INTISARI

Edited Nearest Neighbours (ENN) Dan Synthetic Minority Over-Sampling Technique-Nominal Continuous (SMOTE-NC) Pada Data Tidak Seimbang

Oleh
Nurul Fadilah
23/526313/PPA/06646

Ketidakseimbangan data merupakan masalah umum dalam *machine learning*, terutama ketika salah satu kelas dalam dataset jauh lebih dominan dibandingkan kelas lainnya. Hal ini menyebabkan model cenderung memberikan prediksi bias terhadap kelas mayoritas dan mengabaikan pola pada kelas minoritas. Kombinasi fitur numerik kontinu dan fitur kategorikal dalam dataset semakin memperumit proses penanganan ketidakseimbangan ini. Teknik Edited Nearest Neighbours (ENN) digunakan untuk mengatasi noise dalam data dengan mengeliminasi sampel yang tidak konsisten di sekitar batas antar kelas, sehingga meningkatkan kualitas data sebelum dilakukan oversampling. Sementara itu, Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) diterapkan untuk meningkatkan jumlah sampel kelas minoritas dengan mempertimbangkan atribut numerik dan kategorikal secara lebih spesifik dibandingkan metode SMOTE konvensional.

Penelitian ini mengusulkan kombinasi teknik ENN dan SMOTE-NC untuk mengatasi masalah ketidakseimbangan data dalam dataset yang memiliki fitur numerik kontinu dan kategorikal. ENN digunakan terlebih dahulu untuk menyaring data yang berisik, kemudian SMOTE-NC diterapkan untuk menambah sampel kelas minoritas yang lebih representatif. Kombinasi kedua metode ini bertujuan untuk menghasilkan data sintesis yang lebih akurat, meningkatkan kualitas data, dan meminimalkan potensi overfitting.

Pelatihan model menggunakan algoritma klasifikasi seperti *Extreme Gradient Boosting*, *Random Forest*, dan *Support Vector Machine*. Hasil penelitian menunjukkan bahwa kombinasi ENN + SMOTE-NC dari model klasifikasi XGBoost memberikan performa terbaik dengan *accuracy* sebesar 0,8088, *sensitivity* sebesar 0,5319, *specificity* sebesar 0,9096, dan *average* sebesar 0,7501. Selain itu, model SVM menunjukkan sensitivitas yang lebih tinggi pada kelas minoritas, dengan kinerja yang konsisten pada dataset yang tidak seimbang. Metode ini dapat mengatasi ketidakseimbangan data, mengurangi *noise*, dan meningkatkan kinerja model klasifikasi pada dataset dengan atribut numerik kontinu dan kategorikal.

Kata Kunci: *Data Imbalance*, ENN, SMOTE-NC, XGBoost, *Random Forest*, *Support Vector Machine.*

ABSTRACT

Edited Nearest Neighbours (ENN) and Synthetic Minority Over-Sampling Technique-Nominal Continuous (SMOTE-NC) on Imbalanced Data

By
Nurul Fadilah
23/526313/PPA/06646

Data imbalance is a common problem in machine learning, especially when one class in the dataset is much more dominant than the other. This causes the model to give biased predictions towards the majority class and ignore patterns in the minority class. Combining continuous numerical and categorical features in the dataset further complicates handling this imbalance. The Edited Nearest Neighbors (ENN) technique addresses the noise in the data by eliminating inconsistent samples around the boundaries between classes, thus improving the data quality before oversampling. Meanwhile, the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) is applied to increase the number of minority class samples by considering numerical and categorical attributes more precisely than the conventional SMOTE method.

This study proposes a combination of ENN and SMOTE-NC techniques to address the problem of data imbalance in datasets with continuous numerical and categorical features. ENN is used first to filter out noisy data, and then SMOTE-NC is applied to add more representative minority class samples. Combining these two methods aims to produce more accurate synthesis data, improve data quality, and minimize potential overfitting.

Model training uses classification algorithms such as Extreme Gradient Boosting, Random Forest, and Support Vector Machine. The results showed that the ENN + SMOTE-NC combination of the XGBoost classification model provided the best performance with an accuracy of 0.8088, sensitivity of 0.5319, specificity of 0.9096, and average of 0.7501. In addition, the SVM model showed higher sensitivity to minority classes, with consistent performance on unbalanced datasets. This method can overcome data imbalance, reduce noise, and improve classification model performance on datasets with continuous numerical and categorical attributes.

Keywords: Data Imbalance, ENN, SMOTE-NC, XGBoost, Random Forest, Support Vector Machine.