

共起関係を用いて変更要求から プログラムへの影響分析を行う手法の提案と評価

川畑徹平† 中島毅‡

芝浦工業大学† 芝浦工業大学‡

1 研究背景

大規模なソフトウェア開発では変更要求に対する影響分析が容易ではない。そして長期にわたり製品自体の派生を積み重ねる流用開発の場合その問題はさらに深刻となる。

従来、影響分析にはトレーサビリティを用いた手法が使用されてきた。トレーサビリティ[1]により要求事項の成果物間での対応漏れの確認や、保守品質を底上げすることができる。しかし、トレーサビリティを用いた影響分析の問題点としてトレーサビリティが及ぶ範囲が変更影響範囲よりも大きく直接影響範囲を指し示すものにならないということ、既存要求との関連が明確でない新しい変更要求には対しては利用できないという課題がある。

2 先行研究

先行研究として、大量の変更設計書に基づいて変更要求から影響範囲を推定するシステムがある[2]。先行システムは、過去の変更履歴を畳込みニューラルネット(CNN)により学習して変更要求文から影響範囲のコンポーネントリストを出力するシステムであり、上記課題を解決する。

しかし先行システムは、候補比率20%付近でハズレ率35%、候補内正答率23%と精度が低く、実用化のためにはさらなる精度向上が必要である(上記指標値については後述)。

3 研究目的

本研究では、先行研究の課題であるハズレ率の低減を目標とし、候補内正答率を保ったまま候補比率30%以内でハズレ率5%以下を目指す。

4 提案システム

4.1 システム構成

図1に提案システムの構成図を示す。この構成は機械学習法を除いて先行研究[2]と同じである。

4.2 学習対象となるデータ

本研究で扱うデータとして変更設計書を用い

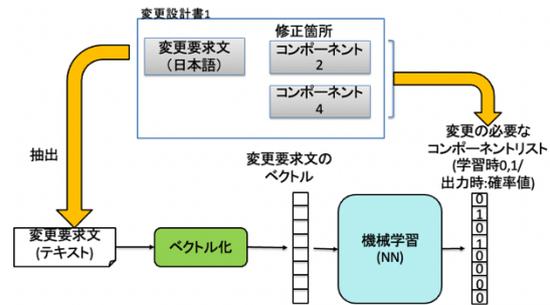


図1 提案システム構成図

る。変更設計書の構成を図2に示す。変更設計書には変更要求文ごとに変更影響範囲が記載されており、流用プログラム母体の修正する必要があるプログラムを示している。

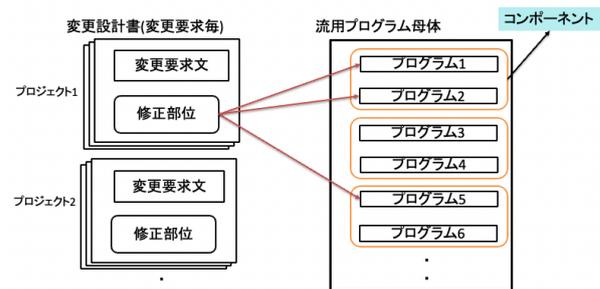


図2 変更設計書の構成

4.2 変更要求文のベクトル化

本研究では先行研究を参考に以下の手順で要求文のベクトル化を行った。

(1)形態素解析(mecab)

(2)ベクトル化

- 2.1) 単語抽出 --- 名詞のみの抽出
- 2.2) 単語のベクトル化 --- word2vec
- 2.3) ベクトル結合化 --- doc2vec

4.3 共起関係を考慮した機械学習

(1)共起関係とその着目理由

本研究では精度改善のために共起関係に注目する。共起関係とは、同時出現するラベルは他のケースでも同時に現れやすいという性質である。共起関係に着目した理由は実験に用いる流用母体プログラムのアーキテクチャ的要因により、あるモジュールが修正される時関連する特定のモジュールが修正されることが多いという仮説が成り立つのではと考えるためである。

Proposal and evaluation of a method for analyzing the impact of change requests on programs using co-occurrence relations

† Teppei Kawabata Shibaura Institute of Technology

‡ Tsuyoshi Nakajima Shibaura Institute of Technology

(2) 学習アルゴリズム

本研究は、文章ベクトルの入力集合を、修正を要する複数のモジュールに対応付ける多ラベル分類の問題に位置付けられる。多ラベル分類に出力間の共起関係を組み込むために多ラベル分類を扱うためのアルゴリズム [3] と分類器には少ないデータ数でも比較的高い精度を出すことができる Support Vector Machine (SVM) を組み合わせるアプローチを採った。

4. 4 多ラベル分類を扱うためのアルゴリズム

(1) Binary Relevance 法 (以下 BR 法) [4]

個々のラベルについて、存在するかしないかの複数の 2 値分類を作成し、各分類器の結果の和集合を出力する手法

以下は共起関係をモデル化するアルゴリズム

(2) Label PowerSet 法 (以下 LP 法) [4]

ラベルのべき集合をクラスとして扱い、マルチクラス分類問題と変換する。その後、各マルチクラス分類器をデータ内で見つかった一意のラベルの組み合わせで学習を行う手法

問題領域の性質や利用可能な学習データ数により、これらのアルゴリズムの適用可能性や精度に違いが出るものとする。本研究では、複数の手法の適用比較を行う。

5 評価実験

5. 1 評価目的

従来手法と比較して、BR 法+SVM および LP 法+SVM の精度が向上するかどうかを調査する。

5. 2 実験方法

実験は以下の手順で行った。

- (1) 変更要求文をベクトル化し特徴量を抽出
- (2) Scikit-learn を用いて BR 法、LP 法の学習モデルを作成。また分類器には SVM を使用した。
- (3) 特徴量とラベルデータを作成した学習モデルに学習させ、テストデータでラベルの予測を行い精度の検証を行う。

5. 3 実験データ

本実験では共同研究先から 405 冊の変更設計書のデータをいただいているため 323 冊を学習用の訓練データに、82 冊をテストデータとして使用した。

5. 4 実験結果の評価方法

本実験では確率値が閾値以上のコンポーネントリストの割合を候補範囲比率、候補範囲内の正答率を候補内正答率、候補範囲外に存在する修正が必要なコンポーネントの割合をハズレ率

として以上の 3 つを評価値として用いる (図 3)。以上の評価値から先行研究の候補範囲比率と候補内正答率を保ったままハズレ値を下げる事ができたか否かを評価する。

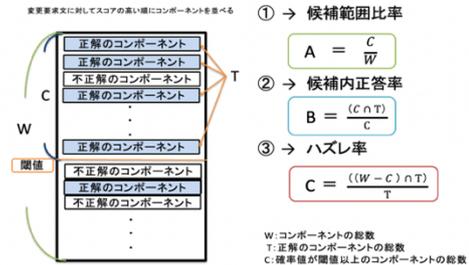


図 3 実験結果の評価方法図

5. 5 実験結果

表 1 は候補範囲比率が 30% に一番近い時の各手法のハズレ率と候補内正答率の精度の一覧を記載している表である。

表 1 評価方法に基づいた実験結果

手法	閾値	候補範囲比率	ハズレ率	候補内正答率
CNN (従来研究)	0.06	30.0%	23.0%	18.0%
BR+SVM	0.06	29.1%	17.1%	19.1%
LP+SVM	0.06	29.3%	18.2%	18.4%

従来研究では候補範囲比率 30% の時のハズレ率は 23.0% であるが BR 法+SVM の時は 17.1% と 6% ほど精度が向上する結果となった。しかし共起関係をモデル化している LP 法+SVM は BR 法+SVM と比較して精度が向上しておらず期待する結果は得られなかった。原因として、LP 法は分類するクラスが多いため学習データ数が少ないと過学習になり、精度低下を招いたと考えられる。

6 まとめと今後の課題

影響分析の精度向上のために共起関係を考慮するアルゴリズムの実装と従来研究との比較を行った。実験結果から多ラベル分類を扱うアルゴリズムと SVM を組み合わせることで従来よりも高い精度を達成できることが分かった。しかし目標値である外れ値 5% 以下には届かなかったため手法のさらなるアルゴリズム改善が必要である。

参考文献

- [1] 宇多川佳久, 他. 情報システム標準におけるトレーサビリティの事例と今後. 情報処理, 51. 2. :150-158 (2010).
- [2] 岩崎 陽也. 過去の変更仕様書を学習することにより要求文からプログラム変更部分の候補を推薦するシステムの提案, (2020).
- [3] Read, Jesse, Antti Puurula, and Albert Bifet. "Multi-label classification with meta-labels" 2014 IEEE international conference on data mining. IEEE, (2014).