

Komparasi Metode SMOTE, SMOTE-ENN, dan SMOTE-CUT dalam Menangani Imbalanced Data pada Klasifikasi Multi-Kelas dengan Support Vector Machine (SVM) = Comparative Analysis of SMOTE, SMOTE-ENN, and SMOTE-CUT in Multi-Class SVM Classification for Imbalanced Data

Kesia Gabriele, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920551862&lokasi=lokal>

Abstrak

Support Vector Machine (SVM) merupakan model klasifikasi yang dikenal dengan keakuratan klasifikasi yang tinggi. Namun, Support Vector Machine (SVM) menghasilkan hasil klasifikasi yang kurang optimal jika data yang digunakan tidak seimbang (imbalanced data). Terdapat beberapa cara dalam menangani data yang tidak seimbang, salah satunya dengan metode resampling. Metode resampling sendiri terbagi dalam dua pendekatan yaitu over-sampling dan under-sampling. Salah satu pendekatan over-sampling yang populer adalah Synthetic Minority Over-sampling Technique (SMOTE). SMOTE bekerja dengan membangkitkan sampel sintesis pada kelas minoritas. Untuk meningkatkan kinerja model, SMOTE dapat digabungkan dengan pendekatan under-sampling seperti Edited Nearest Neighbors (ENN) dan Cluster-based Undersampling Technique (CUT). Dalam kombinasinya dengan SMOTE, ENN berperan sebagai cleaning untuk menghapus data sintesis dari penerapan SMOTE yang tidak relevan dan dianggap sebagai noise. Sementara, CUT berperan dalam mengidentifikasi sub-kelas dari kelas mayoritas untuk menekan angka over-sampling sekaligus meminimalisir hilangnya informasi penting pada kelas mayoritas selama proses undersampling. Kombinasi over-sampling dan under-sampling ini saling melengkapi dan mengatasi kekurangan dari masing-masing metode. Penelitian ini memfokuskan perbandingan performa metode resampling SMOTE beserta variasinya, yaitu SMOTEENN dan SMOTE-CUT dalam mengklasifikasikan data multi-kelas yang tidak seimbang menggunakan Support Vector Machine. Dari analisis yang dilakukan, diperoleh kesimpulan bahwa SMOTE-CUT cenderung menghasilkan performa klasifikasi yang lebih baik dibandingkan dengan SMOTE ataupun SMOTE-ENN. Walaupun demikian, keseluruhan metode resampling (SMOTE, SMOTE-ENN, dan SMOTE-CUT) mampu meningkatkan kinerja dari model klasifikasi Support Vector Machine (SVM).

.....Support Vector Machine (SVM) is popular classifier that is known for its high accuracy value. However, Support Vector Machine (SVM) may not perform well on imbalanced datasets. There are several ways to handle imbalanced data, one of them is through resampling methods. Resampling methods itself divided into two approaches, oversampling and under-sampling. One of the popular over-sampling methods is Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by generating synthetic samples for the minority class. SMOTE can be combined with under-sampling methods such as Edited Nearest Neighbors (ENN) or Cluster-based Under-sampling Technique (CUT). In combination with SMOTE, ENN acts as a cleaning role to remove synthetic data generated from SMOTE application that is not relevant and considered as noise. Meanwhile, CUT plays a role in identifying sub-class from the majority class to reduce over-sampling while minimizing the loss of important information in the majority class during the under-sampling process. The combination of over-sampling and undersampling is needed to complement and overcome the weakness of each method. This research mainly focuses on comparing the performance of the

resampling method SMOTE and its variations, SMOTE-ENN and SMOTE-CUT, in classifying multi-class imbalanced data using Support Vector Machine. From the analysis conducted, it was concluded that data with resampling SMOTE-CUT shows better classification performance compare to data with resampling SMOTE or SMOTE-ENN. However, any resampling method (SMOTE, SMOTE-ENN, and SMOTE-CUT) can handle imbalanced data and improve Support Vector Machine performance.