変数として用いた．14 種類のメトリクスの詳細について表 1 に示す．このデータから，目的変数である Effort Test が欠損しているプロジェクトと欠損値が大半を占めるプロジェクトを除去し，512 件のプロジェクトを実験に用いた．

表 1 のメトリクスのうち，Resource Level は工数の測定方法に関わるメトリクスである．工数の単位は通常人月で与えられるが，Resource Level はこのうち人的要素に関わる変数であり，プロジェクトとしての作業内容の定義である．このメトリクスはどのような業務を行っていた人員をプロジェクトの開発に関わっているとして工数に算入したかを 1〜4 で表しており，Level 1 であればコーディングやドキュメント作成など，プロジェクトの中心として作業を行ったメンバーのみの作業時間を工数として算入し，Level 4 であればプロジェクトで開発されたソフトウェアを実際に納入する際に，納入先で利用者に教育を行うような，プロジェクトとの関わりが比較的薄いメンバーの作業時間をも工数として算入している．

4.2　評価基準

予測精度の評価基準として一般的に用いられている，絶対誤差，相対誤差それぞれの平均値，中央値，および Pred(25) の 5 種類の評価基準を用いて評価した．誤差の各評価基準は値が小さいほど予測精度が高いことを表し，Pred(25) は値が大きいほど予測精度が高いことを表す．

それぞれの評価基準は次の式 (5)〜(10) で計算される．ここで，M 件のプロジェクトがあるとする．また，実測値と予測値をそれぞれ X_i，$\hat{X}_i (i = 1〜M)$ とし，$A_i = \left|\hat{X}_i - X_i\right|$, $R_i = \frac{\left|\hat{X}_i - X_i\right|}{X_i}$ とおく．

絶対誤差平均値 (MAE)

$$MAE = \frac{\sum_{i=1}^{M} A_i}{M} \tag{5}$$

絶対誤差中央値 (MdMAE)

$$MdMAE = \begin{cases} A_n & M = \text{奇数} \\ (A_1 \leq A_2 \leq \cdots \leq A_n \leq \cdots \leq A_{2n-1}) & \\ \dfrac{A_n + A_{n+1}}{2} & M = \text{偶数} \\ (A_1 \leq \cdots \leq A_n \leq A_{n+1} \leq \cdots \leq A_{2n}) & \end{cases} \tag{6}$$

相対誤差平均値 (MMRE)

$$MMRE = \frac{\sum_{i=1}^{M} R_i}{M} \tag{7}$$

相対誤差中央値 (MdMRE)

$$MdMRE = \begin{cases} R_n & M = \text{奇数} \\ (R_1 \leq R_2 \leq \cdots \leq R_n \leq \cdots \leq R_{2n-1}) & \\ \dfrac{R_n + R_{n+1}}{2} & M = \text{偶数} \\ (R_1 \leq \cdots \leq R_n \leq R_{n+1} \leq \cdots \leq R_{2n}) & \end{cases} \tag{8}$$

表 1. 実験データに含まれるメトリクス

変数の種類	名称	詳細	欠損率 (%)
説明変数	Adjusted Function Points	調整済み FP 数	0
	Effort Plan	計画工数	55.4
	Effort Specify	要件定義工数	14.6
	Effort Build	コーディング工数	0.4
	Resource Level	工数に含める作業内容 (4 段階)	0
	Input count	入力機能の FP 数	37.8
	Output count	出力機能の FP 数	37.8
	Enquiry count	照会機能の FP 数	39.0
	File count	ファイル更新機能の FP 数	37.8
	Interface count	インタフェースの FP 数	39.0
	Added count	新規または追加された機能の FP 数	36.6
	Changed count	変更された機能の FP 数	36.6
	Deleted count	削除された機能の FP 数	36.6
目的変数	Effort Test	テスト工数	0

Pred(25)

$$Pred(25) = \sum_{i=1}^{M} isAccurate(R_i) \qquad (9)$$

$$isAccurate(R) = \begin{cases} 1 & R \leq 0.25 \\ 0 & R > 0.25 \end{cases}$$

4.3 実験手順

評価実験は次の手順で行った.

1. 4.1 で述べたデータセットを，欠損値を含むプロジェクトのみのデータセット (プロジェクト数 400 件) と欠損値を含まないプロジェクトのみのデータセット (プロジェクト数 112 件) の 2 つのデータセットに分割した.

   前者を予測モデルを作成するデータ（以降フィットデータと呼ぶ）とし，後者をフィットデータを用いて実際に予測を行うデータ（以降テストデータと呼ぶ）とした．欠損値を含まないデータセットをテストデータとしたのは，欠損値によって予測精度が影響されることを防ぐためである.

2. フィットデータに対し，提案手法，および，比較対象であるペアワイズ除去法，平均値挿入法によって欠損値処理を行った．本稿の評価実験では，欠損値処理を施すデータセットが欠損値を含むプロジェクトのみで構成されており，リストワイズ除去法では全てのプロジェクトが除去されるため，ペアワイズ除去法および平均値挿入法を用いた．提案手法において式 (3) の *k-nearestProjects* は 10 とした.

3. それぞれの手法で欠損値を補完あるいは削除したフィットデータに対して，ステップワイズ重回帰分析で予測を行い，各評価基準を算出した．予測対象のメトリクスを Effort Test とし，テストデータの Effort Test は未知数として予測を行った．分散拡大要因 (VIF) は 10 とした.

5. 評価実験の結果

提案手法及び従来法によって欠損値処理を行い，ステップワイズ重回帰分析で予測した時の各評価基準の値を表 2 に示す.

表 2 より，提案手法を用いて欠損値を補完した場合，評価基準の 4 種類の誤差の値が最も小さく，Pred(25) の値が最も大きい．このことから，提案手法を用いて欠損

表 2. 各手法のステップワイズ重回帰分析での予測精度

	MAE	MdMAE	MMRE	MdMRE	Pred(25)
提案手法	917.055	351.802	2.358	1.003	16%
ペアワイズ除去法	1425.046	439.840	3.767	1.308	7%
平均挿入法	1060.311	390.888	5.818	1.427	10%

値を補完した場合が最も高い精度で予測可能であると言える．また，ペアワイズ除去法と平均値挿入法を比較すると，絶対誤差では平均値挿入法が，相対誤差ではペアワイズ除去法の方が精度が高い．

各手法の絶対誤差，相対誤差の箱ひげ図を図 2 に示す．グラフの縦軸は予測誤差を示し，箱の下端は第 1 四分位，上端は第 3 四分位，箱中の+は中央値，線分（ひげ）の下端の横線は最小値を表す．グラフの上方を省略しているため最大値は示されていない．

図 2 の箱ひげ図より，提案手法を用いて欠損値を補完した場合，箱が一番下に来ており，全体的に精度が高いと言える．また，相対誤差の最小値を除く，最小値，第 1 四分位，第 3 四分位，最大値で，比較対象の手法よりも値が最も小さくなっている．ペアワイズ除去法は，提案手法と平均値挿入法と比べて箱の大きさが大きく，誤差のばらつきが大きいと言える．

これらの評価実験の結果から，プロジェクト間の類似性に基づく予測手法によって欠損値を補完する提案手法は，比較対象の手法であるペアワイズ除去法，平均値挿入法によって欠損値を補完，除去した場合よりも高い精度で予測でき，適切な値で欠損値を補完していると言える．

また，平均値挿入法をペアワイズ除去法と比較した場合，絶対誤差では精度が高く，相対誤差では精度が低いことから，平均値挿入法で欠損値を補完した場合には，過小予測が多く起こっていると考えられる．

6. 関連研究

提案手法のように，何らかのモデルを用いて欠損値を補完する方法は，model based，あるいは likelihood な手法と呼ばれる．Model based な手法として，これまでにも k-nn 法 [2] [4] や similar response pattern imputation (SRPI)[6], full information maximum likelihood

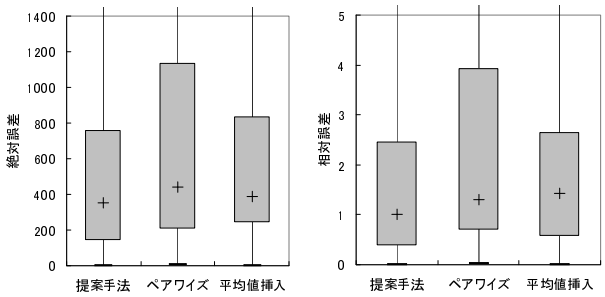

図 2. 各手法ごとの予測精度の箱ひげ図

(FIML) [6]，HotDeck 法 [9]. などを用いた手法などが提案されている．

従来の研究では，欠損値を正確に補完することに注目しており，欠損値を補完したデータセットを用いて工数予測などを行う点に関しては触れられていない．欠損値補完の正確さを評価するためには，欠損値の実測値が必要なため，従来研究の評価実験では，欠損値を含まないデータセットに対して，実験者側で欠損値処理などの手法によって欠損値を与える必要がある．したがって，評価実験で使用しているデータセットが実際の欠損値を正確に表しているとは言えない．

従来研究を踏まえ，本稿では，実際に欠損値を含むデータセットに対して手法を適用することで欠損値を含まないデータセットを作成している．また，補完の正確さによって評価するのではなく，工数を予測することによって評価を行っている．

7. おわりに

本稿では，データセットに含まれる欠損値を補完する方法として，プロジェクト間の類似性に基づいて欠損値を推定し，補完を行う手法を提案した．評価実験の結果，提案手法によって欠損値を補完することで，従来手法で

あるペアワイズ除去法，平均値挿入法よりも高い精度で予測でき，欠損値を適切な値で補完できることが確認できた．

今後は，欠損値をより適切な値で補完できるように手法の改善を行う予定である．また，k-nn 法などの提案手法以外の model based な欠損値補完法と比較する予定である．

謝辞

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] B.W. Boehm, Software engineering economics, Prentice Hall, New Jersey, 1981.

[2] M. Cartwright, M.J. Shepperd, and Q. Song, "Dealing with Missing Software Project Data," Proc. 9th IEEE International Software Metrics Symposium (Metrics'03), pp.154-165, Sydney, Australia, 2003.

[3] "ISBSG Estimating, Benchmarking and Research Suite Release 9," International Software Benchmarking Standards Group, 2004, http://www.isbsg.org/

[4] P. Jonsson and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likert data," Proc of the 10th IEEE International Software Metrics Symposium (Metrics'04), pp.108-118, Chicago, Illinois, 2004.

[5] J. Kromrey, and C. Hines, "Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments," Educational and Psychological Measurement, vo.54, no.3, pp.573-593, 1994.

[6] 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. Software Eng., vol.27, no.11, pp.999-1013, 2001.

[7] M. Shepperd, and C. Schofield,"Estimating software project effort using analogies," IEEE Trans. Software Eng., vol.23, no.12, pp.736-743, 1997.

[8] K. Srinivasan, and D. Fisher, "Machine learning approaches to estimating software development effort," IEEE Trans. Software Eng., vol.21, no.2, pp.126-137, 1995.

[9] K. Strike, K. El Eman, and N. Madhavji, "Software cost estimation with incomplete data," IEEE Trans. Software Eng., vol.27, no.10, pp.890-908, 2001.

[10] 角田雅照, 大杉直樹, 門田暁人, 松本健一, 佐藤慎一, "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," 情処学論, Vol.46, No.5, pp.1156-1164, 2005.