

CCDP 手法の WCDP への適用について

天 寄 聡 介^{†1}

他組織のプロジェクトデータを用いる cross-company 不具合予測 (CCDP) の研究が盛んに行われている。近年は自組織のプロジェクトデータを用いる不具合予測 (WCDP) と遜色ない精度で予測できることが報告されている。幾つかの CCDP 手法は WCDP にも適用可能であるが先行研究ではこの点について報告されていない。ワークショップでは WCDP の予測精度向上の観点から CCDP 手法の効果について議論したい。

On applying CCDP method to WCDP

SOUSUKE AMASAKI^{†1}

Cross-company defect prediction (CCDP) studies has been popular in recent years. Some studies demonstrated that its performance is now comparable to that of within-company defect prediction (WCDP). While some CCDP methods look applicable to WCDP, CCDP studies have not cared the applicability. In the workshop, we would like to discuss the applicability in terms of the improvement of WCDP performance.

1. はじめに

ソフトウェア開発において製品の品質を高めることは重要な課題である。効率的に不具合を発見・修正するための技法として不具合予測手法の研究が広く行われている。不具合予測手法の多くは不具合修正の履歴情報を含む過去のプロジェクトデータの利用を前提とするが、そのような情報が存在しない状況は多い。そのため、他の組織やプロジェクトのデータを利用する Cross-company 不具合予測 (CCDP) の研究が注目されている。

CCDP 手法の多くは、他組織のプロジェクトデータの中から自組織のプロジェクトデータに類似したものを選別したり¹⁾、自組織のプロジェクトデータに類似するよう変換を行う²⁾ ことで予測精度の向上を図っている。近年の研究では、CCDP 手法の予測精度が自組織のデータを用いた予測 (WCDP) の予測精度と遜色ない性能を示す場合があることが報告されている。

CCDP 手法の多くがプロジェクトデータの選別や変換に基づいており、WCDP でも適用可能である場合が多い。例えば、Turhan らの提案法¹⁾ で用いられているのは単純な k-近傍法であり、プロジェクトデータの出自に関わらず適用可能である。しかしながら、

CCDP 手法の研究においてはこの適用可能性が考慮されておらず、WCDP へ適用した場合の予測精度の変化については議論されていない。

本研究では、文献 1) の手法を WCDP に適用した結果について報告する。ワークショップでは WCDP の予測精度の向上の観点から CCDP 手法について議論していきたい。

2. 実験方法

文献 1) の提案手法では、他の組織の複数のプロジェクトデータを結合したものから、予測対象プロジェクトの個々のモジュールに類似したデータを k-近傍法によって選別する。この際、重複して選別されたモジュールは一つのみが利用され、重複分は捨てられる。本実験では、同文献と同じく $k = 10$ とした。

プロジェクトデータには、Peters³⁾ らが実験に用いたプロジェクトデータの一部を利用した。Peters らは PROMISE Repository から 35 件の OSS のプロジェクトデータを他組織のプロジェクトデータとして、21 件を自組織のプロジェクトデータとして用いている。本実験では他組織のプロジェクトデータはそのまま、自組織のプロジェクトデータは規模が小さい 10 件を省いた 11 件を利用した。11 件のデータ名、モジュール数、不具合の割合を表 1 に示す。

実験方法は文献 4) を参考に以下のようにした。

^{†1} 岡山県立大学
Okayama Prefectural University

表 1 自組織のデータ

No.	データ名	ファイル数	不具合の割合 (%)
1	berek	43	37
2	e-learning	64	8
3	forrest-0.7	29	17
4	forrest-0.8	32	6
5	pbeans2	51	20
6	pdftranslator	33	45
7	serapion	45	20
8	skarbonka	45	20
9	systemdata	65	14
10	termoproject	42	31
11	workflow	39	51

表 2 実験結果 (AUC 中央値)

	cc_lr	wc_lr	cc_rf	wc_rf	cc_nb	wc_nb
1	0.656	0.744	0.906	0.805	0.787	0.850
2	0.500	0.525	0.724	0.500	0.491	0.640
3	0.500	0.438	0.479	0.542	0.500	0.771
4	0.500	0.417	0.833	0.500	0.500	0.500
5	0.550	0.626	0.664	0.563	0.576	0.516
6	0.536	0.641	0.732	0.729	0.602	0.673
7	0.500	0.531	0.572	0.508	0.649	0.551
8	0.500	0.501	0.735	0.522	0.444	0.669
9	0.500	0.675	0.700	0.645	0.541	0.504
10	0.500	0.544	0.683	0.721	0.680	0.696
11	0.525	0.442	0.572	0.431	0.567	0.594

- (1) 自組織のデータをランダムに 500 回, 2-fold cross-validation の形式に分割する
- (2) 各テストデータに対して, 他組織のプロジェクトデータと対応するトレーニングデータそれぞれに文献 1) の方法を適用する
- (3) (2) の結果を用いた予測の精度を比較する

予測にはロジスティック回帰モデル, RandomForests, NaiveBayes を用いた. 性能評価には AUC ROC を用いた. 実験には `scikit-learn` を用いた.

3. 結果と考察

表 2 は予測精度の中央値を 11 件の自組織データのそれぞれについてまとめたものである. 「cc」と「wc」はそれぞれ CCDP と WCDP を表す. 「lr」「rf」「nb」はそれぞれロジスティック回帰モデル, RandomForests, NaiveBayes を表す.

ロジスティック回帰モデルの結果をみると, WCDP の方が性能が高い精度を示すプロジェクトが 7 件存在した. 一方, CCDP の方が高い精度を示す場合も 3 件存在する. RandomForests の結果をみると, CCDP の方が高い性能を示すプロジェクトが 10 件存在した. WCDP の方が高い精度を示したのは 1 件のみであった. NaiveBayes の結果をみると, WCDP の方が高い精度を示すプロジェクトが 7 件存在した. 一方, CCDP

の方が高い精度を示す場合は 2 件であった.

以上の結果から, 文献 1) の手法を WCDP に適用した場合, CCDP よりも高い精度で予測ができる場合が多いと言える. 一方で, RandomForests の結果が示すように, CCDP の方がより高い精度を示す場合があることも確認された. さらに, テストプロジェクト毎に結果を見ると, 7 件のプロジェクトにおいて CCDP と RandomForests の組み合わせが最も高い精度を示している.

Turhan らの提案法は予測対象モジュール群に類似したモジュールを選別することで精度向上を図っている. WCDP の方が予測対象により類似したモジュールをトレーニングデータに含んでおり, より高い精度の予測モデルを学習できると考えられるが, RandomForests の結果はこの推測に反する. この原因として, WCDP では類似したモジュールが意外と少なく, 複数のプロジェクトデータから選別する方がより多くの類似モジュールを集められる可能性が考えられる. このことは, CCDP の利点が WCDP では得られない可能性を示す.

4. おわりに

本研究では, CCDP 手法を WCDP に適用した場合の予測精度について比較評価を行った. Turhan らの手法を WCDP に適用した結果, CCDP の予測精度の方が高い場合があることを確認した. またこのことから, CCDP 手法の利点が WCDP では十分に得られない可能性を示した.

今後の研究では, 適用する CCDP 手法を増やすなどの方法で WCDP への適用可能性を探っていきたい.

参 考 文 献

- 1) Turhan, B., Menzies, T., Bener, A. B. and DiStefano, J.: On the relative value of cross-company and within-company data for defect prediction, *Empir. Softw. Eng.*, Vol.14, No.5, pp.540–578 (2009).
- 2) Nam, J., Pan, S.J. and Kim, S.: Transfer defect learning, *Proc. of 35th Int'l Conf. on Softw. Eng. (ICSE)*, IEEE, pp.382–391 (2013).
- 3) Peters, F., Menzies, T. and Marcus, A.: Better cross company defect prediction, *Proc. of 10th IEEE Working Conf. on Mining Softw. Repositories (MSR 2013)*, IEEE, pp.409–418 (2013).
- 4) Nam, J.: Heterogeneous defect prediction, *Proc. of the 11th Joint Meeting on Foundations of Softw. Eng. (ESEC/FSE)*, New York, New York, USA, ACM Press, pp.508–519 (2015).