情報処理学会論文誌 Vol.53 No.2 662–671 (Feb. 2012)

欠損率の高いプロジェクトデータを利用したプロジェクトの成否予測

出張 純也[1,a)] 菊野 亨[1] 菊地 奈穂美[2] 平山 雅之[3]

受付日 2011年6月14日, 採録日 2011年11月7日

概要：ソフトウェア開発プロジェクトから収集したデータを利用して，品質やコストなどを予測する研究が多く行われている．本研究では，通常のプロジェクトから収集される欠損の多いデータを利用して，プロジェクトの成否の予測を試みる．欠損率が高いので，2段階の方法を提案する．最初に，未記入項目の多いメトリクスを削除し，次に予測に影響を与えると考えられるメトリクスだけに絞り込む．メトリクスの絞り込みには相関ルールマイニングを適用する．適用実験として，IPA/SEC のデータ白書として公開されているプロジェクトデータを利用して，プロジェクトの成否を設計工程の終了時に予測した．まず，設計工程終了時点ではまだ値が定まらないメトリクスを削除した．その時点でのデータの欠損率は 43.8%になった．提案法を適用した結果，メトリクスを 7 個にまで絞り込み，予測精度 82.8%が達成できた．

キーワード：ソフトウェア工学，プロジェクトマネジメント，メトリクス

On Prediction of Project Success Using Incomplete Project Data

Junya Debari[1,a)] Tohru Kikuno[1] Nahomi Kikuchi[2] Masayuki Hirayama[3]

Received: June 14, 2011, Accepted: November 7, 2011

Abstract: Many researches tried to predict quality and cost using project data set. Note that project data set is usually assured to be complete in the sense that all metrics data is filled out. But actually we are facing with public project data set which contain many incomplete data. In this paper we try to predict, after design phase, if a project will finish successfully or not based on such a public project data set. We propose two phases of refinements upon data set: (1) reduction of incomplete data and (2) extraction of meaningful metrics. The first reduction is just deletion of such metrics that contain many missing data. We then apply association rule mining for metrics extraction. For prediction of a project,we employ Bayesian Classifier as usual. We conducted an experimental evaluation on IPA/SEC data set which is collected from Japanese companies. The IPA/SEC data set consists of 237 projects and 69 metrics,and contains 43.8% of missing data. By applying the proposed method, 82.8% of accuracy was finally realized with only 7 metrics.

Keywords: software engineering, project management, metrics

1. はじめに

ソフトウェア開発プロジェクトから収集したプロジェクトデータを対象として，品質，コスト，成否などを予測する研究が多く行われてきている．これまでのプロジェクトデータは周到に計画・設計されたプロジェクトから収集されたデータ（ケース 1）か，あるいはよく整備された開発組織におけるプロジェクトから収集されたデータ（ケース 2）が利用されていた．

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565–0871, Japan
[2] 沖電気工業株式会社
Oki Electric Industry Co., Ltd, Warabi, Saitama 335–8510, Japan
[3] 日本大学大学院理工学研究科
Graduate School of Science and Technology, Nihon University, Funabashi, Chiba 274–8501, Japan
[a)] j-debari@ist.osaka-u.ac.jp

しかし現実には，こうした条件を満たさない多くのプロジェクトでも様々なメトリクスデータが収集されている．たとえば，International Software Benchmarking Standards Group（ISBSG）や情報処理推進機構ソフトウェア・エンジニアリング・センターがソフトウェア開発プロジェクトのデータを収集している [6], [17]．こうしたデータには欠損が多く含まれている．IPA/SEC は 2008 年の時点で国内の企業 20 社からソフトウェア開発プロジェクトのデータを収集している．本研究では，こうしたデータをケース 3 と呼び，その活用について 1 つの提案を行う．ケース 1，2 とケース 3 の違いはデータの欠損率に特徴的に現れる．ケース 1 での欠損率は 0%，ケース 2 では 10%未満であるのに対して，ケース 3 では 30%以上になる．

ケース 1 やケース 2 の研究では，予測精度を向上させるためにメトリクスの絞り込みを行うものが多く存在する．本研究においても，予測精度の向上のためにメトリクスの絞り込みを行う．ただし，ケース 3 では欠損率が高いためケース 1 やケース 2 の方法とは異なるアプローチをとる．提案法の主な特徴は次の 3 つである．

(1) まず未記入項目の多いデータを削除して欠損率をある程度改善させておいて，メトリクスの絞り込みを行う．
(2) 未記入項目の多いデータの削除は，未記入項目の多い順番に機械的に行う．
(3) メトリクスの絞り込みには，欠損があるデータに対しても利用できる相関ルールマイニング法を適用する．

本研究では IPA/SEC のデータ白書として公開されているプロジェクトデータを利用して，プロジェクトの成否を設計工程の終了時に予測することを試みる．そのため，提案法のフェーズ 1 として，設計工程が終了した段階でまだ入手が不可能なメトリクスを削除した．こうして求まったプロジェクトデータの欠損率は 43.8%になっていた．適用実験の結果，未記入項目データの削除の操作とメトリクスの絞り込みを行うことによって，予測精度は最も高い場合で 7 つのメトリクスによって 82.8%が達成できることを確認した．

2 章では，関連研究を紹介する．3 章では本研究で扱う予測問題について述べる．4 章では提案手法について述べる．5 章では適用実験とその分析結果について述べる．6 章では研究のまとめを述べる．

2. 従来の研究

プロジェクトデータを利用して種々の予測を行う研究は，おおむね，次の 2 種類に分類することができる．

(1) 可能な限り注意深く設計された開発プロジェクトから，事前に選定されたメトリクスのデータを収集して，それらを用いた科学的，あるいは統計的な分析を行う．データには欠損がほとんどない．
(2) 特に設計されたわけではない，通常の開発プロジェクトから事前に選定されたメトリクスのデータを収集して，それらに種々の統計的な分析を行う．

文献 [3], [7], [11] の研究は，欠損のないプロジェクトデータを対象として行われた研究である．つまり典型的なケース 1 の研究である．文献 [11] では，複数のデータセットに対して，類似度に基づいた工数見積りを適用した．その結果，すべてのデータセットにおいて高い精度で見積りが可能であることが分かった．文献 [3] では，COCOMO 法でコストを見積もる際に，見積りの役に立たないメトリクスを削除することで，コスト見積り精度が上昇することが明らかにされた．文献 [7] は，CoBRA 法に基づいて品質予測を行っているが，品質予測のためのパラメータ決定を行い，予測モデルを構築している．この手法では，予測に用いる特徴の決定を専門家の知見と統計手法に基づいて行っている．

一方，文献 [1], [13], [19] の研究は，少量の欠損を含むプロジェクトデータを対象として行われた研究で，ケース 2 に属している．文献 [13] では，プロジェクトマネージャからのアンケート結果に対してロジスティック回帰分析を適用し，失敗プロジェクトの予測を試みている．この研究では，欠損（アンケートに回答がないこと）を潜在的なリスク要因であると考えて欠損を補完している．文献 [1] では，ベイズ識別器をソフトウェア開発データに適用して，ソフトウェアプロジェクトの最終状態を予測している．ベイズ識別器は欠損のあるデータに対しても適用可能な手法であるため，欠損の削除や補完は行っていない．文献 [19] では開発の現場から得られたアンケートの回答に対して相関ルールマイニングを適用して，ソフトウェアプロジェクトが混乱するリスク要因の特定を行っている．

欠損を扱う方法についての研究としては文献 [2], [8], [12], [14] がある．文献 [12] では，無欠損のデータに人工的に欠損を発生させて，欠損データの削除法，欠損データの補完法について評価実験を行っている．その結果，欠損が少ない場合には欠損データを削除するのが最も効率が良いと結論している．文献 [2] では，工数見積りに使うデータについて，K 近傍法と中央値法を比較し，K 近傍法で補完する場合に精度が最も高くなることを明らかにした．文献 [14] では，複数の欠損補完法を比較した結果，多重補完法が最も優れていると報告している．

上述の研究以外に特徴のある研究として文献 [18], [20] がある．文献 [18] では協調フィルタリングを用いてソフトウェア開発工数を予測している．60%の欠損があるデータに対して協調フィルタリングを適用することで，欠損の除去や補完をしたうえで重回帰分析を行う場合よりも高い精度が達成できることを明らかにした．しかし，文献 [18] ではすべてのメトリクスを用いた分析を行っており，変数の選択を今後の課題としている．文献 [20] ではプロジェクトの類似性に基づいて工数を予測している．この研究では，

ISBSG のデータなどを利用しているが，ユークリッド距離に基づいて類似性を計算しているため，欠損のないデータを作成して利用している．欠損のあるデータを利用する場合については述べられていない．

3. 本研究での予測問題

3.1 プロジェクト成否予測の考え方

本研究ではプロジェクトの成否を設計工程の終了時に予測することを考える．文献 [16] では，見積りの際には開発するシステムだけでなく，プロジェクトの前提条件，プロジェクトにかかわる組織の特徴，その組織内のプロセス，開発チームや個々のメンバの特性など，様々な要因を考慮する必要があるとされている．そのため，本研究では，こうしたメトリクスのうち設計工程の終了時に利用できるものを利用することにした．

また，文献 [9] では，プロジェクトの見積りには，開発期間．工数，機能量，信頼性，生産性の 5 つの中核メトリクスが重要とされている．さらに，これらの中核メトリクスは相互に関係しているため，いくつかが分かれば残ったメトリクスは推定が可能とされている．そのため，これら中核メトリクスについても予測モデルを作成する際に利用することにした．

これらのメトリクスを利用してプロジェクト成否予測を行う場合，過去のプロジェクトデータを用いて予測モデルを作成し，そのモデルに予測対象プロジェクトのデータをあてはめて予測していく．この場合，以下の 2 点が大きな課題となる．

課題 1：欠損 予測モデルを作成する際に利用する過去のプロジェクトデータに欠損が存在する．

課題 2：予測の時期 予測を行う時期によっては，予測に用いるメトリクスに計画値を用いなければならない場合が存在する．

3.1.1 予測における計画値の利用（課題 2 への対応）

これらのメトリクスの中には，プロジェクトの終了後でないと確定しない実績値が存在する．そのため．設計工程終了後などのプロジェクトの前半でこれらのメトリクスの実績値を用いて成否予測を行うのは難しい．そこで，我々はプロジェクトの前半でプロジェクト成否を予測するために，これらのメトリクスについては計画値を用いることを考える．この考えを実現するために，以下のような手順で検討していく．

(1) 工数，期間，規模などのメトリクスの実績値とプロジェクト成否の実績値を用いて，その関連から予測モデルを作成する．

(2) 工数，期間，規模などのメトリクスの実績値と計画値がどの程度乖離しているかを評価する．

(3) 実際のプロジェクト成否予測を行う場合には，(1) で作成したモデルのうち，実績値となっている部分に，計画値の値を代入する．ここで，計画値とは予測の段階で見積りを行うものとする．

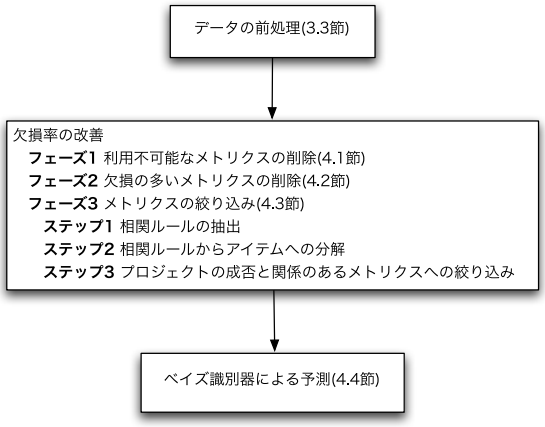

図 1 提案法の構成
Fig. 1 An overview of the proposed method.

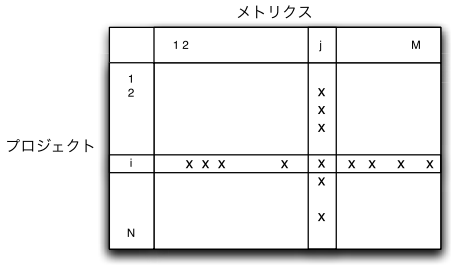

図 2 データ欠損の説明図
Fig. 2 Explanation of missing data.

上記のような手順でプロジェクト成否予測を行う場合，(2) の評価で得られた実績値と計画値の乖離が予測結果に影響を及ぼす可能性がある．このため，実際に提案法を用いてプロジェクト予測を行う場合には，入力とするメトリクスの乖離を予測変動要因として考慮したうえで，プロジェクトの進行にあわせて逐次詳細な情報が確定するたびに繰り返し予測し，予測精度の向上を図り，そのつど，プロジェクトの成否を考えるきっかけを与える運用を想定している．

3.1.2 データ欠損への解決の指針（課題 1 への対応）

データに欠損が多すぎる場合には分析の結果が不正確になる可能性があるため，欠損のあるデータを削除する．しかし，データに欠損が多いため，データの欠損をすべて削除することができない．そのため，本研究では，2 段階で，(1) 機械的に欠損の多いメトリクスを削除する，(2) プロジェクトの成否と関係のあるメトリクスだけに絞り込みを行う，という方法を採用する．この提案法の概要を**図 1** に示す．

図 2 は M 種類のメトリクス，N 件のプロジェクトからなるデータのイメージ図である．3.2 節の IPA/SEC データの場合，$M = 633$，$N = 1{,}397$ となる．x とマークしてある箇所はその値が欠損しているとする．j（$1 \leq j \leq M$）番目の

メトリクスに着目すると，6 個の欠損がある．$i\,(1 \leq i \leq N)$ 番目のプロジェクトには 9 個の欠損がある．このような欠損を削除するには，(1) 欠損の多いメトリクスを削除する，(2) 欠損の多いプロジェクトを削除する，の 2 種類の方法が考えられる．本研究では，(1) の方法を採用する．これは，分析に利用するプロジェクト数が少なくなると統計的な分析が正確に行えない可能性があるからである．

メトリクスの絞り込みの方法としては，ロジスティック回帰分析や相関ルールマイニングなどが考えられる．ロジスティック回帰分析は欠損のないデータへの適用を前提としているため，本研究では利用せず，欠損のあるデータに対しても適用可能な相関ルールマイニングを利用する．

プロジェクトの予測に関しても，欠損のあるデータに対して利用可能な方法を利用する必要がある．そのため，本研究ではベイズ識別器を利用する．

3.2 プロジェクト成否の定義

本研究では IPA/SEC のデータ白書として公開されているプロジェクトデータ [17] を利用して，プロジェクトの成否を設計工程の終了時に予測することを試みる．

このデータは国内の企業 20 社から収集されたものである．2007 年までに終了したエンタープライズ系のソフトウェア開発プロジェクトのうち，1397 プロジェクト[*1]から 633 種類のメトリクスを収集している．データの欠損率は 83.8%となっている．ここで，プロジェクトの成否の判断にはメトリクス「実績の評価（品質）」を利用する．このメトリクスの値は，稼働後 6 カ月の不具合数の実績値が計画値に対してどの程度大きかったかを示す．a，b，c，d，e の 5 種類の値をとり，a が「稼働後不具合数が計画値より 20%以上少ない」，b が「稼働後不具合数が計画値以下である」，c が「稼働後不具合数が計画値の 50%以内の超過である」，d が「稼働後不具合数が計画値の 100%以内の超過である」，e が「稼働後不具合数が計画値の 100%を超える超過である」という意味である．a，b を稼働後不具合数が計画値よりも少なかったため「成功」，c，d，e を稼働後不具合数が計画値よりも多かったため「失敗」と考える．

3.3 前処理

IPA/SEC のデータに収集されているメトリクスには，開発プロジェクトの特徴を示すメトリクスと，規模・工期・工数・信頼性などの実績を示すメトリクスが含まれているが，生産性を示すメトリクスが含まれていない．そのため，生産性に関するメトリクスは，含まれているメトリクスをもとに計算して追加する必要がある．たとえば，「月あたりの工数」，「月あたりの SLOC」，「月あたりの FP」などである．ここでは，「月あたりの SLOC」=「SLOC 実績値」÷「実績月数_プロジェクト全体」[*2]とする．

また，相関ルールマイニングおよびベイズ識別器は連続値を取り扱うことができないため，連続値のメトリクスは離散値に変換する必要がある．たとえば，「実績開発工数」は中央値以上の場合に “High”，中央値未満の場合に “Low” とした．他の連続値のメトリクスについても同様に中央値で 2 分割した．同じ前処理を行ったメトリクスとしては，「SLOC 実績値」，「実績月数_プロジェクト全体」，「月あたりの工数」，「月あたりの SLOC」，「月あたりの FP」などがある．

相関ルールマイニングでは，メトリクスの属性値ごとの件数が少ない場合には，その属性値がルールに現れないという問題がある．メトリクスの属性値の種類が多い場合には属性値ごとの件数が少なくなるため，いくつかの属性値をまとめる必要がある．たとえば，「開発対象プラットフォーム」というメトリクスは「Windows95/98/Me 系」，「Windows NT/2000/XP 系」など 17 種類の属性値をとり，少ないものでは 1 件しかないものもある．そこで，「Windows クライアント」「Windows サーバ」「UNIX 系」「Linux 系」「その他」という 5 種類の属性値にまとめた．同じ前処理を行ったメトリクスとしては，「Web 技術の利用」，「オンライントランザクション処理」，「主開発言語」，「DBMS の利用」がある．

4. 提案手法

提案手法は，図 1 のような構成になっている．まず，データに対する前処理（3.3 節で述べた）を行う．次に，データセットの欠損率の向上を実現する．最後に，予測を行う．欠損率の改善は次の 3 つのフェーズからなる．**図 3** に提案法の適用によってデータセットが変化する様子を描いている．

4.1 フェーズ 1（利用不可能なメトリクスの削除）

与えられるデータ D0 は M0 個のメトリクスと N0 件のプロジェクトからなるとする．本研究ではプロジェクトの設計終了段階での予測を想定している．プロジェクトの設計が終了した段階でまだ入手が不可能なメトリクスを利用不可能なメトリクスと呼び，それらをすべて削除する．ただし，見積り値がその時点で求まるもの（規模，工期，工数，生産性など）については削除しない．さらに，メトリクス「実績の評価（品質）」の値が欠損しているプロジェクトをすべて削除する．この結果，N1 プロジェクト，M1 メトリクスからなるデータ D1 が準備される．

[*1] 実際には 2,056 件のプロジェクトから収集されているが，IPA/SEC やデータ提出企業が「信頼性が低い」と判断しているプロジェクトのデータについては削除している．判断の基準については付録 A.1 で述べる．

[*2] 「SLOC 実績値」，「実績月数_プロジェクト全体」は，いずれもデータ白書に含まれているメトリクスである．

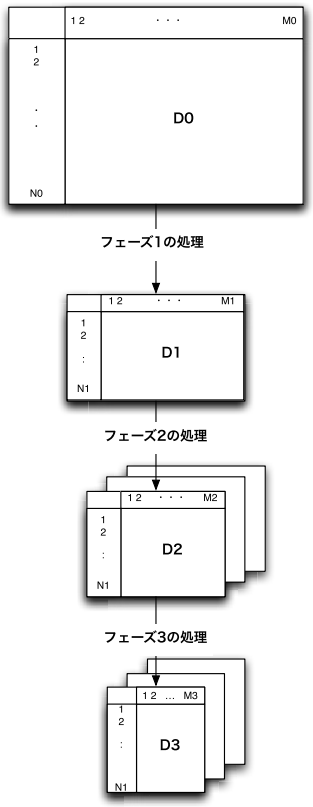

図 3 提案法の適用によるデータセットの変化
Fig. 3 Changing the data set by applying the proposed method.

4.2 フェーズ 2（欠損の多いメトリクスの削除）

データに含まれる欠損が多いメトリクスから順番に削除していき，欠損率が $n\%$ になった時点で削除をやめる．n の値については事前に確定できないので，たとえば $n = 20\%, 15\%, \cdots$ を考える．なお，欠損率が同じメトリクスが複数存在する場合は，同時に削除する．メトリクスの削除が終了した時点でのデータを D2 とすると，そのプロジェクト数は N1 に変化はなく，メトリクス数は減少して M2（$M2 < M1$）となる．

4.3 フェーズ 3（メトリクスの絞り込み）

D2 からプロジェクトの成否に関係のあるメトリクスに絞り込みを行う．絞り込んだ後のメトリクス数は M3 となる．プロジェクト数は N1 のまま変化しない．絞り込みを行った後のデータを D3 とする．フェーズ 3 は次のステップ 1〜ステップ 3 からなる．

ステップ 1 相関ルールの抽出

データ D2 に対して相関ルールマイニング[*3]を行う．相関ルールマイニングとは，相関ルールと呼ばれるデータセットを抽出するデータマイニング手法の 1 つである [5]．相関ルールは，$X_1 \wedge \cdots \wedge X_k \Rightarrow Y$ のような形で表現される．各 X_i（$1 \leq i \leq k$）はアイテム（「メトリクス」=「属性値」）で表現されている．本研究では，プロジェクトの成否と関係のあるメトリクスを選択するために，Y を「実績の評価（品質）」=「成功」に限定している．マイニングの際には，信頼度，支持度と呼ばれるパラメータを変化させてマイニングを行う．

[*3] 相関ルールマイニングについての詳細な説明は付録 A.2 に記載した．

表 1 データの例
Table 1 An example of data.

ID	M_1	M_2	M_3	M_4	品質
1	a	m	p	x	成功
2	a	n	q	y	成功
3	a	m	p	x	成功
4	b	m	p	y	成功
5	b	m	q	x	失敗
6	b	m	q	y	失敗
7	a	m	p	x	失敗

ここでは**表 1** に示す簡単な例を用いてステップ 1〜ステップ 3 の説明を行う．表 1 のデータに対して，最低信頼度 0.7，最低支持度 0.4 として相関ルールマイニングを適用すると，次の 3 つのルールが抽出される．

- R_1：「$M_1 = a$」⇒「品質=成功」
- R_2：「$M_3 = p$」⇒「品質=成功」
- R_3：「$M_2 = m$」∧「$M_3 = p$」⇒「品質=成功」

ステップ 2 相関ルールからアイテムへの分解

抽出した相関ルールの前提部をすべてアイテムに分解する．ここで得られるアイテムの集合を I とする．

たとえば，ステップ 1 の例で得られた R_1，R_2，R_3 から得られる I は {「$M_1 = a$」,「$M_2 = m$」,「$M_3 = p$」} となる．

ステップ 3 プロジェクトの成否と関係のあるメトリクスへの絞り込み

最後に，目的変数との関係がより強いアイテムに絞り込む．成功プロジェクトと失敗プロジェクトについて，I に含まれる各アイテムの条件が満たされる割合を計算する．あるアイテム i について，成功プロジェクトで条件が満たされる割合を成功寄与率，失敗プロジェクトで条件が満たされる割合を失敗寄与率と呼ぶ．成功寄与率と失敗寄与率の差を寄与率とし，寄与率が大きいアイテムをプロジェクトの成否と関係の強いアイテムと考える．この成否と関連の強いアイテムに含まれるメトリクスを，プロジェクトの成否と関係のあるメトリクスとする．

たとえば，ステップ 2 の例で得られた「$M_1 = a$」の成功寄与率は $3 \div 4 = 0.75$ となり，失敗寄与率は $1 \div 3 = 0.33$ となる．寄与率は $0.75 - 0.33 = 0.42$ となる．同様に計算すると，「$M_2 = m$」の寄与率は -0.25，「$M_3 = p$」の寄与率は 0.42 となる．よって，ここでは「$M_1 = a$」と「$M_3 = p$」が成否と関連の強いアイテムであるとし，M_1 と M_3 を成否と関連の強いメトリクスとする．

4.4 ベイズ識別器による予測

D3 に対してベイズ識別器 [4] を利用して予測モデルを構築する*4．ベイズ識別器を用いる主な理由としては，ベイズ識別器が確率として予測結果を示すこと，欠損のあるデータに対しても適用可能であること，そして，先行研究 [1] から判断すると，ある程度の適用可能性が期待できること，があげられる．

5. 適用実験

3.2 節で説明したように，本研究では IPA/SEC のデータ白書として公開されているプロジェクトデータ [17] を利用してプロジェクトの成否を設計工程の終了時に予測する．実験では，4 章で述べたフェーズ 1〜3 を逐次適用し，フェーズ 3 で最終的に絞り込まれた指標を用いてプロジェクト成否予測を行った．この際に，この成否予測で利用したメトリクスについて，実績値と計画値の乖離についても評価した．

以下，5.1 節ではこの実験の狙いと実験全体のスキームを整理する．5.2 節ではフェーズ 1〜3 の適用によってデータ欠損を処理しながら予測モデルを作成し，そのモデルを用いてプロジェクト成否を予測し，その予測精度の変化を説明する．次に，予測に用いたメトリクスに関する実績値と計画値の乖離について調べ，予測モデルと予測結果の妥当性を検討する．さらに，5.3 節ではこの実験を通して確認した予測に利用するメトリクスの削減が予測精度に与える影響について考察を加える．

5.1 実験のスキーム

5.1.1 実験の狙いと構成

実験では 4 章のフェーズ 1〜3 を逐次適用することによってプロジェクト成否予測に用いるデータを準備し，予測モデルを作成して成否予測を行い，予測精度によってその有用性を評価することを目的とする．このため，IPA/SEC のデータ白書として公開されている 237 プロジェクトのデータセットに対し 3.3 節に示した前処理を施したうえで，以下の 4 つの実験を実施した．

実験 1 対象プロジェクトのデータセットに手を加えずにプロジェクト成否を予測する．

実験 2 フェーズ 1 のみを適用してデータセット（D1）を作成する．これを利用して予測モデルを作成し，プロジェクト成否を予測する．

実験 3 フェーズ 1 およびフェーズ 2 を適用してデータセット（D2）を作成する．これを利用して予測モデルを作成し，プロジェクト成否を予測する．

実験 4 フェーズ 1〜3 を適用してデータセット（D3）を作成する．これを利用して予測モデルを作成し，プロジェクト成否を予測する．

ただし，実験 1 の予測モデルにはプロジェクトの設計段階では手に入らないプロジェクト終了時の実績値（例：テスト工程での検出バグ数，テストケース数，実績の評価（工期），実績の評価（コスト）など）の情報が多く含まれる．そのため，実験 1 で得られる予測精度は参考値として位置づける．実験 2 で得られる予測精度を基準に考えることとする．

予測精度は，評価用データのうち，成功と予測して実際に成功していたプロジェクトと，失敗と予測して実際に失敗していたプロジェクトの和をプロジェクトの総数で割った値とする．予測には Weka [15] を使用する．

5.1.2 実験の手順

実験 1〜4 のプロジェクト成否予測は下記の流れで実施する．

(1) 予測モデルは各データセットのデータを用いる．
(2) まず各データセットを学習用データと評価用データを用意するためにランダムに 4 分割し，4-fold cross validation によって下記 (3)，(4) を繰り返し行う．
(3) 学習用データを用いてベイズ識別器により予測モデルを作成する．
(4) 作成した予測モデルに対して，評価用データを入力しプロジェクトの成否予測を行う．

なお，4-fold cross validation の各試行で予測精度を求め，それらの平均を算出する．

5.2 実験結果

5.2.1 実験 1

実験 1 では対象プロジェクトデータに手を加えずにプロジェクト成否を予測した．前述のとおりプロジェクトの成否を予測するので，メトリクス「実績の評価（品質）」に注目する．分析対象となるプロジェクトの数は 237 となった．237 プロジェクトのうち，186 プロジェクトが成功プロジェクト，51 プロジェクトが失敗プロジェクトであった．

値のまったく記入されていないメトリクスが存在すると予測ツールが利用できないため，値のまったく記入されていないメトリクスをすべて削除した．その結果，メトリクスの総数は 511 となった．この 511 個のメトリクス，237 件のプロジェクトからなるデータを D0’ とする．この時点での欠損率は 74.5%となった．

この D0’ を用いてベイズ識別器により予測モデルを構築し，予測精度を評価した結果を**表 2** に示す．予測精度の平均値は 0.757 となった．

5.2.2 実験 2

実験 2 ではフェーズ 1 のみを適用してプロジェクト成否予測を試みた．提案法ではプロジェクトの設計段階で予測するので，その時点までに（見積りなども含めて）情報が入手できないメトリクスを削除する．その結果，メトリクス

*4 ベイズ識別器についての詳細な説明は付録 A.3 に記載する．

表 2 予測精度（実験 1，実験 2，実験 3）
Table 2 Accuracy (Experiment 1, 2, and 3).

	実験 1	実験 2	実験 3				
			欠損 20%	欠損 15%	欠損 10%	欠損 5%	欠損 0%
試行 1	0.753	0.672	0.695	0.703	0.794	0.746	0.694
試行 2	0.760	0.733	0.783	0.667	0.733	0.759	0.730
試行 3	0.752	0.721	0.690	0.673	0.789	0.738	0.767
試行 4	0.765	0.741	0.733	0.733	0.646	0.717	0.783
平均値	0.757	0.717	0.725	0.694	0.741	0.740	0.743

数が 511 から 69 に変化した．この 69 個のメトリクス，237 件のプロジェクトからなるデータを D1 とする．なお，D0' の欠損率は 74.5%であったが，D1 の欠損率は 43.8%であった．このデータセット D1 を用いて，ベイズ識別器によって予測モデルを構築した．各試行で観測された予測精度を表 2 に示す．これらの予測精度の平均は 0.717 となった．

5.2.3 実験 3

実験 3 ではフェーズ 1 および 2 を適用しデータセット D2 を作成し，予測モデルを作成してプロジェクト成否を予測した．ここでは，データに欠損が多く含まれるメトリクスに着目し，欠損の多いメトリクスを段階的に削除していく．具体的には，欠損率が 20%，15%，10%，5%，0%となる 5 種類のデータセット D2 を作成した．その結果，残ったメトリクスの数はそれぞれ 32，27，24，22，11 となった．このデータセットを用いてプロジェクトの成否予測を行うと，その予測精度は 0.725，0.694，0.741，0.740，0.743 となる．各試行で観察された予測精度は表 2 に示す．この結果から，基本的にはフェーズ 2 までの欠損処理を行うことにより，フェーズ 1 のみを実施した実験 2 の結果（0.717）よりも良好な予測精度が得られている．例外的に欠損 15%にしたときに予測精度が実験 2 よりも下がっている．これは，欠損の多いメトリクスを削除するときに目的変数と関係の強いメトリクスが削除されたことが原因であると考えられる．

5.2.4 実験 4

実験 4 では，フェーズ 1〜3 を適用しデータセット D3 を作成し，これを利用して予測モデルを作成し，プロジェクト成否を予測した．ここでは，各学習用データに対して，相関ルールマイニングを適用した．最低信頼度は 0.9 と設定した．最低支持度は各学習用データの成功プロジェクトの 1/3 で成立するルールが抽出されるように設定した．相関ルールマイニングには R [10] を使用した．次に，得られた相関ルールをアイテムに分解し，得られたアイテムに基づいてメトリクスを絞り込んだ．最後に，絞り込まれたメトリクスに基づいて D3 を作成した．絞り込んだ結果のメトリクスの数を**表 3** にまとめた．たとえば，欠損 10%の試行 1，2，3，4 で 63，9，101，43 個のルールが抽出された．ルールを分解して得られたアイテムはそれぞれ 17，12，16，14 個であった．最終的に絞り込まれたメトリクスの数はそれぞれ 7，4，9，5 個であった．

表 3 絞り込まれたメトリクスの数
Table 3 Number of extracted metrics.

	欠損 20%	欠損 15%	欠損 10%	欠損 5%	欠損 0%
試行 1	6	7	7	7	2
試行 2	4	7	4	4	0
試行 3	9	7	9	6	2
試行 4	6	7	5	7	2

表 4 予測精度（実験 4）
Table 4 Accuracy (Experiment 4).

	欠損 20%	欠損 15%	欠損 10%	欠損 5%	欠損 0%
試行 1	**0.776**	**0.800**	**0.828**	0.700	0.783
試行 2	0.717	0.729	0.717	0.768	—
試行 3	0.738	0.700	0.738	0.763	0.780
試行 4	0.690	0.707	0.690	**0.793**	**0.793**
平均値	0.730	0.734	0.743	0.756	0.785

このデータセット D3 を用いて予測をした結果が**表 4** である．表 4 の—は，絞り込みの結果としてメトリクスが 0 個となったために予測を行っていないことを表している．また，それぞれの欠損率について，最も高い精度を太字で表している．すべてのデータのうち最も精度が高かったのは欠損率が 10%の場合で，7 個のメトリクスによって 82.8%を達成している．予測精度の平均値は，欠損率 20%，15%，10%，5%，0%のそれぞれの試行に関して，0.730，0.734，0.743，0.759，0.785 となった．この結果から，実験 4 では実験 2（0.717）よりも高い予測精度が達成できていることが確認できる．

また，同じ欠損率のデータを利用した実験では，実験 3 よりも実験 4 のほうが高い予測精度を達成できていることも確認できる．たとえば，欠損率 15%の場合では，実験 3 では 0.694 となっているが，実験 4 では 0.734 となっている．

これらの事実から，フェーズ 3 が最も有効に働いていると結論できる．

5.3 予測結果の妥当性に関する検討

提案法ではすでに終了したプロジェクトの実績データを用いて予測モデルを構築している．このモデルの中で利用

表 5 絞り込まれたメトリクスの関係
Table 5 Extracted metrics.

メトリクス	欠損 20% 試行 1	欠損 15% 試行 1	欠損 10% 試行 1	欠損 5% 試行 4	欠損 0% 試行 4
開発プロジェクトの種別（新規か改良か）		x		x	x
稼働後品質の目標は妥当か	x		x		
アーキテクチャ		x	x	x	
主開発言語	x	x	x	x	x
DBMS の利用	x	x	x	x	
SLOC	x	x	x	x	
開発工数		x			
月数_プロジェクト全体	x		x		
月あたりの SLOC	x	x	x	x	
月あたりの工数				x	

しているメトリクスには，A. プロジェクトの性質を示すものと，B. 文献 [9] で中核メトリクスと呼ばれているものが含まれる．**表 5** の上から 5 つ「開発プロジェクトの種別」，「稼働後品質の目標が妥当か」，「アーキテクチャ」，「主開発言語」，「DBMS の利用」がプロジェクトの性質を示している．これらのメトリクスは，いずれも見積り値ではなく，設計が終了した時点で計測が可能なものである．一方，B の中核メトリクスとしては残りの 5 つ「SLOC（機能量に該当）」，「月数_プロジェクト全体（開発期間に該当）」，「開発工数（工数に該当）」，「月あたりの SLOC（生産性に該当）」，「月あたりの工数（生産性に該当）」が該当するが，これらは設計終了時点では実績値は確定しないため，計画値を利用する必要が生じる．ここで重要となるのは，これらのメトリクスの実績値と計画値の間にどの程度の乖離があるかという問題である．

本研究で用いたプロジェクトデータについて分析を加えているデータ白書 [17] から，SLOC，工期，工数の計画値と実績値の乖離は，25 パーセンタイル～75 パーセンタイルの範囲で SLOC では 1～1.34 倍，工数では 0.98～1.30 倍，工期では 1～1.05 倍であることが分かっている．これは，たとえば SLOC ではデータ分布の中央の 50%のプロジェクトで実績値は計画値の 1～1.34 倍の範囲に収まっているということである．そのため，多くのプロジェクトでは計画値と実績値の乖離はそれほど大きくなく，実績値のかわりに計画値を利用することが可能であると考えられる．ただし，計画値をそのまま実績値として使うのではなく，1.x 倍などの係数をかけて利用する必要があると考えられる．実データの分析から，失敗プロジェクトの方が成功プロジェクトより乖離が大きくなるという傾向が観測されているため，乖離が大きくなるという前提で計算する必要があると考えている．

実際に適用する際には，蓄積された計画値と実績値から，乖離の範囲を計算して係数を考える．分析に利用したデータの場合，計画値と実績値のずれは，25 パーセンタイル～75 パーセンタイルの範囲で工数が 0.90～1.20，SLOC が 0.93～1.21，月数が 0.95～1.25 の範囲で変動している．そこで，計画値に，工数であれば 1.20 倍，SLOC であれば 1.20 倍，月数であれば 1.25 倍して分析に利用するとよいと考えている．なお，月あたりの SLOC，月あたりの工数に関しては，計画値に関するデータが十分に集まらなかったため，計画値と実績値の乖離が計算できなかった．

5.4 メトリクス削減が予測精度に与える影響

プロジェクト成否の予測を考える場合，予測に用いるメトリクスを削減すると予測精度が低下する可能性が考えられる．たとえば，実験 1 と実験 2 を比較すると，予測の精度は低下している．しかし，実験 2 と実験 4，あるいは実験 3 と実験 4 を比較すると，メトリクスを削減しているにもかかわらず，予測精度は向上している．この事実を引き起こした原因については次のように考えている．

付録 A.3 に示す式のとおり，ベイズ識別器は各メトリクスの値とプロジェクト成否の条件付き確率に基づいてプロジェクトの成否を予測している．フェーズ 3 ではメトリクスの絞り込みを行う際に，相関ルールマイニングを行ったうえで，成功への寄与が大きいメトリクスだけに絞り込みを行っている．フェーズ 3 で計算している成功寄与率と失敗寄与率は，プロジェクト成否の条件付き確率と同じものである．したがって，成功寄与率と失敗寄与率の差が大きいメトリクスを選択することによって，プロジェクトの成否予測の精度が上がるようなメトリクスが選択されることになる．このため，実験 4 ではメトリクス数を削減しているにもかかわらず予測精度が向上したと考えられる．

同様に予測精度の向上を目指してメトリクスを削減した研究としては文献 [1] がある．

6. まとめ

本研究では，現実に企業で収集されているソフトウェア開発プロジェクトのデータを利用してプロジェクトの成

否を設計段階で予測するための方法を提案した．提案法では，まず収集されているデータから欠損の多いメトリクスを削除し，次に相関ルールマイニングを利用してプロジェクトの成否と関係のあるメトリクスだけに絞り込む．プロジェクト成否の予測モデルの構築にはベイズ識別器を利用する．相関ルールマイニングとベイズ識別器はいずれも欠損のあるデータに対しても適用可能な手法であるため，プロジェクトデータからすべての欠損を削除できないような場合であっても適用可能である．

IPA/SEC が収集しているソフトウェア開発プロジェクトのデータを利用して，提案法の適用実験を行った．その結果，最も予測精度が高い場合でメトリクスは 7 個まで絞り込まれ，精度は 82.8%となった．

今後の課題としては，他のデータに対する提案手法の適用，メトリクスを削減して精度が向上したことに対する理論的な考察，従来のメトリクス削減手法との比較があげられる．

謝辞 この研究の一部は，日本学術振興会科学技術研究費補助金基盤研究（C）（課題番号：21500035），および日本学術振興会科学技術研究費補助金特別研究員奨励費（課題番号：21・3963）の助成を受けている．

参考文献

[1] Abe, S., Mizuno, O., Kikuno, T., Kikuchi, N. and Hirayama, M.: Estimation of Project Success Using Bayesian Classifier, *Proc. 28th International Conference on Software Engineering* (*ICSE2006*), Shanghai, China, pp.600–603 (2006).
[2] Cartwright, M.H., Shepperd, M.J. and Song, Q.: Dealing with Missing Software Project Data, *IEEE International Symposium on Software Metrics*, p.154 (2003).
[3] Chen, Z., Boehm, B., Menzies, T. and Port, D.: Finding the Right Data for Software Cost Modeling, *IEEE Software*, Vol.22, pp.38–46 (2005).
[4] Dura, R.O., Hart, P.E. and Stork, D.G.: *Pattern Classification*, John Wiley & Sons, Inc. (2001).
[5] Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers (2001).
[6] Internatinal Software Benchmarking Standards Group: ISBSG Estimating, Benchmarking and Research Suite Release 11 (2009), available from ⟨http://www.isbsg.org/⟩.
[7] Kläs, M., Nakao, H., Elberzhager, F. and Münch, J.: Predicting Defect Content and Quality Assurance Effectiveness by Combining Expert Judgment and Defect Data – A Case Study, *Proc. 2008 19th International Symposium on Software Reliability Engineering*, Washington, DC, USA, IEEE Computer Society, pp.17–26 (2008).
[8] Myrtveit, I., Stensrud, E. and Olsson, U.H.: Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, *IEEE Trans. Softw. Eng.*, Vol.27, pp.999–1013 (2001).
[9] Putnam, L.H. and Myers, W.: *Five Core Metrics: The Intelligence Behind Successful Software Management*, Dorset House Publishing Company, Incorporated (2003).
[10] R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2010).
[11] Shepperd, M. and Schofield, C.: Estimating Software Project Effort Using Analogies, *IEEE Trans. Softw. Eng.*, Vol.23, pp.736–743 (1997).
[12] Strike, K., Emam, K.E. and Madhavji, N.: Software Cost Estimation with Incomplete Data, *IEEE Trans. Softw. Eng.*, Vol.27, pp.890–908 (2001).
[13] Takagi, Y., Mizuno, O. and Kikuno, T.: An Empirical Approach to Characterizing Risky Software Projects Based on Logistic Regression Analysis, *Empirical Software Engineering*, Vol.10, No.4, pp.495–515 (2005).
[14] Twala, B., Cartwright, M. and Shepperd, M.: Comparison of various methods for handling incomplete data in software engineering databases, *International Symposium on Empirical Software Engineering*, p.10 (2005).
[15] Weka Machine Learning Project: Weka, available from ⟨http://www.cs.waikato.ac.nz/˜ml/weka⟩.
[16] （独）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（編）：ソフトウェア開発見積りガイドブック—IT ユーザとベンダにおける定量的見積りの実現，オーム社 (2006).
[17] （独）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（編）：ソフトウェア開発データ白書 2008，日経 BP 社 (2008).
[18] 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，情報処理学会論文誌，Vol.46, No.5, pp.1155–1164 (2005).
[19] 浜野康裕，天嵜聡介，水野 修，菊野 亨：相関ルールマイニングによるソフトウェア開発プロジェクト中のリスク要因の分析，コンピュータソフトウェア，Vol.24, No.2, pp.79–87 (2007).
[20] 瀧 進也，戸田航史，門田暁人，柿元 健，角田雅照，大杉直樹，松本健一：プロジェクト類似性に基づく工数見積もりに適した変数選択法，情報処理学会論文誌，Vol.49, No.7, pp.2338–2348 (2008).

付 録

A.1 データの信頼性について

実験に用いるプロジェクトのデータを選択する際には，データ白書 [17] の「本データの信頼性」「各社評価の本データの信頼性」の 2 つのメトリクスも参考にした．これらのメトリクスは，プロジェクトデータ信頼度を評価したメトリクスであり，a は「データに合理性があり，完全に整合していると認められる」，b は「基本的には合理性があると認められるが，データの整合性に影響を及ぼす要因がいくつか存在する」，c は「重要なデータが提出されていないため，データの整合性を評価できない」，d は「データの信頼性に乏しいと判断できる要因が 1 つもしくは複数見受けられる」という意味である．

適用実験では，「102_本データの信頼性」「10085_本データの信頼性」のいずれかの値が c または d になっているようなプロジェクトはすべて削除した．

A.2 相関ルールマイニング

ある1つのプロジェクトのデータをアイテムの集合で表現するとき，その集合をトランザクションと呼ぶ．全アイテムの集合を $I = \{i_1, i_2, \cdots, i_m\}$，$I \neq \emptyset$ として，その部分集合をアイテムセットと呼ぶ．D を全トランザクションの集合とする．本手法の場合，トランザクション化された分析データが D となる．各トランザクション T は I の部分集合である．$\emptyset \neq X$，$Y \in I$ で，かつ $X \cap Y = \emptyset$ を満たすものを，相関ルール $X \Rightarrow Y$ と呼ぶ．ここで，X を相関ルールの前提部，Y を結論部と呼ぶ．結論部を固定することで，結論部を説明するルールのみを抽出することができる．

相関ルールを評価するパラメータとして，支持度（support）と信頼度（confidence）がある．アイテムセット X の支持度 $sup(X)$ とは，D 中の X を含むトランザクションの割合であり，ルール $X \Rightarrow Y$ の支持度 $sup(X \Rightarrow Y)$ は $sup(X \cup Y)$ で表される．また信頼度 $conf(X \Rightarrow Y)$ は $sup(X \cup Y)/sup(X)$ で定義される．

相関ルールを抽出する際には，信頼度と支持度に最低値を設定して，条件を満たすルールのみを抽出する．

A.3 ベイズ識別器

A.3.1 ベイズの定理

ベイズの定理とは，事前確率を事後確率に変換するもので，あるデータが得られたとき，その結果を反映した下での事後確率を求めるのに使われる [4]．

確率変数 A, B において，

- 事前確率：$P(B)$ = 事象 B が発生する確率
- 事後確率：$P(B|A)$ = 事象 A が起きた後に事象 B が発生する確率

とすると，$P(A) > 0$ の条件の下で

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

が成り立つ．これをベイズの定理という．

A.3.2 ベイズ識別器

データ d の属性集合を $\{q_1, q_2, \cdots, q_n\}$ とする．属性集合が，$q_1 = Q_1, q_2 = Q_2, \cdots, q_n = Q_n$ と与えられたとき，名義変数 c が $c = C$ となる確率

$$P(c = C | q_1 = Q_1 \wedge q_2 = Q_2 \wedge \cdots \wedge q_n = Q_n)$$

は，ベイズの定理を用いて次のように表される．

$$\frac{P(c = C)P(q_1 = Q_1 \wedge \cdots \wedge q_n = Q_n | c = C)}{P(q_1 = Q_1 \wedge \cdots \wedge q_n = Q_n)}$$

この数式を利用して，それぞれのプロジェクトが成功，失敗のいずれのクラスに分類されるかを計算し，予測に用いている．

出張 純也 （学生会員）

2009 年大阪大学大学院情報科学研究科修了．現在，大阪大学大学院情報科学研究科博士後期課程 3 年．ソフトウェア開発プロジェクトのデータに対する定量的な分析に関する研究に従事．IEEE 学生会員．

菊野 亨 （フェロー）

1975 年大阪大学大学院博士課程修了．工学博士．同年広島大学工学部講師．同大学助教授を経て，1987 年大阪大学基礎工学部情報工学科助教授．1990 年同大学教授．現在，大阪大学大学院情報科学研究科教授．大阪大学国際交流センター・センター長．主にフォールトトレラントシステム，ソフトウェア開発プロセスの定量的評価に関する研究に従事．電子情報通信学会，情報処理学会各フェロー．ACM，IEEE 各会員．日本信頼性学会前会長．

菊地 奈穂美 （正会員）

1985 年新潟大学理学部数学科卒業．1993 年 Stanford 大学コンピュータ・サイエンス修士．2006 年大阪大学大学院基礎工学部研究科博士後期課程修了．博士（工学）．沖電気工業（株）にて，通信ソフトウェア仕様記述言語 SDL，MSC 等の設計・検証システムの研究開発，プロジェクトマネジメント，ソフトウェアのプロセス・品質の評価・改善，設計・管理技法等の研究・新技術導入等に従事．2004～2008 年 IPA/SEC にて定量的ソフトウェア管理，見積り手法の研究開発に携わる．IEEE 会員．

平山 雅之 （フェロー）

1986 年早稲田大学大学院理工学研究科修了．1986 年東芝入社．2003 年大阪大学大学院基礎工学研究科修了．博士（工学）．2007 年東海大学専門職大学院客員教授．2011 年日本大学理工学部教授．組込みソフトウェアに関するエンジニアリング手法の研究と普及に従事．専門はソフトウェア品質・信頼性．情報処理学会フェロー．日本品質管理学会所属．