ABSTRACT

CLASSIFICATION OF CANCER TYPES BASED ON TUMOR SIGNATURE DNA USING XGBOOST METHOD

By

Rachelia Dita Anggraini

Cancer diagnosis using liquid biopsy focuses on tumor DNA signatures in the form of circulating tumor DNA (ctDNA) derived from primary tumors obtained from patient blood samples. Somatic mutations and the copy number alterations in ctDNA can be used as key markers for early detection and monitoring of cancer. The DNA mutation data when combined with bioinformatics and biomedical data analysis using machine learning algorithms can help simplify the classification of cancer. XGBoost or extreme gradient boosting is an ensemble learning method that combines several models to produce one optimal prediction model. This study aims to determine the results of the XGBoost algorithm classification using mutations that occur in DNA tumors in the patient's body. This study performs a classification using data from the research of Soh et al., 2017 using the XGBoost algorithm with two types of data sharing, namely 10-fold cross-validation and holdout cross-validation with training data distribution of test data of 60:40, 70:30, 75:25, 80:20, 90:10, 95:5 and 97:3. Each classification experiment carried out in this study uses two types of data, namely data with overall features and the top 900 features as a result of feature selection.

The highest accuracy result was obtained by the XGBoost holdout cross-validation experiment with 90:10 900 features with an accuracy of 89.91%, while the lowest accuracy value was obtained by the XGBoost 10-fold cross-validation 900 feature experiments, which was 70.01%. Based on the experiments that have been carried out, the accuracy value tends to be higher in the 900-feature experiment than the overall feature, and the accuracy of the experiment using the holdout cross-validation division gets a better accuracy value than the 10-fold cross-validation experiment. For holdout cross-validation, the greater the percentage of training data, the higher the accuracy obtained.

Keywords: classification, machine learning, bioinformatics, cancer, mutation, XGBoost

ABSTRAK

KLASIFIKASI TIPE KANKER BERDASARKAN *SIGNATURE TUMOR DNA* MENGGUNAKAN METODE XGBOOST

Oleh

Rachelia Dita Anggraini

Diagnosis penyakit kanker menggunakan biopsi cair berfokus pada *signature tumor DNA* berupa *circulating tumor DNA* (*ctDNA*) yang berasal dari tumor primer yang didapatkan dari sampel darah penderita. Mutasi somatik dan perubahan nomor salinan pada ctDNA dapat digunakan sebagai penanda utama untuk deteksi dan pemantauan dini dari penyakit kanker. Data mutasi DNA tersebut apabila dikombinasikan dengan bidang bioinformatika dan analisis data biomedis menggunakan algoritma *machine learning* dapat membantu mempermudah klasifikasi penyakit kanker. XGBoost atau *extreme gradient boosting* adalah metode pembelajaran *ensemble* yang menggabungkan beberapa model untuk menghasilkan satu model prediksi optimal. Penelitian ini bertujuan untuk mengetahui hasil klasifikasi algoritma XGBoost dengan menggunakan mutasi yang terjadi pada tumor DNA dalam tubuh penderita. Penelitian ini melakukan klasifikasi menggunakan data penelitian Soh et al., 2017 menggunakan algoritma XGBoost dengan dua jenis pembagian data yaitu *10-fold cross-validation* dan *holdout cross-validation* dengan pembagian data latih data uji sebesar 60:40, 70:30, 75:25, 80:20, 90:10, 95:5, dan 97:3. Masing-masing percobaan klasifikasi yang dilakukan dalam penelitian ini menggunakan dua jenis data yaitu data dengan keseluruhan fitur dan 900 fitur teratas hasil seleksi fitur.

Hasil akurasi paling tinggi didapatkan percobaan klasifikasi XGBoost *holdout cross-validation* 90:10 900 fitur dengan akurasi sebesar 89,91%, sedangkan nilai akurasi paling kecil diperoleh percobaan XGBoost *10-fold cross-validation* 900 fitur yaitu sebesar 70,01%. Berdasarkan percobaan yang sudah dilakukan, nilai akurasi cenderung lebih tinggi pada percobaan 900 fitur dibandingkan keseluruhan fitur dan akurasi percobaan menggunakan pembagian *holdout cross-validation* mendapat nilai akurasi lebih baik dibandingkan percobaan *10-fold cross-validation*. Untuk *holdout cross-validation* semakin besar persentase data latih semakin tinggi pula akurasi yang diperoleh.

Kata kunci : klasifikasi, *machine learning*, bioinformatika, kanker, mutasi, XGBoost