氏　　名	甘　茂樺
授与した学位	博　士
専攻分野の名称	工　学
学位授与番号	博甲第　　６８４０　　号
学位授与の日付	２０２３年　３月　２４日
学位授与の要件	自然科学研究科　　産業創成工学専攻 （学位規則第４条第１項該当）
学位論文の題目	Data Generation and Evaluation Methods for Mining Software Engineering Data Sets （ソフトウェア工学データマイニングのためのデータ生成・評価法）
論文審査委員	教授　門田　暁人　　教授　太田　学　　教授　高橋　規一　　准教授　YUCEL ZEYNEP

学位論文内容の要旨

Predictive data mining consists of five steps: setting objectives, data gathering, data preparation, applying data mining algorithms, and evaluating results. This thesis proposes a data generation method and two data evaluation methods for mining software engineering data sets that support data gathering, preparation and evaluation.

The first proposal is a method for artificially generating a “mimic” software project data set, whose characteristics are very similar to a given confidential data set. Instead of using the original (confidential) data set, researchers are expected to use the mimic data set to produce similar results as the original data set. To evaluate the efficacy of the proposed method, software development effort estimation is considered as potential application domain for employing mimic data. Estimation models are built from 8 reference data sets and their concerning mimic data. The experiments confirmed that models built from mimic data sets show similar effort estimation performance as the models built from original data sets, which indicate the capability of the proposed method in generating representative samples.

The second proposal is a data quality metric called Similar Case Inconsistency Level (SCIL). Using SCIL, researchers can assess the quality of input data (training data) before conducting any data mining techniques. An empirical evaluation with 54 data samples derived from six large project data sets showed that SCIL can distinguish between consistent and inconsistent data sets, and that prediction models for software development effort and productivity built from consistent data sets can achieve relatively high accuracy.

The third proposal is a set of evaluation metrics called neg/pos-normalized accuracy measures to address the class imbalance issue in assessing performance of defect prediction models. The proposed measures enable researchers to compare defect prediction results across different data sets with different neg/pos ratios. A case study of defect prediction based on 19 defect data sets shows that the proposed measures enable us to provide a ranking of predictions across different data sets, which can distinguish between successful predictions and unsuccessful predictions.

論文審査結果の要旨

ソフトウェア開発プロジェクトを成功に導くためには、過去の開発実績データを活用し、プロジェクトの計画立案や制御に役立てることが重要となる。そのためには、多数のプロジェクトの開発実績データが必要となるが、昨今では機密保持が重要視されているために、データを集めること自体が困難となっている。また、集められたデータは、その品質が担保されている必要があるが、開発実績データの品質の評価尺度は従来ほとんど提案されていない。さらに、過去の開発実績データをソフトウェアバグ予測に用いることもよく行われるが、データ中のバグを含む個体の比率（neg/pos ratio）によって予測精度が大きくばらつき、予測の成否の評価が難しくなるという課題があった。

本論文では、これらの課題を解決するための3つの方法を提案している。まず、1つめとして、ソフトウェア開発に関する機密データの利用を促進するために、データそのものではなく、データの特徴量のみをデータ保有企業から受け取ることを想定し、特徴量の類似するデータを人工的に生成する方法を提案している。評価実験により、生成されたデータから性能の良いソフトウェア開発工数予測モデルの構築が可能であることを示している。

次に、2つめとして、ソフトウェア開発実績データの品質評価を目的として、データの矛盾性の評価尺度Similar Case Inconsistency Level (SCIL)を提案している。評価実験により、SCILの値の小さな（矛盾の少ない）データセットほど、性能の良いソフトウェア開発工数予測モデルが得られることを示している。

最後に、3つめとして、ソフトウェアバグデータのneg/pos ratioの影響を除外した評価尺度を提案している。評価実験により、提案尺度を用いることでバグ予測の成否の判断や、異なるデータセット間の予測結果の優劣をつけることが可能であることを示している。

これらの研究成果は、ソフトウェア開発実績データの利活用の障害となっていた重要な課題を解決し、また、当該領域の学術研究の発展に大きく貢献するものであり、博士（工学）の学位に値するものと認める。