ABSTRACT

Dave Harry (01679210010)

SOFTWARE DEFECT PREDICTION USING HETEROGENEOUS PARALLEL ENSEMBELS WITH SMOTE AND LASSO FOR FEATURE SELECTION AND IMBALANCE DATA(xiv + 62 pages, 8 figures; 2 tables, 1 appendices)

In the field of software development, early and accurate defect prediction is crucial for improving product quality, reducing cost, and managing testing resources effectively. However, conventional single-model prediction methods often exhibit limitations due to the complex, non-linear nature of defect data. Additionally, imbalanced class distribution in defect datasets presents another challenge, leading to overfitting towards the majority class. The Synthetic Minority Over-sampling Technique (SMOTE) addresses this problem, but an efficient feature selection method is also necessary for the high-dimensional defect data. This thesis presents an innovative approach, applying Heterogeneous Parallel Ensembles with SMOTE and LASSO for feature selection, to enhance software defect prediction.

The research adopts an empirical study design, using MDP NASA software defect dataset. Heterogeneous Parallel Ensembles are used as the primary machine learning framework, incorporating different base models operating in parallel to exploit their individual strengths. To counter the class imbalance problem, SMOTE is used to synthetically augment the minority class, thus facilitating a more balanced and effective model training. Furthermore, LASSO regression, renowned for its ability to handle high-dimensional data, is employed for feature selection. It efficiently identifies the most predictive features, thus improving model interpretability and preventing overfitting.

The experimental results demonstrate the superior performance of the proposed approach. Heterogeneous Parallel Ensembles, when combined with SMOTE and LASSO, exhibit significantly enhanced predictive accuracy for software defects. The use of SMOTE ensures the model's robustness against the imbalance in defect data, while LASSO assists in recognizing the most pertinent features contributing to the prediction.

The results of the evaluation show that our proposed model leads to an increase of up to 6% in accuracy and an increase of up to 15% in AUC, compared to when SMOTE and LASSO were not used.

Keywords: Software Defect Prediction, Heterogeneous Ensembles, Machine Learning, Software Development, Early Detection
51 Refences (1951- 2021)

ABSTRAK

Dave Harry (01679210010)

PREDIKSI CACAT PERANGKAT LUNAK MENGGUNAKAN ENSEMBEL PARALLEL HETEROGEN DENGAN SMOTE DAN LASSO UNTUK SELEKSI FITUR DAN KETIDAKSEIMBANGAN DATA
(xiv + 62 halaman; 8 gambar; 2 tabel; 1 lampiran)

Di bidang pengembangan perangkat lunak, prediksi terhadap kesalahan yang akurat dan sesegera mungkin sangat penting untuk meningkatkan kualitas produk, mengurangi biaya, dan mengelola sumber daya pengujian secara efektif. Namun, metode prediksi model tunggal konvensional sering menunjukkan keterbatasan karena sifat data cacat yang kompleks dan non-linear. Selain itu, distribusi kelas yang tidak seimbang dalam kumpulan data cacat menghadirkan tantangan lain, yang menyebabkan overfitting ke kelas mayoritas. The Synthethic Minority Over-sampling Technique (SMOTE) mengatasi masalah ini, tetapi metode pemilihan fitur yang efisien juga diperlukan untuk data cacat dimensi tinggi. Tesis ini menyajikan pendekatan inovatif, menerapkan Ensemble Paralel Heterogen dengan SMOTE dan LASSO untuk pemilihan fitur, untuk meningkatkan prediksi kerusakan perangkat lunak.

Penelitian ini mengadopsi desain studi empiris, menggunakan dataset cacat perangkat lunak MDP NASA. Ansambel Paralel Heterogen digunakan sebagai kerangka kerja pembelajaran mesin utama, menggabungkan berbagai model dasar yang beroperasi secara paralel untuk mengeksploitasi kekuatan masing-masing. Untuk mengatasi masalah ketidakseimbangan kelas, SMOTE digunakan untuk menambah kelas minoritas secara sintetis, sehingga memfasilitasi pelatihan model yang lebih seimbang dan efektif. Selain itu, regresi LASSO, yang terkenal karena kemampuannya menangani data berdimensi tinggi, digunakan untuk pemilihan fitur. Ini secara efisien mengidentifikasi fitur yang paling prediktif, sehingga meningkatkan interpretasi model dan mencegah overfitting.

Hasil percobaan menunjukkan kinerja unggul dari pendekatan yang diusulkan. Ansambel Paralel Heterogen, bila dikombinasikan dengan SMOTE dan LASSO, menunjukkan akurasi prediksi yang ditingkatkan secara signifikan untuk kerusakan perangkat lunak. Penggunaan SMOTE memastikan ketangguhan model terhadap ketidakseimbangan dalam data cacat, sementara LASSO membantu dalam mengenali fitur paling relevan yang berkontribusi pada prediksi. Hasil evaluasi menunjukkan bahwa model yang kami usulkan menghasilkan peningkatan akurasi hingga 6% dan

peningkatan AUC hingga 15%, dibandingkan dengan saat SMOTE dan LASSO tidak digunakan.

Kata Kunci: Software Defect Prediction, Heterogeneous Ensembles, Machine Learning, Software Development, Early Detection
51 referensi (1951- 2021)