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 usage 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, Recursive Feature Elimination (RFE) and Wrapper Method (WM), are trained with dataset of 7,331 samples and 667 features from PubChem Bioassay and DUD-E decoys. To evaluate the selected features, a dataset of 368 Indonesian herbal chemical compounds labeled by manually docking to PDB HIV-1 protease is used to benchmark the performance of linear SVM classifier using different sets of features. Our experiments show that set of 471 features selected by RFE and WM achieve reduction of classification time by 4.0 and 8.2 seconds respectively. Although the accuracy and sensitivity are also increased by 8% and 16%, no improvement observed for precision and specificity.**

1. Introduction

The evolution of viruses can makes them resistant to existing drugs. One of the most popular case is HIV (Human Immunodeficiency Virus) which caused AIDS (Acquired Immunodeficiency Syndrome), which has been a global issue for years. HIV possesses a high drugs resistance due to its high replication and mutation abilities. Since drug discovery is a very complicated, expensive and time-consuming, curing AIDS and other illness caused by evolving virus become very challenging [1].

In order to discover new drugs, first, one needs to find a set of chemical compound candidates by observing reaction to drug target in the lab. This process is usually called high-throughput screening (HTS). Despite of its importance, this process is considered inefficient and expensive because most of chemical compounds consumed in the experiments. One way to make this process more efficient is by reducing the number of compounds that need to be tested in lab by performing virtual screening beforehand [2]. By having the number of lab experiments reduced, ultimately it will reduce overall time and cost needed in drug discovery [3].

Virtual screening applies computer algorithms to find chemical compounds that have high probability of reaction to drug's target. One of its approach is ligand-based screening (LBS), where new candidates are chosen based on their structural or characteristic similarity to known drug's chemical compounds. This implies that LBS approach relies on previous drug discovery results, which usually obtained using HTS such as PubChem BioAssay [4], CHEMBL [5], PubChem Compound [6] and ZINC [7].

Since LBS is also a pattern matching problem, supervised learning algorithms can be used to classify chemical compounds using database of known drug descriptions as training dataset. The number of features required to described each compounds also affects the performance of both

supervised and unsupervised learning algorithms. This phenomenon is usually addressed as the curse of dimensionality [8]. Two techniques commonly applied to solve this phenomenon are feature extraction and feature selection. While the first one extracts or processes existing features to get set of new ones, the last one selects a subset of features from the existing ones. This research focuses on observing the performance of two feature selection methods, SVM Recursive Feature Elimination (SVM-RFE) and Wrapper Method (WM), to select subset of features from Indonesian herbal chemical compounds that react to HIV-1 protease.

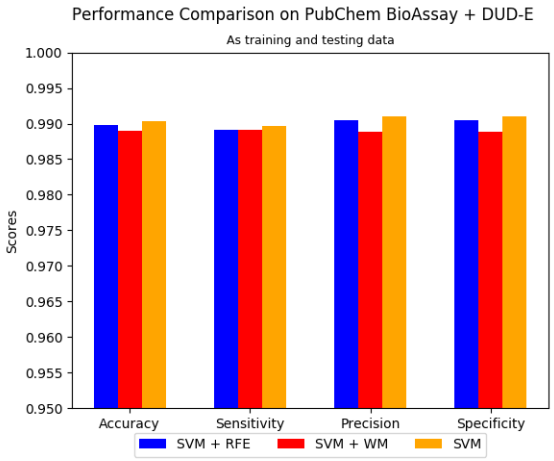

Figure 1. Classification performance comparison on PubChem BioAssay + DUD-E dataset

2. Related Work

Related research in virtual screening used a method that consist of two phases: First, machine learning based LBS is used to select potential chemical compound candidates, and second, molecular docking is done with between potential candidates and drug's target [9]. Since molecular docking requires a lot of computational resources, high LBS precision is required to improve efficiency. In the other hand, low recall or sensitivity causes potential candidates excluded [3]. This research shows Support Vector Machine (SVM) performs well to classify potential candidates in LBS. Using this as basis, we explore usage of feature selections to improve SVM performance in LBS.

Molecular descriptor is a numerical value representing chemical information encoded within a symbolic representation of a molecule. This numerical value can also be obtained by some standardized experiment on a molecule [10]. At least there are 701 types of molecular descriptor that can be extracted from a chemical compounds. Therefore, it is difficult to analyze manually all correlations between descriptors [3]. In machine learning based LBS, not all molecular descriptors directly affect the result of classification. For instance, the number of Bromin (Br) atom is always 0 for every compounds in PubChem BioAssay database. There are even around 500 descriptors behaving in such way in the same database. Therefore, it is also recommended to reduce the number of features by using techniques like Feature Selection [3].

Feature selection can improve accuracy of classification task, and also improves its efficiency by reducing computational costs. On top of that, it can give better understanding about the resulted model as suggested by another related research [8]. But it should also be noted that improvement given by application of feature selection is depending on the type of data. Hence, the result of its application may vary between datasets [8]. To anticipate this, our experiments use datasets from two different sources: public source (PubChem BioAssay + DUD-E) and Indonesian Herbal DB.

3. Dataset

TODO

4. Feature Selection

5. Experiments

TODO

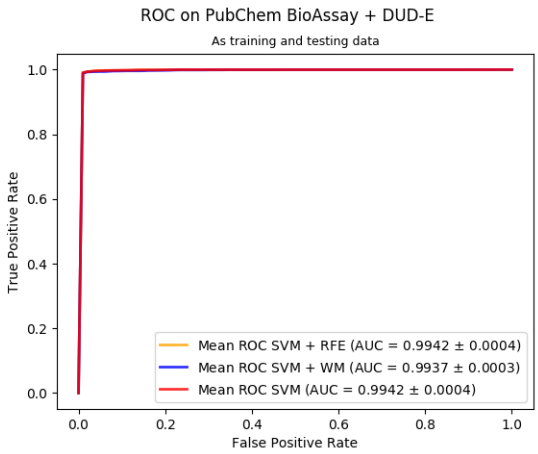

Figure 2. AUC ROC comparison on PubChem BioAssay + DUD-E dataset

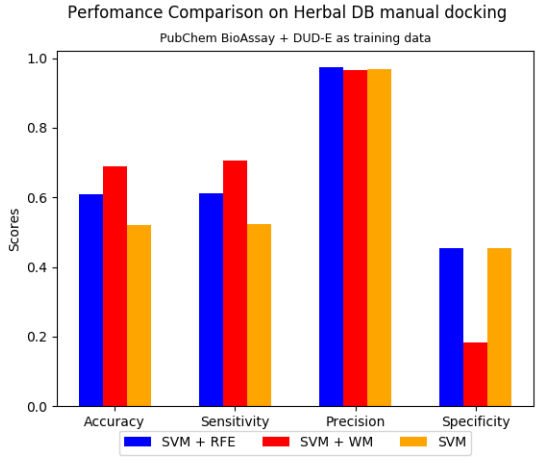

Figure 3. Classification performance comparison on Herbal DB dataset

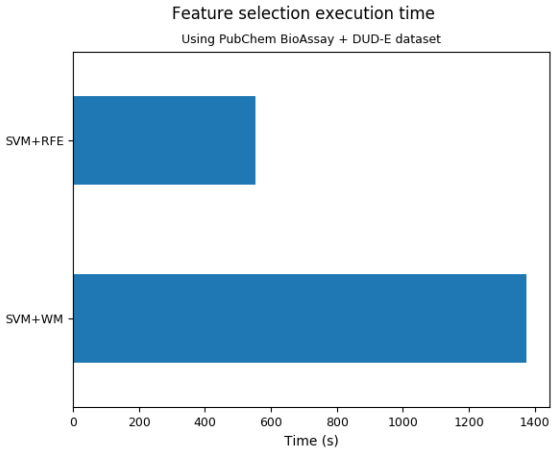

Figure 4. Feature selection time comparison

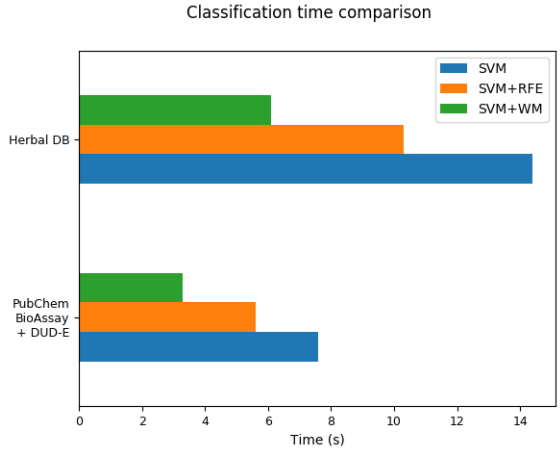

Figure 5. Classification time comparison

6. Analysis

TODO

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

7. Conclusion

TODO

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.

[2] J. J. F. Chen and D. P. Visco Jr, "Developing an in silico pipeline for faster drug candidate discovery: Virtual high throughput screening with the signature molecular descriptor using support vector machine models," *Chemical Engineering Science*, vol. 159, pp. 31–42, 2017.

[3] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "Drug/nondrug classification using support vector machines with various feature selection strategies," *computer methods and programs in biomedicine*, vol. 117, no. 2, pp. 51–60, 2014.

[4] P. BioAssay, "update wang," 2014.

[5] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey *et al.*, "The chembl bioactivity database: an update," *Nucleic acids research*, vol. 42, no. D1, pp. D1083–D1090, 2014.

[6] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[7] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "Zinc: a free tool to discover chemistry for biology," *Journal of chemical information and modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.

[8] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 2008, pp. 90–105.

[9] M. Hilman, "Analisis teknik data mining dan kinerja infrastruktur komputasi cloud sebagai bagian dari sistem perancangan obat terintegrasi," *Graduate Thesis*, 2012.

[10] C. W. Yap, "Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.