INTISARI

Implementasi CatBoost pada Data Multikelas Tidak Seimbang (Studi Kasus: Klasifikasi Tutupan Hutan)

Oleh

Febyanita Sari
19/439211/PA/19034

Pada analisis klasifikasi, keakuratan model dapat dipengaruhi oleh keseimbangan banyak sampel dalam tiap kelas. Data dengan karakteristik ini banyak ditemukan karena secara natural sulit untuk mendapat data yang seimbang di dunia nyata. Ketidakseimbangan pada data multikelas tentu lebih mungkin terjadi dibandingkan data biner karena kepemilikan kelas yang lebih banyak. Data tidak seimbang dapat ditangani melalui pendekatan tingkat data (*resampling*) dan pendekatan tingkat algoritma. Beberapa penelitian empiris telah menyebutkan bahwa *ensemble learning* dapat meningkatkan akurasi pada klasifikasi data tidak seimbang. Seiring perkembangan *ensemble learning*, muncul beberapa inovasi yang populer digunakan karena menunjukkan performa yang baik dalam kasus klasifikasi, salah satunya adalah CatBoost. Tugas Akhir ini membahas komparasi penanganan data multikelas tidak seimbang dengan teknik *resampling* dan teknik *ensemble learning*. Perbandingan dilakukan secara empiris dengan data Covertype dari UCI Machine Learning Repository. Teknik Undersampling dipilih karena data Covertype merupakan data yang cukup besar sehingga tidak perlu replikasi data seperti pada *oversampling*. Teknik algoritma yang dipilih adalah CatBoost karena merupakan teknik *ensemble learning* khususnya *boosting* yang secara empiris terbukti unggul dalam beberapa penelitian. Model dasar pada penelitian ini adalah Decision Tree tanpa penanganan apapun. Kesimpulan penelitian ini adalah CatBoost mengungguli model Decision Tree, Decision Tree dengan penanganan Undersampling, dan Gradient Boosting. CatBoost menghasilkan nilai akurasi 82%, F1-Score Macro 72%, dan F1-Score Weighted 82%. Selain itu, diperoleh variabel terpenting dalam model yaitu elevasi.

Kata kunci : CatBoost, Ketidakseimbangan Data, Multikelas, Undersampling.

ABSTRACT

CatBoost Implementation on Imbalanced Multiclass Data (Case Study: Forest Cover Classification)

By

Febyanita Sari
19/439211/PA/19034

The balance of samples in each class can have an impact on a classification model's accuracy. Naturally, it is difficult to get balanced data in the real world. Due to the ownership of additional classes, imbalanced class in multiclass settings arise more frequently than in binary settings. Imbalanced class can be handled by data level method and algorithm level method. Many empirical research has stated that ensemble learning can improve accuracy in imbalanced data classification. Along with the development of ensemble learning, several innovations emerged and popularly used because they show great performance in classification cases, one of which is CatBoost. This project will compare data level method and algorithm level method in handling imbalanced multiclass data. Comparison will be carried out empirically using Covertype Data from UCI Machine Learning Repository. Undersampling is chosen as the data level method because the Covertype data is large enough thus there is no need for data replication as in oversampling. CatBoost is chosen as the algorithm level method because it has been empirically proven to be superior in several studies. The base model used in this study is Decision Tree. The study's findings show that CatBoost works better than the Gradient Boosting, Decision Tree, and Decision Tree with Undersampling. CatBoost yield an accuracy score of 82%, F1-Score Macro of 72%, and F1-Score Weighted of 82%. In addition, the most important variable in the model is elevation.

Keywords : CatBoost, Imbalanced Class, Multiclass, Undersampling.