INTISARI

PENGEMBANGAN MODEL PREDIKSI EPITOP KONFORMASI SEL B BERBASIS *MACHINE LEARNING* DENGAN PENANGANAN *CLASS IMBALANCE*

Oleh
Binti Solihah
15/392460/SPA/00554

Epitop konformasi antigen Sel B merupakan komponen vaksin berbasis protein. Pengembangan vaksin dengan metode *rational design* memanfaatkan pendekatan komputasi untuk memprediksi epitop. Pendekatan berbasis *machine learning* lebih baik dibandingkan pendekatan lain, namun metode ini sensitif terhadap masalah *class imbalance*. Beberapa metode penanganan *class imbalance* sudah diterapkan, namun kinerja dari model-model yang dihasilkan masih menengah sehingga belum memenuhi kebutuhan pengguna.

Penelitian ini mengusulkan metode sampling CluSMOTE yang menggabungkan *undersampling* berbasis klaster dengan SMOTE. Algoritme HDBSCAN digunakan sebagai pembentuk klasternya. Metode ini digunakan pada pengembangan model prediksi epitop konformasi level residu. Dua model prediksi epitop level residu yaitu: (1) CluSMOTEDT yaitu gabungan CluSMOTE dan D*ecision Tree*, (2) CluSMOTESVM yaitu gabungan CluSMOTE dan SVM. Metode CluSMOTE digunakan untuk memodifikasi *bootstrap* pada *Bagging* untuk membentuk model CluSMOTEBag DT. Selanjutnya, luaran model prediksi diproses dengan metode *spatial clustering* berbasis *graph* untuk memperoleh gugus epitop.

Model dibangun dengan Dataset Rubinstein dan dievaluasi dengan metode *leave-one-out-cross-validation* (LOOCV) berbasis *complex*. Parameter kinerja yang digunakan adalah *precision, recall, AUC, Gmean, Adjusted Gmean (AGm*) dan F-*score*. Model dikomparasikan dengan beberapa metode *state of the art* dengan Dataset Kringelum dan Dataset SEPPA3. Berdasarkan hasil pengujian didapatkan AUC dan AGm tertinggi pada model CluSMOTE DT dicapai pada r=2 dengan TPR= 0,797 dan TNR= 0,834. Metode CluSMOTEBag DT memberikan kinerja terbaik. Terjadi peningkatan kinerja AUC sebesar 0.045. CluSMOTE DT dan CluSMOTEBag DT pada Dataset Kringelum lebih unggul dibandingkan beberapa metode lain. CluSMOTEBag DT juga lebih unggul dibandingkan CluSMOTE DT dan metode lainnya. Kinerja metode prediksi gugus epitop masih perlu ditingkatkan.

Kata kunci: *class imbalance*, cluster-based undersampling, SMOTE, *decision tree*, *Support Vector Machine*, epitope konformasi

ABSTRACT

DEVELOPMENT OF MACHINE LEARNING-BASED CONFORMATIONAL EPITOPE PREDICTION MODEL WITH CLASS IMBALANCE HANDLING

by

Binti Solihah
15/392460/SPA/00554

Conformational B cell epitope is a protein-based vaccine component. The development of vaccines using a rational design method utilizes a computational approach to predict the epitope. The machine learning-based approach is better than other approaches, but this method is sensitive to the imbalanced class problem. Several methods of handling the imbalance class have been implemented, but the resulting models' performance is still medium to meet user needs..

There are two prediction models of the conformational epitope developed in this study: the residue level prediction model and the patch level prediction model. Class imbalance in the residue level prediction model is solved by the CluSMOTE sampling method, which combines cluster-based undersampling with SMOTE. HDBSCAN is used to form the cluster. Apart from resampling, CluSMOTE is also used to get a balanced bootstrap for Bagging. The residual level prediction model is built using a decision tree or SVM. The decision tree is used in the CluSMOTEBagDT and non-ensemble CluSMOTEDT models. SVM is used in the non-ensemble CluSMOTESVM approach. The best-performing models are combined in the patch level prediction model, where the residue that is predicted as an epitope will be processed using the graph-based spatial clustering method to form epitope clusters.

The model was built with the Rubinstein dataset and tested with the Kringelum and SEPPA 3 datasets. Validation was carried out with complex-based LOOCV. The highest AUC and AGM in the CluSMOTE DT model were achieved at r = 2 with TPR = 0.797 and TNR = 0.834. The CluSMOTEBag DT method gives the best performance. There was an increase in AUC performance of 0.045. The CluSMOTE DT and CluSMOTEBagDT in the Kringelum dataset are superior to several other methods. CluSMOTEBagDT is also superior to CluSMOTE DT and other methods. The performance of the patch level of the epitope prediction model still needs to be improved..

Keyword: class imbalance, cluster-based undersampling, SMOTE, decision tree, support vector machine, conformational epitope