Universitas Indonesia Library >> UI - Skripsi Membership

Analisis Perbandingan Metode SMOTE, SMOTE-ENN, dan SMOTE-Tomek Link Dalam Menangani Imbalanced Data pada Klasifikasi = Comparative Analysis of SMOTE, SMOTE-ENN, and SMOTE-Tomek Link Methods in Handling Imbalanced Data in Classification

Valery Ongso Putri, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920542848&lokasi=lokal

---

Abstrak

Ketidakseimbangan data merupakan masalah umum yang terjadi dalam bidang analisis data. Data menjadi tidak seimbang karena terdapat perbedaan antara jumlah sampel pada setiap kelasnya. Masalah ketidakseimbangan ini menyebabkan model klasifikasi menjadi bias, dimana model akan cenderung memprediksi kelas mayoritas secara efektif dibandingkan dengan kelas minoritas dan dapat menyebabkan kesalahan interpretasi dalam pengambilan suatu keputusan. Terdapat beberapa cara dalam menangani data yang tidak seimbang, yaitu random undersampling dan random oversampling. Salah satu metode dari random oversampling yang populer adalah Synthetic Minority Oversampling Technique (SMOTE). SMOTE dapat digabungkan dengan metode random undersampling, yaitu Edited Nearest Neighbors (ENN) dan Tomek link. Pada metode gabungan SMOTE-ENN dan SMOTE-Tomek link, SMOTE bekerja terlebih dahulu dengan membuat sampel sintetis pada kelas minoritas. ENN dan Tomek link berperan sebagai cleaning untuk menghapus data yang tidak relevan dan dianggap sebagai noise. Untuk melihat pengaruh ketiga metode resampling tersebut, yaitu SMOTE, SMOTEENN, dan SMOTE-Tomek Link, dilakukan simulasi data. Simulasi data dapat melihat pengaruh ukuran sampel, ukuran proporsi kelas, dan metode resampling terhadap model klasifikasi decision tree, random forest, dan XGBoost pada data yang tidak seimbang. Simulasi data juga dijalankan sebanyak 100 iterasi yang menunjukkan bahwa iterasi pertama cukup untuk mewakili hasil dari 100 iterasi. Hasil menunjukkan bahwa ketiga metode cenderung mampu memberikan hasil yang baik dengan adanya peningkatan nilai metrik precision, recall, ROC-AUC, dan G-Mean. Metode SMOTE dengan XGBoost bekerja dengan baik pada ukuran sampel kecil dengan adanya peningkatan nilai metrik yang cukup signifikan. Pada SMOTE-ENN, nilai recall cenderung meningkat yang diikuti oleh menurunnya nilai precision pada proporsi 1:9, 2:8, dan 3:7 dengan sampel yang relatif kecil. SMOTE-Tomek Link juga meningkatkan nilai metrik pada sampel yang relatif kecil dengan proporsi memberikan nilai metrik tertinggi.

.....Data imbalance is a common problem that occurs in the field of data analysis. The data becomes unbalanced because there is a difference between the number of samples in each class. This imbalance problem causes the classification model to be biased, where the model will tend to predict the majority class effectively compared to the minority class and can cause misinterpretation in making a decision. There are several ways to handle imbalanced data, namely random undersampling and random oversampling. One of the popular random oversampling methods is Synthetic Minority Over-sampling Technique (SMOTE). SMOTE can be combined with random undersampling methods, namely Edited Nearest Neighbors (ENN) and Tomek link. In the combined SMOTE-ENN and SMOTE-Tomek link method, SMOTE works first by creating a synthetic sample in the minority class. ENN and Tomek link act as cleaning to remove irrelevant data and are considered as noise. To see the effect of the three resampling methods, namely SMOTE, SMOTE-ENN, and SMOTE-Tomek Link, data simulation was conducted. Data simulation can see the effect

of sample size, class proportion size, and resampling method on decision tree, random forest, and XGBoost classification models on imbalanced data. The data simulation was also run for 100 iterations which shows that the first iteration is sufficient to represent the results of 100 iterations. The results show that the three methods tend to be able to provide good results with an increase in the precision, recall, ROC-AUC, and G-Mean metric values. The SMOTE method with XGBoost works well on small sample sizes with a significant increase in metric values. In SMOTE-ENN, the recall value tends to increase followed by a decrease in precision value at proportions 1:9, 2:8, and 3:7 with relatively small samples. SMOTE-Tomek Link also increases the metric value on relatively small samples with proportions of 1:9 and 2:8. In addition, the resampling method was also used on data available on Kaggle.com, namely Pima Indian Diabetes and Give Me Some Credit:: 2011 Competition. In the Pima Indian Diabetes data, it can be seen that the recall, ROC-AUC, and G-Mean values are the highest using SMOTE-ENN with the XGBoost model. On the Give Me Some Credit:: 2011 Competition also shows that the SMOTE-ENN method with the XGBoost model provides the highest metric value.