NN-filter と project selection の比較評価

天 嵜 聡 介[†1] 河 田 和 也[†1]

他のプロジェクトのデータを用いる cross-company 不具合予測手法の研究において，予測対象プロジェクトと類似した特性を持つデータを選別する手法が提案されている．選別の単位として，モジュールを考える手法とプロジェクトを考える手法が提案されているが，それらの関係については十分な比較が行われてない．ワークショップでは予測精度の観点からプロジェクトデータの選別方法について議論したい．

Comparisons of NN-Filter and Project Selection

Sousuke Amasaki [†1] and Kazuya Kawata[†1]

Cross-company defect prediction studies proposed two types of data selection methods: module-selection and project-selection methods. owever, they have not been compared in detail in terms of predictive performance. In the workshop, we would like to discuss a way to selecting project data for better cross-company defect prediction.

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 不具合予測手法 (CCDP) の研究が注目されている．

CCDP における精度向上のアプローチの一つとして，予測対象モジュールの特性を考慮して他の組織のプロジェクトデータの中から有用なもののみを選別する手法が提案されている[1),2)]．Turhan ら[1)] は他の組織のプロジェクトデータを複数結合し，その中からモジュール単位で予測に役立つと考えられるデータを選別している．一方，He ら[2)] は複数の他の組織のプロジェクトデータから予測に役立つと考えられるプロジェクトデータを選別している．しかしながら，これらの手法の予測精度については十分な比較が行われておらず，また，これらの手法を組み合わせた場合に予測精度がどのように変化するのかについても明らかでない．

†1 岡山県立大学
Okayama Prefectural University

そこで本研究では，これらの手法の比較および組み合わせについて予測精度の観点から比較評価を行なった．

2. プロジェクトデータの選別手法

文献 1) で提案されている手法 (NN-filter) では，他の組織のプロジェクトデータを結合したものから，k-近傍法によって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選別する手法である．

一方，文献 2) で提案されている手法 (project selection) では，他の組織のプロジェクトデータそれぞれについて代表値を求めて，予測対象プロジェクトの代表値と類似したプロジェクトのデータを複数選別して結合したものを予測に用いる．

3. 実験方法

PROMISE Repository に登録された 14 件のプロジェクトデータを実験に利用した．プロジェクト名，モジュール数，不具合の割合を表 1 に示す．

実験では各プロジェクトを予測対象とみなし，他のプロジェクトのデータを不具合予測モデルの学習データとして用いた．この際，学習データを以下の 4 通りで作成した．

- プロジェクトデータを結合
- プロジェクトデータを結合+NN-filter
- project selection (全体の 60%)

表 1 実験に利用したプロジェクト一覧

データ名	ファイル数	不具合の割合 (%)
ant-1.3	125	16
arc	234	11
camel-1.0	339	3
ivy-1.1	111	57
jedit-3.2	272	33
log4j-1.0	135	25
lucene-2.0	195	46
poi-1.5	237	59
redaktor	176	15
synapse-1.0	157	10
tomcat	858	8
velocity-1.4	196	75
xalan-2.4	723	15
xerces-1.2	440	16

- project selection (全体の 60%) + NN-filter

予測には RandomForests を用い，性能評価については AUC ROC を用いた．

4. 結果と考察

実験結果を図 1 に示す．実験では 14 種類のプロジェクトに対して予測をおこなったため，個々のプロジェクトに対する AUC ROC が 14 通り出力される．箱ヒゲ図はこの 14 通りの結果から作成した．

左端の箱ヒゲ図は，単純にプロジェクトデータを複数結合した場合の予測結果である．中央値が 0.7 より高く一定の予測精度が実現できていると考えられる一方で，予測対象となるプロジェクトもしくは予測手法の学習に用いたプロジェクトデータによって精度に大きなばらつきがある．

右隣の箱ヒゲ図は，このデータに NN-filter を適用した結果である．予測精度のばらつきが減少しており，NN-filter が明らかに予測に不適切なデータを排除できていると考えられる．一方で，中央値は 0.7 付近であり，予測精度の向上は明確とは言えない．

左から 3 番目の箱ヒゲ図は，project selection によって選別したプロジェクトデータを用いた予測結果である．単純に全てのプロジェクトデータを結合する場合と比べて予測精度のばらつきが減少しており，適切なプロジェクトデータを選択できていると言える．一方で，中央値は 0.7 付近であり，予測精度の向上は明確とは言えない．

右端の箱ヒゲ図は，project selection によって選別したプロジェクトデータを結合した後，さらに NN-filter を適用した結果である．NN-filter および project selection を単体で使用した場合と予測精度のばらつきが同程度であり，また，中央値も 0.7 付近である．下方の外れ値の予測精度がやや向上しているものの組み合わせによる相乗効果は明確であるとは言えない．

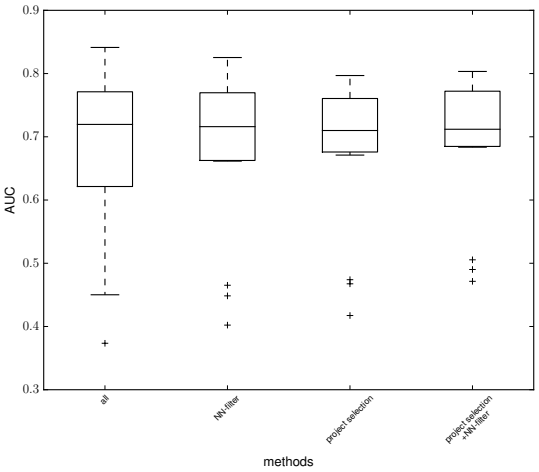

図 1 実験結果

以上の結果から，NN-filter と project selection の効果は同程度であると考えられる．また，組み合わせによる予測精度の向上が観察できないことから，これらの手法では類似したデータを選別している可能性が考えられる．

5. おわりに

本研究では，NN-filter と project selection の比較および組み合わせの影響について調査した．その結果，それぞれの手法による予測精度の向上は同程度であること，組み合わせによる相乗効果が見られないことを明らかにした．

ワークショップでは，予測精度の向上に繋がる手法の組み合わせ方や組み合わせに適した選別手法について議論していきたい．また，今後の研究では，使用するプロジェクトデータの数を増やすなどの方法で更なる検証を行っていきたい．

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, *Empirical Software Engineering*, Vol.14, No.5, pp.540–578 (2009).
2) He, Z., Shu, F., Yang, Y., Li, M. and Wang, Q.: An investigation on the feasibility of cross-project defect prediction, *Automated Software Engineering*, Vol.19, No.2, pp.167–199 (2012).