

令和 6 年 6 月 8 日現在

機関番号：25301

研究種目：基盤研究(C)（一般）

研究期間：2021～2023

課題番号：21K11833

研究課題名（和文）新たな領域に踏み出すソフトウェア開発組織のための予測・見積もりモデル

研究課題名（英文）Prediction approaches for software organizations diving into new application domains

研究代表者

天崎 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） 3,200,000円

研究成果の概要（和文）：先行研究で提案されたCPDP手法及びCCSEE手法を網羅的に調査して、CPDPからCCSEE及びCCSEEからCPDPの双方向で適用可能性を明らかにし、また、予測性能の観点から有用性を明らかにした。ソースファイルを想定したCPDP手法について、ソースコードの変更差分を想定した場合の有用性を明らかにし、また、Cross-Personalized Defect Prediction という新しい適用先を示した。さらに、直接ソースコードや機能要求のテキストを入力とする工数見積もり手法及び不具合モジュール予測手法において、他の開発組織のプロジェクトデータの有用な利用形態についての可能性を示した。

研究成果の学術的意義や社会的意義

本研究で得られた成果のうち、CCSEEとして有用な手法及びCPDPとして有用な手法は、新しい領域に踏み出す過程で発生するプロジェクト固有データの不足を補うために多くのソフトウェア開発組織で活用できる知見である。また、Cross-Personalized Defect Predictionは本研究で新たに見出された適用先であり、学術的にも今後の発展が期待される。さらに近年発達しているテキストを直接埋め込みベクトルに変換して行う工数見積もりや不具合モジュール予測の研究において、事前学習のデータ量が重要であるという知見は今後の同分野の研究の方向に影響を及ぼすものである。

研究成果の概要（英文）：This study revealed the applicability of CPDP to CCSEE and CCSEE to CPDP based on the approaches proposed in past studies. The effectiveness of the applicable CPDP and CCSEE approaches on different domains was also shown regarding the predictive performance. The applicability of CPDP for file-level prediction to change-level prediction was also discussed and evaluated empirically. Cross-Personalized Defect prediction was identified as a new application area of CPDP approaches. Furthermore, effort estimation and defect prediction based on text embedding vectors are examined to suggest effective usages of others' datasets.

研究分野：ソフトウェア工学

キーワード：工数見積もり 不具合モジュール予測 組織横断データの活用

## 様式 C - 19、F - 19 - 1、Z - 19 (共通)

### 1. 研究開始当初の背景

ソフトウェア開発では、プロジェクトの将来に見通しが立つよう様々な予測・見積もりが行われる。このことは予測・見積もりの正確さがプロジェクトの成否に大きな影響を及ぼすことを意味する。例えば、実際より工数が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。昨今、開発規模やコード行数など、定量的尺度(メトリクス)によって数値化されたソフトウェア開発プロジェクトの特徴を統計的手法や機械学習と組み合わせる予測・見積もりモデルの研究が盛んである。

一般的には、高性能なモデルの構築には予測・見積もり対象のプロジェクトと類似した特性のプロジェクトデータが必要である。しかしながら、新しい領域での開発に踏み出した組織にとって過去のプロジェクトデータはこの条件を満たさない。また、同じ領域の他の開発組織のプロジェクトデータを単純に融通するだけでは必ずしも十分な性能は得られない。

近年、対象と特性が異なるプロジェクトデータを活用する手法の研究に注目が集まってきている。特に、工数見積もりと不具合モジュール予測に関する研究が顕著である。不具合モジュール予測ではこのような問題設定を Cross-Project Defect Prediction (CPDP) と呼び、既に多数の手法が提案されサーベイ論文[1]も出版されている。工数見積もりでは Cross-Company Software Effort Estimation (CCSEE) と呼ばれる研究分野となっており、例えば[2]などの手法が提案されている。

### 2. 研究の目的

組織固有のプロジェクトデータが十分に蓄積されていない状況において、ソフトウェア開発タスクに関連したメトリクスを入力とする予測・見積もりモデルの予測性能向上に有益な知見をもたらすことが目的である。ソフトウェア開発組織が新たなビジネス領域に踏み出した際の状況を想定している。ソフトウェア開発タスクとして工数見積もりと不具合モジュール予測を対象とする。

### 3. 研究の方法

- (1) CPDP 向けに提案されている手法のうち CCSEE でも有用な手法の調査：  
先行研究、特にサーベイ論文を参照して広い範囲で様々な CPDP 手法を調査・収集した。ただし、CPDP 手法は毎年数多く提案されていることを考慮して、実装が公開されている CPDP 手法を優先して評価した。また、アプローチの多様さを基準として選別を行った。有用性を評価するために工数見積もりモデルの入力となるプロジェクトデータを調査・収集した。先行研究で利用されているものを中心に収集した。
- (2) CCSEE 向けに提案されている手法のうち CPDP でも有用な手法の調査：  
先行研究を参照して広い範囲で様々な CCSEE 手法を調査・収集した。CPDP に比べて CCSEE に関する研究は相対的に少ないことを考慮して、実装が公開されているものに限らず手法を収集した。有用性を評価するために不具合モジュール予測モデルの入力となるデータを調査・収集した。先行研究で利用されているものを中心に収集した。不具合モジュール予測では予測の粒度がファイルや変更差分など様々である。異なる粒度のデータを収集した。
- (3) CPDP 及び CCSEE の双方で有用な手法の調査：  
前述の(1)と(2)で有望と判断された手法を比較・分析して双方で有用な手法の共通点などを調査した。また、異なるアプローチをとる手法を組み合わせる方法を検討した。

### 4. 研究成果

- (1) 先行研究で提案された CPDP 手法及び CCSEE 手法を網羅的に調査して、CPDP から CCSEE 及び CCSEE から CPDP の双方向で提案手法の適用可能性を明らかにし、また、予測性能の観点から実証実験により適用可能な手法の有用性を明らかにした。

CCSEE に適用可能な CPDP 手法について先行研究、特にサーベイ論文を重点的に調査した。予測精度の観点から実証実験を行う必要があるため、実装が公開されている 24 の CPDP 手法が採用しているアプローチを特に重点的に分析した。同様の調査を逆の方向でも実施した。その結果、分析対象となった 24 の CPDP 手法のうち、16 の CPDP 手法が CCSEE に適用可能であることが明らかとなった。また、CCSEE 手法の調査では CPDP 手法に適用可能なものはアンサンブル学習など一般的な枠組みで定義されたアプローチが多く、CPDP 手法としてすでに検証が行われているものがほとんどであることが明らかとなった。適用不可能な手法はドメイン知識に依存していることも確認できた。

CCSEE に適用可能な CPDP 手法は CPDP のサーベイ研究[3]で定義された戦略のほぼ全てを含むものであることが確認できた。適用可能な CPDP 手法に含まれていない戦略は他組織のプロジェクトデータを単純にマージする戦略であった。この戦略を加えて予測性能の観点から実証実験による CPDP 手法の有用性を調査した。CCSEE の様々な状況を模すために、開発組織や使用言語、開発ドメインなど複数の観点でプロジェクトデータを分割して評価実験を行った。評価実験で最も性能が高いとみなされた手法は単純なマージであった。他の多くの手法はより複雑な手順をとる一方で工数見積もりの性能は単純なマージを上回るものではなかった。単純なマージは従来の工数見積もりで最も広く採用されている方法であり、この方法が CCSEE において有用であることは新たな発見である。先述のサーベイ研究では、CPDP においては単純なマージは他の手法より劣っているとみなされており、CPDP と CCSEE の違いが明らかとなった。

- (2) プロジェクト間におけるソースファイルを想定した CPDP 手法について、ソースコードの変更差分を想定した場合の有用性を明らかにし、また、プロジェクトに従事する開発者間における Cross-Personalized Defect Prediction という新しい適用先を示した。

不具合モジュール予測は異なる粒度を対象とする研究が広く行われている。ソースファイルは最も基本的な粒度であるが、ファイルに含まれる機能ごとに不具合の有無を予測したい需要がある。また、近年ではソースファイルを細かくバージョン管理するプラクティスが一般化したことから、ソースコードの変更差分を対象に不具合の混入を防ぐ手段として不具合モジュール予測を行うユースケースが現れている。変更差分は開発者と紐づけられて記録されており、不具合の確認を当該開発者に依頼できる利点がある。変更差分を対象とした場合におけるソースファイルを対象とした CPDP 手法の適用可能性を確認した。また、変更差分が開発者と紐づけられているため、CPDP 手法における他のプロジェクトはプロジェクト内の他の開発者と解釈することも可能である。新たな領域に踏み出すソフトウェア開発組織には新たな領域を専門とする開発者がプロジェクトに参加する可能性があるため、後者の解釈による CPDP 手法の適用も有用であると考えられる。そのため、この二通りの観点で適用可能性を調査した。

変更差分を対象とした不具合モジュール予測は先行研究で定義されたメトリクス[4]を入力とするものが多い。このため、他のプロジェクトで収集された変更差分を用いた場合でもソースファイルを単位とした CPDP 手法をそのまま適用した。実証実験による予測性能の評価には変更差分を対象とした不具合モジュール予測の先行研究で用いられた 16 プロジェクトのデータを利用した。21 の CPDP 手法の性能を比較した結果、2 つの CPDP 手法が同程度に他の手法より有意に性能が高いことが明らかとなった。これらは組織内のデータを用いる場合と同程度に良い性能を示していたことから、実用性も高いと考えられる。

他の開発者の変更差分履歴を用いた場合における CPDP 手法の有用性についても、ソースファイルを単位とした CPDP 手法をそのまま適用することで検証した。実証実験による予測性能の評価には変更差分を対象とした不具合モジュール予測の先行研究で用いられた 9 プロジェクトのデータを利用した。23 の CPDP 手法の性能を比較した結果、予測性能に影響を及ぼさず CPDP 手法が複数観測された。9 プロジェクト全てにおいて予測性能を向上できた CPDP 手法は存在しなかった。しかしながら、1 つの CPDP 手法は全体として性能向上に寄与しており、実用的な観点からはこの手法を利用することがよいと言える。

- (3) 直接ソースコードや機能要求のテキストを入力とする工数見積もり手法及び不具合モジュール予測手法において、他の開発組織のプロジェクトデータの有用な利用形態についての可能性を示した。

メトリクスはヒューリスティクスにより人為的に作成される入力であるのに対して、ソースコードや機能要求を記述したテキストを入力とする工数見積もりや不具合モジュール予測は入力を固定長のベクトルに変換する。メトリクスと比べて次元数が多く次元ごとの意味的な解釈も難しい。そこで、これらの対照的な入力を組み合わせた場合とそうでない場合において予測性能を比較した。評価実験では、不具合の一種である脆弱性に関する9つデータセットを用いた。その結果、これらを組み合わせることが性能の向上に繋がる可能性が明らかとなった。また、固定長のベクトルの計算には、ソフトウェア開発に関するデータのみでは不十分である可能性を示した。このことはCPDPで一般的な他のプロジェクトのデータ変換手法の有効性が相対的に低下している可能性を示唆している。

テキストを変換した固定長のベクトルを工数見積もりに利用する場合においてもソフトウェア開発に関するデータの有用性の評価を行った。近年一般的となったアジャイル開発における見積もりで使用されるストーリーポイントを予測の対象として評価実験を行った。脆弱性予測の場合と同様にソフトウェア開発に関するデータが見積もり精度の向上に寄与する度合いは低かった。CCSEEで固定長のベクトルを入力とする場合もCPDPと同様に一般的なデータで事前学習されたモデルからベクトルを計算することが有利に働く可能性を示唆している。

### 参考文献

- [1] Y. Zhou, Y. Yang, H. Lu, L. Chen, Y. Li, Y. Zhao, J. Qian, and B. Xu, "How Far We Have Progressed in the Journey? An Examination of Cross-Project Defect Prediction," *Trans. Softw. Eng. Methodol.* vol. 27, no. 1, 51 pages, ACM.
- [2] L. L. Minku, "A novel online supervised hyperparameter tuning procedure applied to cross-company software effort estimation," *Empirical Software Engineering*, vol. 24, pp. 3153 - 3204, Springer.
- [3] S. Hosseini, B. Turhan, and D. Gunarathna, "A systematic literature review and Meta-Analysis on cross project defect prediction," *IEEE Trans Softw Eng* 45(2), pp.111-147, 2019.
- [4] Y. Kamei, E. Shihab, B. Adams, A. E. Hassan, A. Mockus, A. Sinha, and N. Ubayashi, "A large-scale empirical study of just-in-time quality assurance," *IEEE Trans Softw Eng* 39(6), pp.757-773, 2013.

## 5. 主な発表論文等

〔雑誌論文〕 計5件（うち査読付論文 5件/うち国際共著 1件/うちオープンアクセス 2件）

1. 著者名 OHARA Kosuke, AMAN Hirohisa, AMASAKI Sousuke, YOKOGAWA Tomoyuki, KAWAHARA Minoru	4. 巻 E106.D
2. 論文標題 A Comparative Study of Data Collection Periods for Just-In-Time Defect Prediction Using the Automatic Machine Learning Method	5. 発行年 2023年
3. 雑誌名 IEICE Transactions on Information and Systems	6. 最初と最後の頁 166 ~ 169
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2022MPL0002	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 Leelaprute Pattara, Amasaki Sousuke	4. 巻 150
2. 論文標題 A comparative study on vectorization methods for non-functional requirements classification	5. 発行年 2022年
3. 雑誌名 Information and Software Technology	6. 最初と最後の頁 106991 ~ 106991
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.infsof.2022.106991	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 該当する
1. 著者名 Amasaki Sousuke, Aman Hirohisa, Yokogawa Tomoyuki	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 -
1. 著者名 Aman Hirohisa, Amasaki Sousuke, Yokogawa Tomoyuki, Kawahara Minoru	4. 巻 28
2. 論文標題 An automated detection of confusing variable pairs with highly similar compound names in Java and Python programs	5. 発行年 2023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/s10664-023-10339-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

1. 著者名 積田 静夏、天寄 聡介、林 晋平	4. 巻 65
2. 論文標題 情報検索を用いたBug Localization手法にモジュール粒度の違いが与える影響	5. 発行年 2024年
3. 雑誌名 情報処理学会論文誌	6. 最初と最後の頁 792 ~ 807
掲載論文のDOI (デジタルオブジェクト識別子) 10.20729/00233609	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13件 (うち招待講演 0件 / うち国際学会 11件)

1. 発表者名 Tenma Kita, Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, Minoru Kawahara
2. 発表標題 Have Java Production Methods Co-Evolved With Test Methods Properly?: A Fine-Grained Repository-Based Co-Evolution Analysis
3. 学会等名 Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (国際学会)
4. 発表年 2022年

1. 発表者名 S. Amasaki, H. Aman, T. Yokogawa
2. 発表標題 An Evaluation of Effort-Aware Fine-Grained Just-in-Time Defect Prediction Methods
3. 学会等名 Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (国際学会)
4. 発表年 2022年

1. 発表者名 K. Wayama, T. Yokogawa, S. Amasaki, H. Aman, K. Arimoto
2. 発表標題 Verifying Game Logic in Unreal Engine 5 Blueprint Visual Scripting System Using Model Checking
3. 学会等名 International Workshop on Automated Software Engineering for Computer Games (国際学会)
4. 発表年 2022年

1. 発表者名 S. Amasaki, H. Aman, T. Yokogawa
2. 発表標題 An Evaluation of Cross-Project Defect Prediction Approaches on Cross-Personalized Defect Prediction
3. 学会等名 International Conference on Product-Focused Software Process Improvement (国際学会)
4. 発表年 2022年

1. 発表者名 Shizuka Tsumita, Shinpei Hayashi and Sousuke Amasaki
2. 発表標題 Large-Scale Evaluation of Method-Level Bug Localization with FinerBench4BL
3. 学会等名 IEEE International Conference on Software Analysis, Evolution and Reengineering (国際学会)
4. 発表年 2023年

1. 発表者名 天崎 聡介, 阿萬 裕久, 横川 智教
2. 発表標題 他者の開発履歴を利用したパーソナライズ不具合予測における他組織データを利用した不具合予測手法の適用について
3. 学会等名 第29回ソフトウェア工学の基礎ワークショップ
4. 発表年 2022年

1. 発表者名 S. Amasaki, H. Aman, T. Yokogawa
2. 発表標題 A Preliminary Evaluation of CPDP Approaches on Just-in-Time Software Defect Prediction
3. 学会等名 Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (国際学会)
4. 発表年 2021年

1 . 発表者名 S. Amasaki, H. Aman, T. Yokogawa
2 . 発表標題 Searching for Bellwether Developers for Cross-Personalized Defect Prediction
3 . 学会等名 Product-Focused Software Process Improvement ( 国際学会 )
4 . 発表年 2021年

1 . 発表者名 S. Amasaki
2 . 発表標題 On Effectiveness of Further Pre-training on BERT Models for Story Point 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 2023年

1 . 発表者名 Shinnosuke Irie, Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, Minoru Kawahara
2 . 発表標題 A Comparative Study of Hybrid Fault-Prone Module Prediction Models Using Association Rule and Random Forest
3 . 学会等名 World Symposium on Software Engineering ( 国際学会 )
4 . 発表年 2023年

1 . 発表者名 Yuki Fushihara, Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, Minoru Kawahara
2 . 発表標題 A Trend Analysis of Test Smells in Python Test Code Over Commit History
3 . 学会等名 Euromicro Conference on Software Engineering and Advanced Applications ( 国際学会 )
4 . 発表年 2023年



1. 発表者名 S. Amasaki, H. Aman, T. Yokogawa
2. 発表標題 An Evaluation of Word Embeddings on Vulnerability Prediction with Software Metrics
3. 学会等名 Work-in-Progress Session in Euromicro Conference on Software Engineering and Advanced Applications (国際学会)
4. 発表年 2023年

1. 発表者名 積田 静夏、天寄 聡介、林 晋平
2. 発表標題 異粒度情報の統合に基づく細粒度Bug Localization
3. 学会等名 SIGSS2024年3月研究会
4. 発表年 2024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・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
---------	---------