トップエスイー修了制作

ソースコードの静的解析結果への Random Forest適用による不具合予測

三菱電機マイコン機器ソフトウエア株式会社　垣田賢一　kakita.kenichi@mms.co.jp

流用元ソフトウェアの品質評価に関わる課題

製品全体の品質は,流用元ソフトウェアの品質によって決まることが多い.
流用元ソフトウェアのうち,優先的に品質評価するべき箇所を特定する手法が求められる
QACなどのソースコード静的解析ツールが一般に利用されているが,優先的に品質評価するべき箇所の特定に有効とはいえない

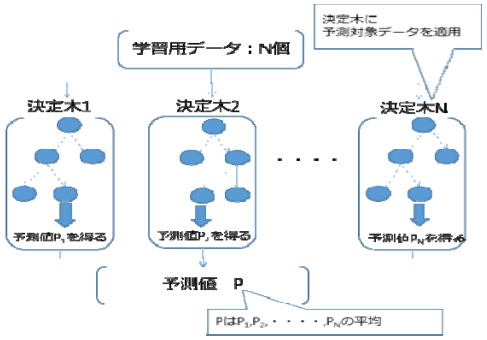
手法の適用による解決

静的解析ツールの出力に対して,機械学習の一種であるRandom Forestを適用することで流用元ソフトウェアの不具合発生箇所の特定を試みた

実施方法

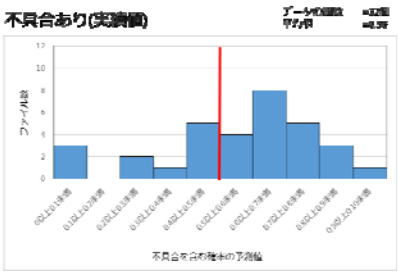
流用元ソフトウェアのメトリクスをファイル単位で取得し
不具合によって改修したか否かというデータを
組み合わせてRandom Forestを実行した

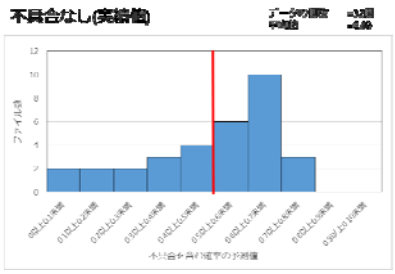
【適用上の工夫】
①不具合ありとなしのデータ不均衡
　→アンダーサンプリングの適用
②説明変数がサンプルサイズに比べて少ない
　→説明変数の絞り込み



Random Forestとは学習用データから複数の決定木を作成し
決定木を予測対象データに適用し,その結果を平均して予測値を得る手法である

結果





・不具合ありのグラフでは,予測値が0.5以上のファイルが全体の約60%近くを占めている
⇒不具合ありの予測において正しい予測は約60%であった

・不具合なしのグラフでは,予測値が0.5以下のファイルが全体の約40%を占めている
⇒不具合なしの予測において正しい予測は約40%であった

課題

構築した予測モデルによって得た
予測値を実績値と比較した

評価基準	今回のデータ	亀井らの研究	纐纈らの研究
再現率	65.6(%)	59.5(%)	62.9(%)
適合率	48.8(%)	28.2(%)	40.2(%)
F1値	56.0(%)	38.2(%)	49.1(%)

今回得た値は過去の不具合予測の研究と比較しても,決して高い精度とはいえないが,一定の水準に達していることがいえる

今後の課題として,予測値に対する不具合有無の判断基準となる境界値の決定方法がある