様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２２ 日現在

機関番号：２５３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１２４６

研究課題名（和文）プロジェクトデータ資産を積極的に活用する工数見積もりモデルに関する研究

研究課題名（英文）Research on Effective Project Data Utilization for Software Effort Estimation Models

研究代表者

天嵜　聡介（Amasaki, Sousuke）

岡山県立大学・情報工学部・准教授

研究者番号：００４３４９７８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：時間的近接性が低い過去プロジェクトデータであっても適切に選別を行うことで工数見積り精度の向上に有用な学習データを構築できることを明らかにした。また、他組織のデータを利用する不具合モジュール予測向けの手法が単一組織の過去のデータを活用する状況でも有用であることを明らかにした。また、異なる特性を持つプロジェクトデータを活用した工数見積り予測に有用なモデルの構築の手法を実証的に示した。

研究成果の学術的意義や社会的意義
開発プロジェクトが少なくデータがすぐ古びてしまう組織でも過去プロジェクトデータを利用して一定の精度で工数を見積もることが可能という実務における新たな知見となる。また、新しいアプリケーション分野に参入した時に他組織のデータで工数見積りを必要とする開発組織にとって有用な知見をもたらした。不具合モジュール予測についての知見は、継続的にソフトウェアを開発する組織における不具合発見の効率化に役立つ。

研究成果の概要（英文）：This project revealed that training data for quality effort estimation can be organized even with far past project data. Also, it was found that cross-project defect prediction approaches were effective for cross-version defect prediction. An empirical experiment shows a useful effort estimation strategy where training data have different characteristics from a target organization.

研究分野：ソフトウェア工学

キーワード：工数見積り　転移学習　不具合モジュール予測

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ソフトウェア開発に要する工数を正確に見積もることは、ソフトウェア開発プロジェクトの成否に大きく影響することが知られている。実際より大きく見積もられたソフトウェア開発プロジェクトは、開発計画自体がキャンセルされることがある。実際より小さく見積もられた場合、計画通りに開発が進まず、納期の遅延や品質の低下などを引き起こす。長年の経験と勘に基づいた専門家による見積りと相補的なものとして、工数見積りモデルの研究が長年に渡り盛んに行われている。

工数見積りモデルは、過去のソフトウェア開発プロジェクトの特徴を定量的尺度（メトリクス）によって数値化し、統計的手法や機械学習と組み合わせて構築する。工数見積もりモデルの入力となるプロジェクトデータの特性は見積り精度に大きな影響を及ぼす要因の一つである。統計的手法や機械学習を用いるため、ある程度のプロジェクトデータの量がないと高い精度は得られない。一方、ソフトウェア開発組織は継続的なプロセス改善や開発者の成長、新規アプリケーション分野への参入など様々な要因に晒されており、プロジェクトの特性は絶えず変化している。その結果、工数見積りモデルの構築に使用したプロジェクトデータの特徴が開発組織の現状と異なるという状況が発生する。

２．研究の目的

工数見積りモデルの精度を向上するために有益な知見をもたらすことが目的である。プロセス改善などに取り組んでいるソフトウェア開発組織はプロジェクトの特徴が時間とともに大きく変化している。古いプロジェクトデータを除外する方法もあるが、本研究では、古いプロジェクトデータを積極的に活用することに主眼を置いて調査した。

３．研究の方法

(1) 時間的近接性の情報を利用した古い過去プロジェクトデータを活用する方法の調査：
先行研究では、過去プロジェクトデータのうち古いものは除外されていた。精度向上のために時間的近接性が高い（直近）プロジェクトからさらに有用なデータを絞り込む（データ数減少）か、古い過去プロジェクトデータのうち有用なデータを取り込む（データ数増加）か、どちらの方法がより有効であるか調査した。

(2) 時間的近接性に着目した場合における他組織のデータを利用する不具合モジュール予測（CPDP）手法の有効性の調査：
プロジェクトデータの特徴の変化は、時間経過だけでなく新規分野への参入によっても生じる。不具合モジュール予測の分野では、異なる分野（体制的には他組織と見做せる）のデータを利用する方法（CPDP）が提案されている。工数見積りへの適用を想定し、時間的近接性の違いに対しても CPDP 手法が有効であるか調査した。

(3) 異なる特性を持つプロジェクトデータを活用した工数見積り予測（CCSEE）に CPDP を適用するための調査：
CPDP を CCSEE に適用するにあたっては、分類問題と回帰問題の違いを考慮する必要がある。CPDP 手法の CCSEE への適用可能性やアプローチの網羅性などについて CPDP 手法を調査した。また、調査に基づいて CCSEE における有効性を実証的に調査した。

４．研究成果

(1) 過去プロジェクトデータの時間的近接性と見積対象プロジェクトに対する特性の類似度を組み合わせた絞り込みよりも、時間的近接性が低い過去プロジェクトデータからのデータ取り込みの方が、工数見積りモデルの学習データ構築に有用であることを明らかにできた。

時間的近接性による絞り込みと見積対象プロジェクトに対する特性の類似度による更なる絞り込みを行う手法 A を作成した。時間的近接性によるデータの絞り込みは、先行研究[1]で用いられている「直近に組織が経験したプロジェクトの数」に基づく方法と「直近の一定期間（カレンダー時間）」に基づく方法の二通りを採用した。更なる絞り込みには k 近傍法を利用する方法を

採用した。また、時間的近接性による絞り込みに加えて、時間的近接性が低い過去プロジェクトデータから有用なデータを絞り込む手法 B を作成した。時間的近接性は手法 A と同じである。絞り込みには手法 A と同じ手法を採用したが、特性の類似度を測る基点として、時間的近接性が高い過去プロジェクトデータを参照した。

単一組織の過去プロジェクトデータを用いた手法 A の実証的実験から以下のことが明らかとなった。まず、k 近傍法を利用する絞り込み単体では、工数見積りモデルの構築に利用する過去プロジェクトデータが均質化され、工数見積りの精度が向上した。次に、時間的近接性と特性の類似性による絞り込みを行った手法 A では、それぞれ単体で用いた場合に得られた工数見積り精度の向上が確認できなくなった。互いの効果が打ち消されたと考えられる。一方、同様の条件における手法 B の実証実験では、それぞれの手法を単独で適用した場合よりも有意に見積り精度が向上する条件が存在することが確認できた。

以上のことから、時間的近接性が低い過去プロジェクトデータであっても適切に選別を行うことで工数見積り精度の向上に有用な学習データを構築できることが明らかにできた。この成果は開発プロジェクトが少なくデータがすぐ古びてしまう組織でも一定の精度で工数を見積もることが可能という実務における新たな知見となる。

(2) 時間経過やアプリケーション分野の違いなどによって特徴が異なったプロジェクトデータを用いる工数見積りへの適用を念頭に、他組織のデータを利用する不具合モジュール予測（CPDP）向けの手法が単一組織の過去のデータを活用する状況（CVDP）でも有用であることを示した。

工数見積り分野と同様に予測モデルの精度向上が研究されている不具合モジュール予測分野では、予測モデルの学習に利用するデータを過去のリリースに求める方法と他組織（他のプロジェクトなど）に求める方法とが検討されてきた。前者の方法では、時間的近接性に着目してデータはそのまま利用することが多い。後者の方法では、他組織のデータをそのまま利用するだけでなく、予測対象の特徴に合わせて変形させる手法が提案されてきた。工数見積りでは前者の場合でもデータの選別（変形）が有効であることを確認できたが、不具合モジュール予測向けの手法が前者の場合でも有効であるか不明であった。後者の手法を工数見積りに適用するにあたって、前者の状況でも有用であるか検証した。

先行研究[2]で実装が公開されている不具合モジュール予測手法向けに提案された 24 種類の手法の有効性を実証的に検証した。データ利用のシナリオとして、(a) 最も古い過去プロジェクトデータ、(b)最も直近の過去プロジェクトデータ、(c) 全ての過去プロジェクトデータ、(d) 他組織のデータ＆(a)、(e)他組織のデータ＆(b)、(f)他組織のデータ＆(c)、(g)他組織のデータのみ、の 7 種類を想定した。

実証実験では、不具合モジュール予測によく用いられる複数の機械学習手法とオープンソースプロジェクトから収集されたデータを用いた。複数のリリースから収集されたデータが時間的近接性についての検証を可能にしている。

実証実験から以下のことが明らかとなった。まず、CVDP で有効な CPDP 手法は一部に限られていた。このことは工数見積りにおいても手法の選別が必要であることを示唆する。次に、過去プロジェクトデータを利用する場合は、最も直近のデータのみを用いるシナリオが最良であった。この知見は工数見積りにおける時間的近接性の議論と一致する。最後に、過去プロジェクトデータがない場合でも、他組織のデータでも一定の精度の予測モデルを構築できる手法が存在する。このことは、工数見積りでも特性が異なるデータで一定の精度の予測モデルが構築できる可能性を示唆する。

以上のことから、他組織のデータを利用する不具合モジュール予測（CPDP）向けの手法が単一組織の過去のデータを活用する状況（CVDP）でも有用であることが明らかにできた。この成果は継続的にソフトウェアを開発する組織における不具合発見の効率化に役立つと考えられる。

(3) 他組織のデータを利用する不具合モジュール予測（CPDP）手法を異なる特性を持つプロジェクトデータを活用した工数見積り予測（CCSEE）向けに変換して、その一部が CCSEE においても有用であることを示した。

CPDP 手法の工数見積りへの適用可能性を調査した。前述の 24 種類の CPDP 手法を調査した結果、16 種類は CCSEE 向けに変換できることが確認できた。変換可能な CPDP 手法について、(a) 半分以上が他組織のデータを 1 つのみ入力として受け取ること、(b)半分以上が予測対象のデータと

類似するように他組織のデータを変換すること、などが明らかとなった。

実証実験に先立ち、上記で確認した CCSEE に適用可能な CPDP 手法で採られている方策が CPDP 全体の方策をどの程度網羅しているか確認した。先行研究[3]を参照した調査の結果、特性が一致するプロジェクトデータを対象とした CPDP 手法の全ての方策が網羅できていることを確認できた。

実証実験では、工数見積り研究でよく使用される複数のデータセットを用いた。また、これらのデータセットの時間的・アプリケーション分野的な特徴の違いを考慮した実験条件を設定した。実証実験の結果、データの変形による効果は限定的であることが確認できた。この結果は CPDP の研究で確認された成果ともある程度整合的であることも確認できた。一方で、工数見積りモデルの構築には集団学習を用いることが大きな役割を果たすことが確認された。この成果は新しいアプリケーション分野で工数見積りを必要とする開発組織にとって有用な知見である。

＜引用文献＞

[1] C. Lokan and E. Mendes, Investigating the use of moving windows to improve software effort prediction: a replicated study, Empirical Software Engineering, 2016, pp. 1-52.
[2] S. Herbold, CrossPare: a tool for benchmarking cross-project defect predictions, in Proc. of ASEW ’15, 2015, pp. 90-96.
[3] S. Hosseini, B. Turhan, D. Gunarathna, A systematic literature review and Meta-Analysis on cross project defect prediction. IEEE Transactions on Software Engineering vol.45, no.2, 2019, pp.111-147.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Sousuke Amasaki	４．巻 25
２．論文標題 Cross-version defect prediction: use historical data, cross-project data, or both?	５．発行年 2020年
３．雑誌名 Empirical Software Engineering	６．最初と最後の頁 1573～1595
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10664-019-09777-8	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sousuke Amasaki, Hirohisa Aman, Tomoyuki Yokogawa	４．巻 27
２．論文標題 An extended study on applicability and performance of homogeneous cross-project defect prediction approaches under homogeneous cross-company effort estimation situation	５．発行年 2022年
３．雑誌名 Empirical Software Engineering	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10664-021-10103-4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　0件／うち国際学会　12件）

１．発表者名 Sousuke Amasaki, Hirohisa Aman, Tomoyuki Yokogawa
２．発表標題 An Exploratory Study on Applicability of Cross Project Defect Prediction Approaches to Cross-Company Effort Estimation
３．学会等名 International Conference on Predictive Models and Data Analytics in Software Engineering（国際学会）
４．発表年 2020年

１．発表者名 Sousuke Amasaki
２．発表標題 Augmenting Window Contents with Transfer Learning for Effort Estimation
３．学会等名 International Workshop on Quantitative Approaches to Software Quality（国際学会）
４．発表年 2020年

１．発表者名
Sousuke Amasaki, Hirohisa Aman, Yokogawa Tomoyuki

２．発表標題
Towards Better Effort Estimation with Cross-Project Defect Prediction Approaches

３．学会等名
Evaluation and Assessment on Software Engineering（国際学会）

４．発表年
2019年

１．発表者名
Sousuke Amasaki, Hirohisa Aman, Yokogawa Tomoyuki

２．発表標題
Applying Cross Project Defect Prediction Approaches to Cross-Company Effort Estimation

３．学会等名
International Conference on Predictive Models and Data Analytics in Software Engineering（国際学会）

４．発表年
2019年

１．発表者名
Sousuke Amasaki

２．発表標題
Exploring Preference of Chronological and Relevancy Filtering in Effort Estimation

３．学会等名
Product Focused Software Process Improvement（国際学会）

４．発表年
2019年

１．発表者名
Sousuke Amasaki

２．発表標題
Cross-version defect prediction using cross-project defect prediction approaches: Does it work?

３．学会等名
The 14th International Conference on Predictive Models and Data Analytics in Software Engineering（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	阿萬　裕久 (Aman Hirohisa) (50333513)	愛媛大学・総合情報メディアセンター・特任教授 (16301)	
研究分担者	横川　智教 (Yokogawa Tomoyuki) (50382362)	岡山県立大学・情報工学部・准教授 (25301)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関