Penerapan ADASYN-Tomek Links dalam Menangani Class Imbalance Menggunakan Model Random Forest = Application of ADASYN-Tomek Links in Handling Class Imbalance Using Random Forest Model

Favian Sulthan Wafi, author

Abstrak

Class imbalance atau ketidakseimbangan jumlah kelas pada dataset merupakan permasalahan yang kerap muncul pada salah satu teknik data mining, yaitu klasifikasi. Hal ini menyebabkan kinerja dari model klasifikasi menjadi buruk karena model menjadi bias terhadap kelas mayoritas. Terdapat beberapa metode untuk menangani permasalahan ini, salah satunya dengan melakukan resampling. Resampling menyeimbangkan jumlah kelas pada dataset dengan membentuk instance (data point) minoritas baru melalui oversampling ataupun menghapus instance mayoritas melalui undersampling. Akan tetapi, pengaplikasian teknik oversampling atau undersampling dapat memunculkan permasalahan baru. Teknik oversampling berisiko mengamplifikasi noise dari kelas minoritas, sedangkan teknik undersampling berisiko menghilangkan informasi penting dari kelas mayoritas. Untuk mengatasi kekurangan satu sama lain, kedua teknik ini dapat digabungkan menjadi hybrid sampling. Penelitian ini akan menggunakan teknik hybrid sampling ADASYN-Tomek Links, yaitu penggabungan antara teknik oversampling ADASYN dan undersampling Tomek Links. ADASYN, sama seperti SMOTE, membentuk instance sintetis minoritas baru di sekitar instance minoritas yang ada, tetapi memfokuskan pembentukannya di daerah kelas minoritas yang lebih sulit untuk dipelajari model. Di sisi lain, Tomek Links menghapus pasangan instance, disebut dengan pasangan Tomek Link, yang dianggap sebagai noise. Dengan begitu, noise yang dihasilkan dari oversampling ADASYN dapat dikurangi melalui Tomek Links. ADASYN-Tomek Links akan diaplikasikan pada 2 dataset, dengan total instance dan derajat class imbalance berbeda, menggunakan model klasifikasi random forest dengan teknik optimalisasi hyperparameter random search. Hasilnya, teknik ADASYN-Tomek Links memberikan performa terbaik dalam memprediksi kedua kelas dengan nilai balanced accuracy tertinggi pada kedua dataset, melebihi teknik ADASYN dan SMOTE.
.....Class imbalance, or the imbalance in the number of classes in a dataset, is a problem that often arises in one of the data mining techniques, namely classification. This causes the performance of the classification model to be poor because the model becomes biased towards the majority class. There are several methods to handle this problem, one of which is by performing resampling. Resampling balances the number of classes in the dataset by forming new minority instances (data points) through oversampling or deleting majority instances through undersampling. However, the application of oversampling or undersampling techniques can give rise to new problems. The oversampling technique risks amplifying noise from the minority class. On the other hand, the undersampling technique risks eliminating important information from the majority class. Therefore, to overcome each other’s shortcomings, these two techniques can be combined, which is called hybrid sampling. This research will use the ADASYN-Tomek Links hybrid sampling technique, which is a combination of the ADASYN oversampling technique and the Tomek Links undersampling technique. ADASYN, similar to SMOTE, generates new synthetic minority instances around the existing minority instances, but focuses their formation in areas of the minority class that are more

difficult for the model to learn. On the other hand, Tomek Links removes pairs of instances, called Tomek Link pairs, which are considered as noise. Thus, noise that might be generated from ADASYN oversampling can be reduced through Tomek Links. ADASYN-Tomek Links will be applied to 2 datasets , with different total instances and degrees of class imbalance, using a random forest classification model with random search hyperparameter optimization. As a result, the ADASYN-Tomek Links technique provides the best performance in predicting both classes with the highest balanced accuracy value on both datasets, surpassing the ADASYN and SMOTE techniques.