INTISARI

IMPLEMENTASI BOBOT KELAS DALAM *RANDOM FOREST* UNTUK PENANGANAN DATA TIDAK SEIMBANG

Zulfikar Hanif
20/459378/PA/20039

Klasifikasi adalah bentuk analisis data yang bertujuan mengekstrak model untuk menggambarkan kelas-kelas kategorik dari data. Untuk meminimalkan tingkat eror dari klasifikasi yang dilakukan, data yang digunakan dalam klasifikasi tidak diperkenankan dalam keadaan tidak seimbang. Sebuah dataset dikatakan tidak seimbang apabila kelas-kelasnya tidak terwakili secara merata. Oleh karena itu, diperlukan adanya penanganan terhadap data tidak seimbang sebelum dilakukan analisis klasifikasi. Penelitian ini akan menerapkan penanganan ketidakseimbangan data dalam klasifikasi *Random Forest* untuk memprediksi apakah suatu tembakan dalam pertandingan sepak bola dapat berujung pada gol. Dataset yang digunakan adalah *Football Events* dari situs Kaggle. Dataset tersebut memperlihatkan bahwa hanya sekitar 10% tembakan yang dapat menghasilkan gol. Angka ini sesuai dengan probabilitas gol pada sepak bola profesional serta mengindikasikan ketidakseimbangan data. Pada penelitian ini, penanganan yang dipakai adalah SMOTE-N dan bobot kelas. Penentuan bobot kelas menggunakan *grid search* serta parameter *balanced*. Klasifikasi dengan bobot kelas menghasilkan *accuracy* sebesar 72,2% untuk bobot hasil *grid search* dan 72,3% untuk bobot *balanced*, *sensitivity* sebesar 76,5% untuk bobot hasil *grid search* dan 76,4% untuk bobot *balanced*, serta *specificity* dan *g-mean* yang identik, yakni *specificity* sebesar 71,7% dan *g-mean* sebesar 74,1%. Nilai *g-mean* tersebut serupa dengan nilai yang dihasilkan dengan penanganan SMOTE-N serta meningkat dari nilai yang dihasilkan oleh klasifikasi tanpa penanganan sebesar 51,2%.

Kata kunci: klasifikasi, data tidak seimbang, *Random Forest*, SMOTE-N, bobot kelas.

ABSTRACT

IMPLEMENTATION OF CLASS WEIGHT IN RANDOM FOREST FOR HANDLING IMBALANCED DATA

Zulfikar Hanif
20/459378/PA/20039

Classification is a form of data analysis that aims to extract models that describe categorical classes from data. To minimize the error rate of the classification, the data used in classification should not be imbalanced. A dataset is said to be imbalanced if the classes are not equally represented. Therefore, it is necessary to handle imbalanced data before classification analysis is carried out. This research will apply data imbalance handling in Random Forest classification to predict whether a shot in a football match can lead to a goal. The dataset used is Football Events obtained from the Kaggle website. The dataset shows that only about 10% of shots can result in goals. This value corresponds to the probability of goals in professional football and indicates an imbalance in the data. In this research, the treatments that will be used are SMOTE-N and class weights. The determination of class weights uses grid search and balanced parameters. Class-weighted classification resulted in accuracy of 72,2% for grid search results' weights and 72,3% for balanced weights, sensitivity of 76,5% for grid search results' weights and 76,4% for balanced weights, and identical specificity and g-mean, with a specificity of 0,717 and a g-mean of 0,741. The g-mean value is similar to the value generated with SMOTE-N treatment and improved from the value generated by classification without treatment of 51.2%.

Keywords: classification, imbalance data, Random Forest, SMOTE-N, class weight.