Metode easy ensemble dengan random forest untuk mengatasi masalah klasifikasi pada kelas data tidak seimbang = Easy ensemble with random forest to handle imbalanced data in classification

Gregorius Vidy Prasetyo, author

Abstrak

ABSTRAK
Pada permasalahan seperti kesehatan atau dunia retail banyak dijumpai data-data yang memiliki kategori yang tidak seimbang. Sebagai contoh jumlah penderita penyakit tertentu relatif langka pada suatu studi atau jumlah transaksi yang terkadang merupakan transaksi palsu (fraud) jumlahnya secara signifikan lebih sedikit ketimbang transaksi normal. Kondisi ini biasa disebut sebagai kondisi data tidak seimbang dan menyebabkan permasalahan pada performa model, terutama pada kelas minoritas. Beberapa metode telah dikembangkan untuk mengatasi permasalahan data tidak seimbang, salah satu metode terkini untuk menanganinya adalah Easy Ensemble. Easy Ensemble diklaim dapat mengatasi efek negatif dari pendekatan konvensional seperti random-under sampling dan mampu meningkatkan performa model dalam memprediksi kelas minoritas. Skripsi ini membahas metode Easy Ensemble dan penerapannya dengan model Random Forest dalam mengatasi masalah data tidak seimbang. Dua buah studi empiris dilakukan berdasarkan kasus nyata dari situs kompetisi hacks.id dan kaggle.com. Proporsi kategori antara kelas mayoritas dan minoritas pada dua data di kasus ini adalah 70:30 dan 94:6. Hasil penelitian menunjukkan bahwa metode Easy Ensemble, dapat meningkatkan performa model klasifikasi Random Forest terhadap kelas minoritas dengan signifikan. Sebelum dilakukan resampling pada data (nhacks.id), nilairecall minority hanya sebesar 0.47, sedangkan setelah dilakukan resampling, nilainya naik menjadi 0.82. Begitu pula pada data kedua (kaggle.com), sebelum resampling nilai recall minority hanya sebesar 0.14, sedangkan setelah dilakukan resampling, nilai naik secara signifikan menjadi 0.71.
<hr>
ABSTRACT
In the real world problem, there is a lot case of imbalanced data. As an example in medical case, total patients who suffering from cancer is much less than healthy patients. These condition might cause some issues in problem definition level, algorithm level, and data level. Some of the methods have been developed to overcome this issues, one of state-of-the-art method is Easy Ensemble. Easy Ensemble was claimed can improve model performance to classify minority class moreover can overcome the deï¬?ciency of random under-sampling. In this thesis discussed the implementation of Easy Ensemble with Random Forest Classifers to handle imbalance problem in a credit scoring case. This combination method is implemented in two datasets which taken from data science competition website, nhacks.id and kaggle.com with class proportion within majority and minority is 70:30 and 94:6. The results show that resampling with Easy Ensemble can improve Random Forest classifier performance upon minority class. This been shown by value of recall on minority before and after resampling which increasing significantly. Before resampling on the first dataset (nhacks.id), value of recall on minority is just 0.49, but then after resampling, the value of recall on minority is increasing to 0.82. Same with the second dataset (kaggle.com), before the resampling, value of recall on minority is just 0.14, but then after resampling, the value of recall on minority

is increasing significantly to 0.71.