Clustering Varian Sekuens Protein SARS-CoV-2 Menggunakan Algoritma BIRCH dengan Seleksi Fitur LASSO = Clustering of SARS-CoV-2 Protein Sequence Variants Using BIRCH Algorithm with LASSO Feature Selection

Situmeang, Jason Nimrod Joshua, author

Abstrak

<p style="text-align: justify;">Penelitian ini bertujuan untuk melakukan pengelompokan varian virus SARS-CoV-2 melalui proses clustering menggunakan metode unsupervised learning. Data yang digunakan adalah sekuens protein SARS-CoV-2 yang diekstraksi fiturnya menggunakan paket Discere dalam bahasa pemrograman Python. Sebanyak 27 fitur dihasilkan dan diseleksi dengan metode seleksi fitur Least Absolute Shrinkage and Selection Operator (LASSO). Metode Elbow digunakan untuk menentukan jumlah cluster yang optimal. Dalam penelitian ini, digunakan metode clustering K-Means dan Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). Evaluasi hasil clustering dilakukan menggunakan metrik evaluasi Silhouette Score dan Davies-Bouldin Index, serta memperhatikan waktu runtime untuk setiap simulasi. Hasil evaluasi kemudian dibandingkan untuk melihat perbedaan performa antara kedua metode clustering yang digunakan, serta pengaruh seleksi fitur terhadap performa clustering. Hasil terbaik diperoleh pada simulasi dengan metode clustering BIRCH + LASSO, dengan nilai Silhouette Score 0,74186 untuk jumlah cluster k=4 dan 0,73207 untuk k=5. Nilai Davies-Bouldin Index terbaik juga diperoleh pada simulasi tersebut, yaitu 0,42697 untuk k=4 dan 0,37949 untuk k=5. Waktu runtime terbaik tercatat pada simulasi dengan metode K-Means + LASSO, yaitu 0,21551 detik untuk k=4 dan 0,17539 detik untuk k=5. Dapat disimpulkan bahwa metode BIRCH menghasilkan cluster yang lebih baik berdasarkan metrik evaluasi, namun K-Means memberikan proses clustering yang lebih cepat. Seleksi fitur dengan metode LASSO juga membantu meningkatkan performa clustering.

.....This study aims to perform clustering of SARS-CoV-2 virus variants using unsupervised learning methods. The data used consists of SARS-CoV-2 protein sequences whose features are extracted using the Discere package in the Python programming language. A total of 27 features are generated and selected using the Least Absolute Shrinkage and Selection Operator (LASSO) feature selection method. The Elbow method is employed to determine the optimal number of clusters for the clustering process. The clustering methods used in this research are K-Means clustering and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). The clustering results are evaluated using the Silhouette Score and Davies-Bouldin Index metrics, while also considering the runtime for each simulation. The evaluation results are then compared to examine the performance differences between the two clustering methods and the impact of feature selection on clustering performance. The best Silhouette Score is obtained in the simulation using the BIRCH + LASSO clustering method, with a value

of 0.74186 for k=4 and 0.73207 for k=5. The best Davies-Bouldin Index is also achieved in the same simulation, with values of 0.42697 for k=4 and 0.37949 for k=5. The fastest runtime is recorded in the simulation using the K-Means + LASSO method, with a time of 0.21551 seconds for k=4 and 0.17539 seconds for k=5. In conclusion, the BIRCH method yields better clustering results based on the evaluation metrics, while K-Means provides faster clustering processes. The LASSO feature selection method also aids in improving clustering performance.</p>