

論文内容の要旨

博士論文題目 Improving Automation in Bug Report Categorization and Defect Prediction

氏名 Nachai Limsettho

本論文では、ソフトウェア開発における欠陥報告書分類と欠陥予測の自動化をより容易なものとし、その適用範囲を拡大する手法を提案する。具体的には、「ノンパラメトリックな自然言語前処理」、「プロジェクト横断型予測」、「教師なし分類」の3つの観点で、自動化を容易にする手法を提案する。欠陥報告書分類と欠陥予測を自動化する手法はこれまでに数多く提案されている。ただし、自動化手法とはいえ、有意な結果を得るためには、パラメータ調整や教師データセットの準備など、その適用には多くの工数を要し、適用範囲が限定される場合も多い。

「ノンパラメトリックな自然言語前処理」では、欠陥報告書に記載されている自然言語から特徴ベクトルを抽出し、欠陥報告書分類に活用する手法を提案した。適用実験の結果、抽出された特徴ベクトルを用いることで、パラメータ調整なしに、欠陥報告書を高い精度で分類できることが分かった。「プロジェクト横断型予測」では、欠陥予測モデルの構築に用いるデータセット（教師データセット）と予測対象データセットにおける欠陥分布の違いを考慮することで、両者に大きな違いがあっても高い予測精度を実現する手法を提案した。適用実験の結果、適用対象の64%において、欠陥予測精度の向上が確認された。「教師なし分類」では、クラスタリングとクラスタへのラベル付け技術を用いることで、教師データなしでも欠陥報告書を高い精度で分類する手法を提案した。適用実験の結果、教師データがある場合と同等の分類精度が実現可能であることが確かめられた。

(論文審査結果の要旨)

本論文は、ソフトウェア開発時に作成される欠陥報告書の分類、及び、現在から将来のある時点において欠陥が含まれると思われるソフトウェアモジュールの予測(推定)において、分類や予測の精度を維持しつつ、必要となる工数を低減することを目的としている。欠陥報告書分類や欠陥モジュール予測の手法はこれまでに数多く提案されており、ツール化も行われているが、パラメータ調整や教師データセットの準備など、その適用には、依然として多くの工数を要する。本論文では、パラメータ調整を不要、もしくは、容易にする手法として「ノンパラメトリックな自然言語前処理」を、また、分類や予測の対象とするデータセットと類似性の高い教師データセットの準備を不要とする手法として「プロジェクト横断型予測」、および、「教師なし分類」を提案し、オープンソースソフトウェアを対象とした適用実験を通じてその有用性を明らかにしている。

「ノンパラメトリックな自然言語前処理」に関しては、欠陥報告書に記載されている自然言語から抽出した特徴ベクトルを用いることで、パラメータ調整を行わなくても欠陥報告書を高い精度で分類可能であることを明らかにしている。パラメータ調整は多くの工数を要するだけでなく、その妥当性の検証も必要となる。提案手法により、欠陥報告書分類はこれまで以上に容易、かつ、より広い対象に適用可能となる。「プロジェクト横断型予測」では、欠陥予測モデルの構築に用いる教師データセットと予測対象データセットにおける欠陥分布の違いを考慮することで、予想対象データセットと類似性の高い教師データセットの準備が必ずしも必要ないことを明らかにしている。この結果は、データセットの類似性は、欠陥分布に基づいて補正、あるいは、担保できることを示しており、モデル適用を容易にすると共に、その妥当性の向上にも貢献するものである。「教師なし分類」では、文字通り、教師データセットがなくても欠陥報告書を高い精度で分類可能であることを明らかにしている。

以上のとおり、本論文は、ソフトウェア開発における欠陥報告書分類、及び、欠陥モジュール予測を、その精度を維持しつつ容易にするものである。適用実験はその有用性を示しており、手法の普及を促進する。欠陥報告書分類や欠陥モジュール予測は、ソフトウェア開発におけるリスク回避技術として大変重要であり、当該領域の学術研究の発展にも大きく貢献することから、博士(工学)論文として価値あるものと認める。