

科学研究費助成事業 研究成果報告書

平成 26 年 6 月 9 日現在

機関番号：13601

研究種目：基盤研究(C)

研究期間：2011～2013

課題番号：23500043

研究課題名(和文) 予測モデルの再利用による、ソフトウェア欠陥予測ツールの開発

研究課題名(英文) Construction of error prone module reusing prediction model

研究代表者

海尻 賢二 (KAIJIRI, Kenji)

信州大学・工学部・教授

研究者番号：80115336

交付決定額(研究期間全体)：(直接経費) 3,400,000円、(間接経費) 1,020,000円

研究成果の概要(和文)：最初は検査プロジェクトに対していくつかの基準で似ているプロジェクトを選択し、プロジェクト間のデータ調整を行う事で最適な再利用ができるのではないかと考えた。そこで再利用の検討を行ったが、プロジェクトの特性の相違や、予測アルゴリズムの選択等に大きく影響する事がわかった。次に新たにデータマイニング手法を適用して、最適な予測器を見出すという手法を考えた。ソフトウェア欠陥予測ツールは訓練プロジェクト、利用するメトリクス、予測のアルゴリズムを決める事で予測器を作る。与えられた検査プロジェクトに対して適切な予測器を選ぶマイニング手法を提案し、その適切性を実証した。

研究成果の概要(英文)：In error prone module prediction, we construct a predictor using some prediction algorithm and some training data, then we predict the error proneness of testing data by using this predictor, so construction of adequate predictor for each testing data is an important problem. At first we tried data adjustment based on the similarity between the training data and the testing data, but we found that other project specific characteristics affect the accuracy of prediction. In the next place, we considered knowledge based predictor mining method. In this method we construct the knowledge base which consists of many project data and their prediction accuracy data using several prediction algorithms and the project specific data. We constructed test data set and by using these data set and public data set, we showed the effectiveness of this method. We showed only the possibility of this method, so further consideration and prototype prediction tool construction are needed.

研究分野：総合領域

科研費の分科・細目：情報学 ソフトウェア

キーワード：ソフトウェア保守 メトリクス ソフトウェア欠陥予測

1. 研究開始当初の背景

ソフトウェア保守において事前に error の有無を調べる事は極めて重要である。一般には error の調査はテストプログラムの実行によって行われるが、それだけでは十分ではないので人手によるプロセスとして inspection が行われる。Inspection は labor intensive な作業であり、対象となるソフトウェアコンポーネントのすべてを対象とすることは無駄であるので、その効率化が求められる。Inspection の効率化のために error prone module を予測する研究が行われている。Error prone module とは誤りを含んでいる可能性の高いモジュールをいう。Error prone と予測されたモジュールのみを予測対象とすることで inspection の効率化を図る。これらの予測手法では

- 予測アルゴリズム
- 利用するメトリクス群
- 訓練のためのデータセット

の3つを決めたいうで、予測器を作り、予測器に対象のデータセット(検査プロジェクトのデータ)を与えて error prone module の予測を行う。予測アルゴリズムと利用するメトリクス群を決定した上で、それらのメトリクス値と error 情報を持つプロジェクトを用意する(これを訓練プロジェクトという)。そして予測アルゴリズムは訓練プロジェクトを使って予測器を作る。検査プロジェクトに対して同様にメトリクス値を求め、これを予測器に与える事で error prone module の予測を行う。

ここでメトリクスとはソフトウェアの種々の状況を定量化したもので、もっともよくつかわれるものはソースコードメトリクスである。ソースコードメトリクスはソースコードの状態を定量化したものであり、行数、演算子の数、分岐の数、オブジェクト指向の場合のメソッドの数等がある。ソースコードメトリクス以外にも変更履歴に着目した change metrics や変化の分布に着目したエントロピメトリクス等が提案されている。また予測アルゴリズムとしては決定木に基づく手法、ベイズネットに基づく手法、回帰分析に基づく手法等多くが提案されている。

与えられた予測対象に対してこれら3者(以降これらを予測手法オプションと呼ぶ)をどのように選べばよいかについては多くの研究が行われている。しかしながらどのような予測器が対象とする検査プロジェクトに対して最良であるかという点については定まった解はなく、一律に最良の予測器は無いであろうという結論になっており、検査プロジェクトに対して最良の予測器を選択するという問題が未解決である。

2. 研究の目的

上述の通り、目的は与えられた検査プロジェクトに対して最良の予測器を作る(選ぶ)手法の考案であるが、2つのアプローチが考

えられる。

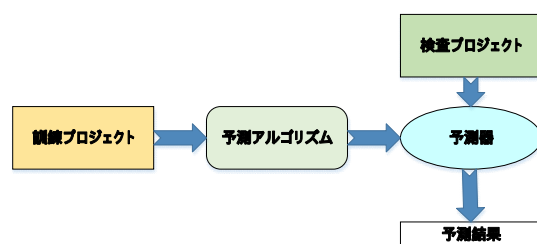
一つは多くの参照プロジェクトに対して最適な予測器をあらかじめ識別しておき、検査プロジェクトに対して最も適合する参照プロジェクトを選択し、データの調整を行ったうえで選ばれた参照プロジェクトを訓練プロジェクトとして作られた予測器を使うという手法である。一般にこのような目的のために多くの参照プロジェクトが公開されている。またそれが不十分な場合は自ら参照プロジェクトを構築する。参照プロジェクトは答としての error の有無の情報を持っているので自らを予測器に与える事で得られる予測結果の評価が可能である。この評価法を交差検定法という。これにより各参照プロジェクトに対する最良の予測器を識別しておくことができる。

他は参照プロジェクトに基づき、(訓練プロジェクト+予測器の仕様+検査プロジェクト+予測結果)の4つ組からなる知識ベースを作り、与えられた実検査プロジェクトからのマイニングにより適切な予測器を選ぶという手法である。

これら2つの手法を具体化し、種々のデータを用意したうえで、その有効性を検証し、ツール化する事を目的とした。

3. 研究の方法

予測器は訓練プロジェクトと予測アルゴリズムを選択し、利用するメトリクスを選ぶことで構築できる。一般には訓練プロジェクトデータは検査プロジェクトの過去データを利用する事が多いが、開発フェーズが異なっていた場合や特性が異なるような場合は必ずしも過去データの利用がベストとは言えない事がわかっている。

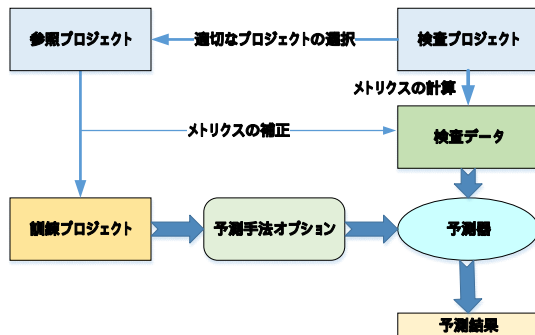


まず最初のアプローチとしては検査プロジェクトに対してメトリクス分布の類似性等から類似の訓練プロジェクトを選択し、さらに検査プロジェクトのデータとの相違を補正するという予測モデルの再利用手法を考えた。この手法のためには

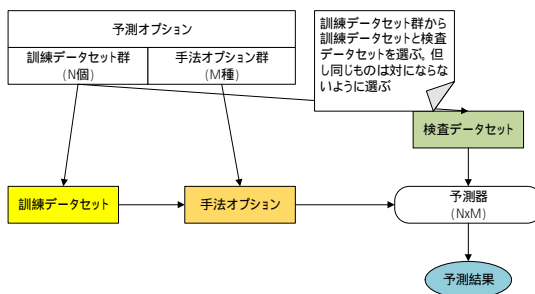
1. 多くの訓練データを集める事
2. 類似のプロジェクトの識別方法
3. データの補正方法

の3点が重要となる。1に関しては科研採択以前に収集していたデータセット及び公開されている参照データセットを利用した。2に関しては検査プロジェクトを含めてデータセットのクラスタリングを行う手法、訓練

プロジェクトと検査プロジェクトの分別の可能性から評価する手法などを試した。3に関しては訓練プロジェクトの各メトリクスの値の平均値と検査プロジェクトの各メトリクスの値の平均値を合わすように正規化するという手法などを試した。



この手法は主に訓練プロジェクトの選択に係っている。しかしながら、予測には予測アルゴリズムやメトリクスも関わり、それらを考慮した予測器の選択が必要となる。そこで以下のマイニング手法を考えた。



本マイニングのアプローチは科研プロジェクト開始後に見つけた論文 (Zhimin: An investigation on the feasibility of cross project defect prediction. Automated Software Engineering, 19(2):167-199, 2012.) にヒントを得ている。検査プロジェクトに対して一律に良好な予測器がない理由としてはプロジェクトの、メトリクスとしては考慮されない特徴が予測の精度に影響しているのではないかと考えた。そこでこのような特徴量も含めて、以下のデータからなる知識ベースを構築した

1. 訓練データのメトリクスの集約値
2. 訓練データの特徴量
3. 予測アルゴリズム
4. 訓練データのメトリクスの集約値
5. 訓練データの特徴量
6. 予測結果

これら1から5の諸量と、6の予測結果は何らかの関係を持っていると考えた。しかしどのような関係であるかは不明である。そこでデータマイニング手法を利用する。1から6のデータに基づき、検査データに対して適切な訓練データと予測アルゴリズムの対を識別する予測器マイナーを作る。この予測器マ

イナーに検査プロジェクトのデータを与える事で、最良の予測器を作るための(訓練データ+アルゴリズム)対、即ち予測器が識別できる。予測器の決定にはメトリクスの選択も大きな要因となるが、科研費期間内にはメトリクスも含めた予測器マイナーの構築法の考案はできなかった。従ってメトリクスとしては予め定めたソースコードメトリクスを固定的に利用している。

4. 研究成果

実験のための次の2つのデータセットを利用した:

- Promise データセット:このような予測実験のために公開されている参照データセットである。この中で NASA のプロジェクトに基づき計測された10種のデータセットを利用した。
- 自前で収集したプロジェクトデータ:公開されているデータセットはメトリクスが限定されている。また公開されている以上の情報を得ることもできない。そこで自前で19個の open source software のデータを収集、分析し、これを利用した。

まずは類似のプロジェクトを見つけ、予測器を再利用する手法であるが、クラスタリングによる手法、メトリクス値の補正による手法等を試したが良好な結果を得ることはできなかった。

そこで次にマイニングに基づく手法を試みる事にした。本手法は前述のように Zhimin の提案手法に基づくものではあるが、原論文ではメトリクスのみを考慮しており、プロジェクトの特徴量を考慮していなかった。そもそもメトリクスのみによる予測が必ずしもうまくいかない理由の一つに考慮されていない特徴量の存在があるのではないかと考えた。そこで特徴量を考慮して、適切な予測器を識別する手法を考案し、その有効性を2つのデータセットを利用して確認した。予測においては precision と recall という尺度を一般に利用する。この2つの尺度は予測結果を下図のような真偽で分けた場合、

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FP})$$

と定義される。

	Error を持つと予測	Error は持たないと予測
実際に error を持つ	TP	FN
Error は持たない	FP	TN

しかしこの2つの尺度は一方を良くすれば一方が下がるという性質をもつ。従って一方だけでは予測器の良否は判定できない。そこでこの2つの尺度を組み合わせて

Precision > and recall >
という条件を設定し、 として種々の値を設定して、良好な予測器の基準として知識ベースを実現した。そしてこの知識ベースの交差検定法により予測器の識別がうまくいくかどうかを確かめた。結果 $\text{precision} = 0.3$ 、 $\text{recall} = 0.7$ という厳しい条件下においても precision 値が 0.6 以上の値での識別が可能であることが判明した。これは識別できた予測器の内 6 割以上が error prone module 予測の性能として $\text{precision} > 0.3$ and $\text{recall} > 0.7$ を満たすという事を意味している。

Error prone module の予測に関しては従来種々の予測器の適否に関する研究は多く行われてきている。そして検査プロジェクト毎に適切な予測器を選択する必要があるという事は判明している。しかしながらこの「適切な予測器の選択」をどのようにして行えばよいかという研究は行われてこなかった。本研究はこの第一歩である。しかしながら識別の精度として $\text{precision} > 0.6$ という結果は決して満足のいくものではない。さらに次に様な点からの改善が必要である

- 適切な予測器を選択するために使いうる特徴量にはまだ試みていないものが多く存在する。これらを試みる事で識別の精度の向上を図る
- Error prone module の予測精度自体の向上も必要となる。 $\text{precision} > 0.3$ and $\text{recall} > 0.7$ のような条件を満たす予測器が多くあれば、その中からの選択は容易になる。現在はこのような条件を満たす識別器は全体の数%であり、それを選択するという事で識別の精度が必ずしも良いとは言えない。この割合を高める事で識別の精度も高める事が可能になる。

また予測結果の提示までを行うツールの開発までも目指していたが、それは実現できなかった。マイニング手法の改善を図ったうえでツール化も目指したい。

研究の中でこのマイニング手法は他の分野でも応用できるのではないかと考え、ソフトウェアドキュメント間の traceability link の再確立問題への適用を考察した。そしてこの内容に関しても 2 つの論文で発表を行った。

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

〔雑誌論文〕(計 1 件)

上田健之、小形真平、海谷治彦、海尻賢二 : IR 手法に基づく traceability link recovery のための method mining、電子情報通信学会論文誌、J97-D, 414-426、2014、査読有

〔学会発表〕(計 3 件)

嶋大輔、小形真平、海谷治彦、海尻賢二 : 学習に基づく error-prone モジュール予測器選択、情報処理学会第 76 回全国大会、2014.3.11、東京

内宮秀明、小形真平、海谷治彦、海尻賢二 : 最適な Error-prone module 予測器を選択する手法の提案、電子情報通信学会 KBSE 研究会、2014.1.27、東京

Takeyuki Ueda, Shinpei Ogata, Haruhiko Kaiya, Kenji Kaijiri (IR based Traceability Link Recovery Method Mining) ICSEA, 2013.10.28, Italy

〔その他〕

ホームページ等

6. 研究組織

(1) 研究代表者

海尻 賢二 (KAIJIRI, Kenji)
信州大学・工学部・教授
研究者番号 : 80115336

(2) 研究分担者

なし

(3) 連携研究者

なし