INTISARI

GRAPH AUTO ENCODER UNTUK DETEKSI OUTLIER PADA MODEL *XGBOOST REGRESSION*

Oleh

Saifudin Rosyid
22/509297/PPA/06455

Deteksi outlier berperan dalam meningkatkan kualitas data untuk memperoleh kinerja model *machine learning* yang optimal. Suatu metode deteksi tertentu tidak dapat diandalkan untuk semua karakteristik dataset. Pada dataset multivariat, deteksi outlier tidak hanya menilai distribusi data variabel secara individu, namun juga mempertimbangkan keterkaitan antar variabel dalam dataset. Batas yang samar antar outlier dengan objek normal juga menyulitkan proses deteksi outlier. *Isolation Forest* dalam hal ini efektif mendeteksi outlier yang mudah dipartisi dengan data normal, namun mempunyai kelemahan menangani outlier yang hanya dapat diisolasi dalam subruang orde tinggi yang mempertimbangkan interaksi antar fitur dataset.

Metode *Graph Auto Encoder* (GAE) memberikan solusi dengan menerapkan *feature value propagation* untuk menangani outlier sekalipun berdekatan dengan area data normal. Metode berbasis graf menganalisis keterhubungan antar data dengan melibatkan semua variabel dataset. Pendekatan ini mengoptimalkan deteksi outlier pada dataset multivariat sehingga meningkatkan kualitas data untuk menghasilkan performa optimal model XGBoost Regression. Kinerja model yang optimal dinilai berdasarkan metrik *mean absolute error* (MAE), *R-squared,* dan *root mean squared error* (RMSE).

Penelitian ini menggunakan 7 dataset sebagai objek analisis. Model XGB-GAE melampaui performa penelitian terdahulu pada dataset *Boston Housing, Concrete Compressive Strenght, Airfoil Self Noise,* dan *Superconductivity Data*. Sedangkan pada dataset *AQI Central Pollution Control Board*, *AQI Open Goverment Data,* model tidak outperform disebabkan oleh banyaknya jumlah data yang hilang dari dataset, sementara pada dataset *QSAR Fish Toxicity,* model gagal mencapai konvergensi selama proses pelatihan maupun pengujian. Implementasi GAE berkontribusi positif dalam mencapai kinerja optimal model *XGBoost Regression*. Namun, operasi graf membutuhkan komputasi yang tinggi dan tingkat akurasi deteksi outlier GAE dipengaruhi oleh pemilihan parameter *k-nearest neighbor* yang tepat.

Kata kunci: *Outlier Detection, Isolation Forest, Graph Auto Encoder, Machine Learning, Extreme Gradient Boosting*

ABSTRACT

GRAPH AUTO ENCODER OUTLIER DETECTION IN THE XGBOOST REGRESSION MODEL

By

Saifudin Rosyid
22/509297/PPA/06455

Outlier detection plays a crucial role in enhancing data quality to achieve optimal machine learning model performance. A specific detection method may not be reliable for all dataset characteristics. In multivariate datasets, outlier detection not only assesses the distribution of individual variables but also considers the interrelationships between variables within the dataset. The ambiguous boundary between outliers and normal objects further complicates the outlier detection process. Isolation Forest is effective in detecting outliers that can be easily partitioned from normal data but struggles with outliers that can only be isolated in high-order subspaces considering the interactions between dataset features.

The Graph Auto Encoder (GAE) method provides a solution by applying feature value propagation to handle outliers, even those that are close to the normal data area. Graph-based methods analyze the interconnectedness between data by involving all dataset variables. This approach optimizes outlier detection in multivariate datasets, thereby improving data quality to yield optimal performance of the XGBoost Regression model. Optimal model performance is evaluated based on the mean absolute error (MAE), R-squared, and root mean squared error (RMSE) metrics.

This research utilizes seven datasets as the analysis objects. The XGB-GAE model outperformed previous research on the Boston Housing, Concrete Compressive Strength, Airfoil Self Noise, and Superconductivity Data datasets. However, the model did not outperform on the AQI Central Pollution Control Board and AQI Open Government Data datasets due to the significant amount of missing data, while on the QSAR Fish Toxicity dataset, the model failed to converge during both training and testing processes. While GAE implementation contributes positively to achieving optimal performance of the XGBoost Regression model, its graph operations demand high computational resources. Additionally, the accuracy of outlier detection in GAE is sensitive to the appropriate selection of the k-nearest neighbor parameter.

Keyword: Outlier Detection, Isolation Forest, Graph Auto Encoder, Machine learning, Extreme Gradient Boosting