INTISARI

PENERAPAN METODE *NEAR MISS UNDERSAMPLING* DAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) PADA ANALISIS KLASIFIKASI DATA TIDAK SEIMBANG

oleh
Evi Iravianti
19/445716/PA/19540

Perkembangan teknologi yang amat pesat sangat berpengaruh terhadap ketersediaan data di pasaran, di mana sebagian data yang tersedia jumlahnya cukup besar dan kompleks. Data dengan jumlah yang besar dan kompleks akan meningkatkan peluang terjadinya kesalahan ketika data tersebut digunakan untuk suatu analisis. Permasalahan yang sering ditemui ketika melakukan sebuah analisis klasifikasi adalah adanya data tidak seimbang, yakni proporsi jumlah data pada setiap kelasnya cukup berbeda. Hal ini dapat mengakibatkan terjadinya bias, yakni model klasifikasi yang dihasilkan cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Beberapa metode yang dapat digunakan untuk mengatasi masalah data tidak seimbang pada analisis klasifikasi aadalah *undersampling* dan *oversampling. Near Miss* merupakan salah satu metode *undersampling* yang dapat digunakan untuk menangani masalah data tidak seimbang dengan cara memilih data berdasarkan jarak data kelas mayoritas terhadap kelas minoritas. Metode ini akan mereduksi data pada kelas mayoritas untuk mendapatkan proporsi kelas yang lebih seimbang. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan salah satu metode *oversampling* yang menyeimbangkan data dengan cara membentuk *instance sintetic* untuk kelas minoritas. Pada tugas akhir ini dilakukan analisis pada data *Heart Failure* dan data *Indian Liver Patient Records* dengan menggunakan metode klasifikasi *Random Forest* yang dikombinasikan dengan motode *Near Miss Undersampling* serta metode SMOTE. Dari analisis yang telah dilakukan, didapatkan kesimpulan

bahwa penerapan metode *Near Miss Undersampling* menghasilkan performa yang lebih baik dari pada metode SMOTE untuk analisis klasifikasi *Random Forest.*

Kata kunci: data tidak seimbang, klasifikasi, *Near Miss*, SMOTE, *Random Forest*

ABSTRACT

IMPLEMENTATION OF NEAR MISS UNDERSAMPLING METHOD AND SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) FOR IMBALACED DATASET CLASSIFICATION

by

Evi Iravianti

19/445716/PA/19540

The rapid development of technology greatly affects the availability of data on the market, where some of the available data is quite large and complex. Large and complex amounts of data will increase the chances of errors occurring when the data is used for an analysis. The problem that is often encountered when carrying out a classification analysis is the presence of unbalanced data, that is the proportion of the amount of data in each class is quite different. This can lead to bias, in which the resulting classification model tends to predict the majority class and ignore the minority class. There are several methods that can be used to overcome the problem of unbalanced data in classification analysis, namely undersampling and oversampling. Near Miss is one of the undersampling methods that can be used to deal with unbalanced data problems by selecting data based on the distance from the majority class to the minority class. This method will reduce the data in the majority class to get a more balanced class proportion. Synthetic Minority Over Sampling (SMOTE) is an oversampling methods that balances data by creating a synthetic instance for the minority class. In this final project, an analysis of Heart Failure dataset and Indian Liver Patient Records dataset which is carried out using the Random Forest classification method combined with the Near Miss Undersampling method and the SMOTE method. From the analysis that has been done, concluded that the application of the Near Miss Undersampling method produces better performance than the SMOTE method for the analysis of Random Forest classification.

Keywords: unbalanced data, classification, Near Miss, SMOTE, Random Forest