Feature selection methods for machine learning based docking prediction of Indonesian medicinal plant compounds and HIV-1 protease

Rahman Pujianto*, Yohanes Gultom†, Ari Wibisono‡, Heru Suhartanto§
Faculty of Computer Science, Universitas Indonesia
*Email: *rahman.pujianto@ui.ac.id, †yohanes.gultom@ui.ac.id, ‡ari.w@cs.ui.ac.id §heru@cs.ui.ac.id*

***Abstract*—This research evaluates feature selection methods to reduce number of features required to predict docking result between Indonesian medicinal plant compounds and HIV protease.**

Two feature selection methods on public dataset from PubChem Bioassay and DUD-E decoy which originally have 667 features of molecular description. linear SVM docking predicti of 368 Indonesian herbal chemical compounds with HIV-1 protease from PDB (3OCX). The dataset, which consists 357 positive and 11 negative samples, was labeled by performing manual docking using Autodock.

This research utilizes machine learning approach to analyze 1,412 chemical compound structures from Indonesian herbal database. From the data, 667 features are extracted from each compounds to build an unlabeled dataset. In order to build a HIV-1 protease inhibitor classification dataset, top 10 of protease inhibitors from related research are used as reference to give positive labels to the unlabeled dataset. Consequently, the remaining 1,402 compounds are labeled as negative. An SVM classifier, which is trained using different public dataset, achieves 66.7% classification accuracy on it.

1. Introduction

Simulasi dinamika molekular membutuhkan sumber daya komputasi dan waktu yang tidak sedikit. Machine learning harusnya dapat membantu memprediksi hasil simulasi dinamika molekular dengan lebih efisien. Riset ini mencoba melatih model machine learning untuk memprediksi hasil docking dari 368 senyawa herbal indonesia dengan protein HIV-1. [1]

2. Related Work

TODO

3. Experiments

Dataset dibuat dengan mencoba melakukan simulasi docking 368 senyawa herbal indonesia dengan protein HIV-1 menggunakan PyRx & Autodock. Senyawa yang berhasil docking diberi label positif (357 senyawa) dan yang gagal diberi label negatif (11 senyawa).

Eksperimen yang dilakukan: 1. Data latih dan data uji PubChem 2. Data latih PubChem, data uji HerbalDB 3. Data latih dan data uji HerbalDB

Model yang digunakan untuk tiap eksperimen adalah 1. SVM 2. SVM + RFE 3. SVM + WM (GA) 4. (ditambah sesuai kebutuhan)

4. Analysis

Model mana yang paling akurat memprediksi hasil docking senyawa dengan HIV-1

5. Conclusion

TODO

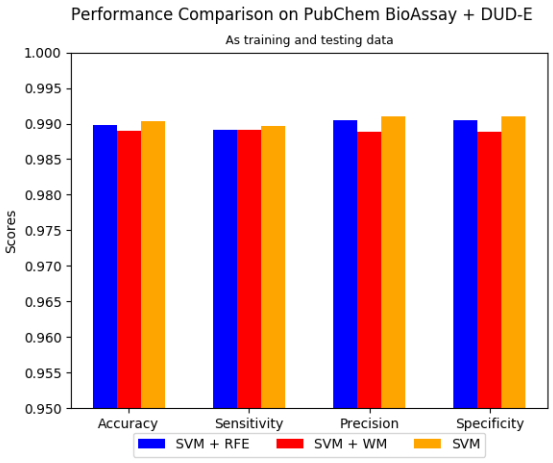

Figure 1. Classification performance comparison on PubChem BioAssay + DUD-E dataset

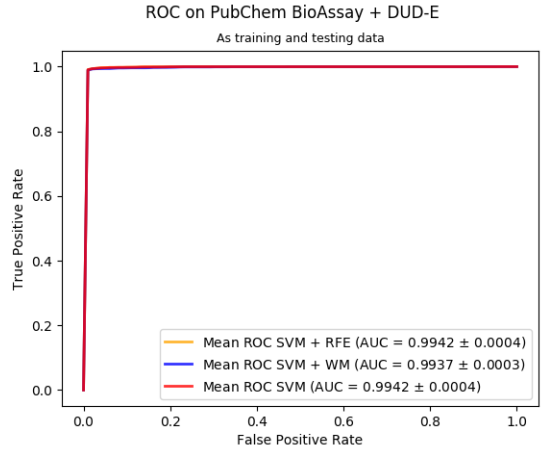

Figure 2. AUC ROC comparison on PubChem BioAssay + DUD-E dataset

References

[1] A. Yanuar, H. Suhartanto, A. Mun *et al.*, "Virtual screening of indonesian herbal database as hiv-1 protease inhibitor," *Bioinformation*, vol. 10, no. 2, p. 52, 2014.

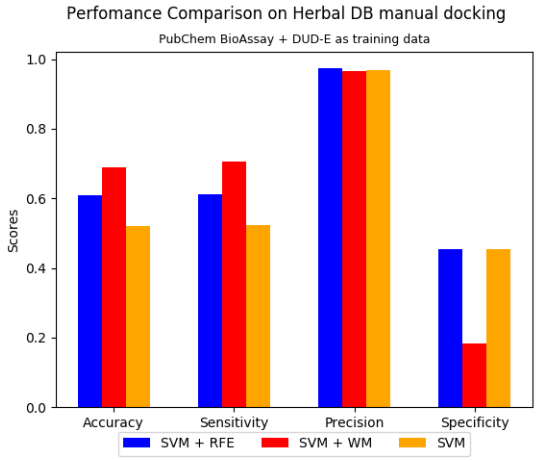

Figure 3. Classification performance comparison on Herbal DB dataset

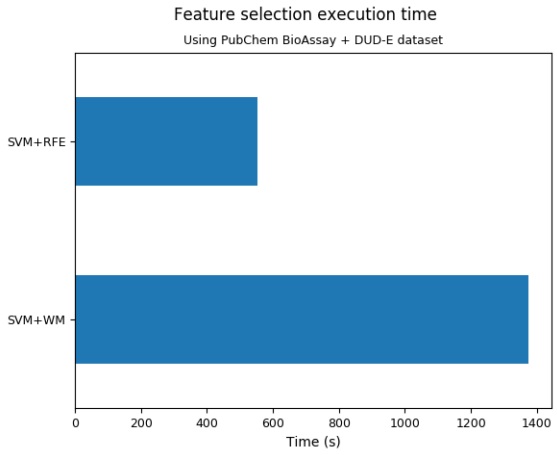

Figure 4. Feature selection time comparison

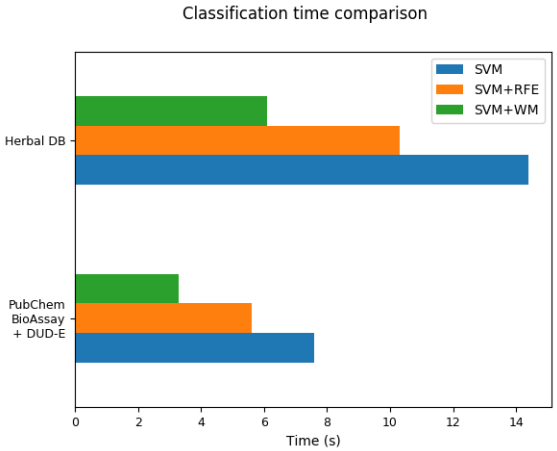

Figure 5. Classification time comparison