INTISARI

APLIKASI SMOTE+ENN PADA ANALISIS KLASIFIKASI DENGAN DATA TIDAK SEIMBANG UNTUK PENGKLASIFIKASIAN KANDIDAT PULSAR

Oleh

Wolfgang May Panca Angga Adi Purna
17/409530/PA/17837

Analisis klasifikasi merupakan suatu teknik untuk memprediksi label kelas dari data obeservasi dengan menggunakan model klasifikasi yang terbentuk. Analisis klasifikasi yang paling sederhana adalah klasifikasi biner. Dalam dunia nyata, pada analisis klasifikasi biner seringkali ditemukan kasus *imbalanced data. Imbalanced data*, dalam klasifikasi biner, merupakan kondisi pada suatu dataset dengan proporsi data antara kelas positif dan kelas negatif yang tidak seimbang, di mana jumlah sampel pada kelas positif (*main class of interest*) jauh kurang dari kelas negatif. Kondisi tersebut dapat menyebabkan penurunan performa klasifikasi. Kondisi *imbalanced data* tidak semata-mata menjadi satu-satunya faktor permasalahan yang menyebabkan penurunan performa pengklasifikasi, meskipun hal tersebut terjadi pada data yang tidak seimbang. Terdapat faktor lain yang juga dapat menurunkan performa dari algoritma pembelajar, salah satunya adalah tingkat data tumpang tindih antar kelas atau *class overlapping*. Maka dari itu, untuk menangani kondisi permasalahan tersebut pada penelitian ini digunakan metode *Synthetic Minority Oversampling Technique + Edited Nearest Neighbor* (SMOTE+ENN).

Metode SMOTE+ENN bekerja dengan menerapkan algoritma SMOTE terlebih dahulu dan dilanjutkan dengan penerapan algoritma ENN. SMOTE membangkitkan sampel sintetis sebagai sampel pengamatan tambahan pada kelas minoritas. Kemudian, ENN melakukan modifikasi dataset dengan cara pembersihan dataset dari sampel pengamatan yang memiliki label kelas berbeda dari mayoritas label kelas *K* sampel tetangga terdekatnya. Studi kasus pada penelitian ini menggunakan dataset *HTRU2 Data Set*. Dalam penelitian ini, diperoleh hasil bahwa performa klasifikasi, yang diukur dengan indikator *g-mean*, dari analisis klasifikasi dengan menerapkan SMOTE+ENN lebih baik daripada analisis klasifikasi tanpa penerapan teknik resampling dan dengan penerapan SMOTE.
Kata kunci: klasifikasi, data tidak seimbang, *class overlapping*, SMOTE, ENN, *g-mean*.

ABSTRACT

APPLICATION OF SMOTE+ENN IN IMBALANCED CLASSIFICATION ANALYSIS FOR PULSAR CANDIDATES CLASSIFICATION

By

Wolfgang May Panca Angga Adi Purna
17/409530/PA/17837

Classification analysis is a technique used for predicting class labels of instances using a constructed classification model. The simplest classification analysis is binary classification. In real, imbalanced data is usually happening in binary classification. In binary classification, imbalanced data is an imbalanced proportion between the positive class and the negative class, where the number of samples in the positive class is too few than in the negative class. That condition can decrease the classification performance. Imbalanced data is not solely being the factor that can decrease the classification performance, even it happens in an imbalanced dataset. Other factors can decrease the classification performance, such as class overlapping. Therefore, Synthetic Minority Oversampling Technique + Edited Nearest Neighbor (SMOTE+ENN) is used in this research.

SMOTE+ENN works by doing the SMOTE algorithm first and then continuing with the ENN algorithm. SMOTE creates synthetic instances as adding instances for the minority class. Then, ENN modifies the dataset by delete instances that have class labels different from the majority class labels of its *K* nearest neighbors. The case study in this research uses *HTRU2 Data Set*. The result in this research is classification analysis with application SMOTE+ENN has better classification performance, based on *g-mean*, than classification analysis without application resampling technique and with application SMOTE.
Keywords: classification, imbalanced data, class overlapping, SMOTE, ENN, *g-mean*